# *Bias and Discrimination in Opaque Automated Individual Risk Assessment Systems: Challenges for Judicial Review under the Equality Act 2010*

Gianna Seglias[*]

**Abstract**—Public authorities' use of automation to assist decision-making poses a novel challenge for administrative law principles, which were developed with reference to human decision-making processes. This article explores one particular aspect of this challenge. It considers how public authorities utilise automated systems to assess the statistical risk of individuals engaging in particular types of undesirable behaviour, and explains how these technologies may facilitate biased outcomes. It then analyses how judicial review under the Equality Act 2010 could be used to challenge the use of such systems, focussing in particular on sections 29 and 149. Finally, it outlines the evidential challenge posed by algorithmic opacity and the potential of the public sector equality duty to mitigate this difficulty.

[*] University College London. I am grateful to the OUULJ editorial team for comments on earlier drafts. All errors remain my own.

# *Introduction*

The combined effect of the public sector's enthusiasm for new technologies[1] and the legacy of New Labour-era governance focussed on strategic risk management[2] is that public authorities have been eager to adopt algorithmic tools to assess the statistical risk of individuals behaving in undesirable ways. While these technologies have the potential to facilitate efficiently targeted intervention, concerns have been raised about the potential for biased outcomes.[3] Where the burden of technological bias falls on a group possessing a protected characteristic under the Equality Act 2010 (EA 2010), decisions made using these tools may constitute unlawful discrimination under UK equality law. This article will consider whether administrative law is sufficiently equipped to protect individuals from discriminatory decisions made with the assistance of machine learning automated individual risk-assessment systems (AIRAS) and argue that the existing grounds of review under the EA 2010 may have significant potency. Part 1 introduces the workings and uses of AIRAS, considers their discriminatory effects, and explains the evidential challenge posed by algorithmic opacity. Part 2 explores the potential of EA 2010 sections 29 and 149 in challenging the use of discriminatory AIRAS, noting in particular the potential of the public sector equality duty (PSED) to create much-needed transparency. Part 3 concludes that although piecemeal review under the EA 2010 may be an effective mechanism for combating

[1] Government Digital Service, *Government Transformation Strategy* (2017) <www.gov.uk/government/publications/government-transformation-strategy-2017-to-2020> accessed 5 April 2021.

[2] Carol Harlow and Rick Rawlings, *Law and Administration* (3rd edn, CUP 2009) 73ff.

[3] For example see Joshua Kroll et al., 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review 633.

discrimination in AIRAS in the short term, the creation of *ex ante* accountability mechanisms could ultimately offer a more comprehensive and principled response to the questions raised by algorithmic bias.

## *1. How AIRAS Discriminate*

Risk assessment systems are a subset of automated decision-making systems which use algorithms to make decisions with reduced or no human input. AIRAS are concerned with providing a decision-maker with information about whether an individual is 'at risk' of having a certain characteristic or carrying out some undesirable activity. These systems calculate this risk by processing known data. For example, if the relevant characteristic or activity, or *output variable*, is child abuse, the model will consider past recorded instances of abuse to determine which other factors, or *input variables*, correlate to the occurrence of this event. Some AIRAS are systems in which the input variables are specified and their statistical weight in contributing to the output is determined by programmers. However, other systems utilise 'machine learning', a form of artificial intelligence through which a system autonomously develops its decision-rules by being run over a set of training data.[4]

[4] ibid 638.

## *A. Uses of AIRAS*

AIRAS are already in use by various UK public authorities.[5] In the criminal justice system, they are used to make decisions about policing, charging, and parole.[6] They have also been used by children's services to identify children at risk of abuse and to make safeguarding decisions.[7] A potential area for expansion of AIRAS is in immigration and border control, where predictive analytics could be used to determine which individuals are likely to become over-stayers or pose a danger to national security. In the future, AIRAS are likely to expand to other areas of citizen-state interaction, particularly where public authorities working with reduced budgets are looking to implement more cost-effective decision-making procedures.

## *B. Potential for Discriminatory Decisions*

Machine learning AIRAS are frequently biased. Under a broad understanding of the term, some form of 'bias' is inherent to such systems because they use data about *groups* of people to predict the behaviour of *individuals*. However, many AIRAS are also 'biased' in a more narrow sense; this latter understanding, which

[5] For examples see Lina Dencik et al., 'Data Scores as Governance: Investigating uses of citizen scoring in public services' (2018) <https://datajustice.files.wordpress.com/2018/12/data-scores-as-governance-project-report2.pdf> accessed 5 April 2021.

[6] See Michael Veale et al., 'Algorithms in the Criminal Justice System' (Law Society 2019) <www.lawsociety.org.uk/en/topics/research/algorithm-use-in-the-criminal-justice-system-report> accessed 5 April 2021.

[7] Hackney Council ended a pilot programme, the Early Help Profiling System, in 2019. For an international comparison, see the Allegheny Family Screening Tool, described in Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (St Martin's Press 2018).

will be adopted in this paper, refers to systems which make predictions that are systematically too high or too low for specific subgroups.[8] Where the relevant group corresponds to individuals possessing a 'protected characteristic' (as defined in sections 4-12 EA 2010), this bias may constitute discrimination in the legal sense. Kroll et al. identify three ways in which automated decision-making systems 'simultaneously systematise and conceal discrimination'.[9] Firstly, they may be trained on data sets reflecting past prejudice, meaning that the model treats prejudiced decisions as an example to learn from.[10] Secondly, they can build in discrimination through model construction. This may arise in particular at the stage where data is selected to be considered by the model (feature selection).[11] For example, it is common in AIRAS to identify membership of a protected class as an input, because protected characteristics may indeed correlate with the relevant outcome risk variable. In a child abuse prediction model, for instance, people of a particular race or gender may be statistically more likely to have their children taken into foster care following child abuse incidents. Even if these characteristics are not fed into the algorithm, they may 'sneak back in' through proxies, which are included variables containing 'signals' predicting the excluded variable.[12] For example, postcode may

[8] Partnership on AI, 'Report on Algorithmic Risk Assessment Tools in the US Criminal Justice System' (2019) <www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/> accessed 5 April 2021, 15.

[9] Kroll et al. (n 3) 680.

[10] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2017) 104 California Law Review 671.

[11] Kroll et al. (n 3) 681.

[12] Michael Veale and Lilian Edwards, 'Slave to the Algorithm? Why a "right to an explanation" is probably not the remedy you are looking for' (2017) 16 Duke Law & Technology Review 18, 29.

predict race.[13] Thirdly, systems may 'mask' intentional discrimination by a prejudiced decision-maker or model developer.[14] Even where the final human decision-maker is themselves unbiased, discrimination in the system is replicated if the decision-maker overly relies on the automated decision.[15]

## *C. Case Studies*

There are two AIRAS about which there has been significant commentary: COMPAS, in the USA, and HART, in the UK. Both are used in the criminal justice system to assess the risk of recidivism and each has been accused of facilitating biased decisions. Reference will be made to these case studies throughout the paper.

### *I. COMPAS*

The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system is used by courts in several American jurisdictions to assess recidivism risk for the purpose of various criminal justice decisions, including sentencing. Among other things, the system provides both a General Recidivism Risk scale, for which the output variable is an arrest within two years of the intake assessment, and a Violent Recidivism scale, for which the output variable is an arrest within

[13] Black and minority ethnic communities are more concentrated in certain postcodes - see Office for National Statistics, '2011 Census' <www.nomisweb.co.uk/census/2011/ks201ew> accessed 5 April 2021.

[14] Kroll et al. (n 3) 682.

[15] See comments on 'automation bias' by Reuben Binns as reported in Alice Irving, 'Rise of the algorithms' (2019) (*UK Human Rights Blog*, 4 November 2019) <https://ukhumanrightsblog.com/2019/11/04/rise-of-the-algorithms/> accessed 5 April 2021.

two years for an offence on a person.[16] Not much is known about the algorithm employed by COMPAS, as the software was privately developed by Northpointe (now trading as Equivant) and continues to be owned by the company. The risk scores are based on public criminal records[17] and answers to a 137-question survey which tracks both static factors, such as criminal history, and dynamic factors, such as attitudes towards right and wrong.[18]

A 2016 investigation by ProPublica on the use of COMPAS in Broward County, Florida, concluded that the system disadvantaged Black defendants.[19] While the tool made errors at roughly the same rate for Black and white defendants, the *direction* of error made differed according to defendants' race. The system was more likely to mislabel Black defendants as high risk and conversely more likely to mislabel white defendants as low risk. Even when accounting for criminal history, age, and gender, Black defendants were 77 percent more likely to receive a higher score on the Violent Recidivism scale and 45 percent more likely to receive a higher score on the General Recidivism Risk scale. This disparity is illustrated in the ProPublica study by comparing two real-life examples of erroneous predictions made in relation to two petty theft suspects: a white defendant who was rated low-risk despite previous convictions for armed robbery and who subsequently committed burglary, and a Black defendant with

[16] Northpointe, 'COMPAS Risk & Need Assessment System: Selected Questions' (2012) <www.northpointeinc.com/files/downloads/FAQ_Document.pdf> accessed 16 June 2020.

[17] *Loomis v Wisconsin* 881 NW2d 749 (Wis 2016) [55].

[18] Northpointe, 'Risk Assessment' (2012) <www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html> accessed 5 April 2021.

[19] Julia Angwin et al., 'Machine Bias' (*ProPublica* 23 May 2016) <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 5 April 2021.

only a juvenile misdemeanour record who was rated high risk, but has not since been charged with any further offences.

In response to the investigation, Northpointe 'strongly reject[ed]' the allegation of bias, maintaining that COMPAS was 'equally accurate' for all races,[20] since it correctly classified people as recidivists at the same rate.[21] These contradictory framings of the model—as either unbiased because it makes errors at an equal rate for both groups, or as biased because the burden of false positives falls disproportionately on one group—reflect the 'impossibility theorem of fairness'. This is a statistical rule which holds that where there are different rates of occurrence of an output variable between groups, a model can either have equal predictive accuracy for both groups *or* achieve equal rates of false positives and negatives for both groups, but cannot simultaneously satisfy both fairness criteria.[22] Accordingly, even an instrument seemingly free from predictive bias may result in a 'disparate impact' in systems where a higher risk score results in disadvantage for the subject.[23]

## *II. HART*

Durham Constabulary's Harm Assessment Risk Tool (HART) is used at the point of arrest to identify individuals at low or moderate risk of recidivism, who are to be referred to the

[20] William Dieterich, Christina Mendoza and Tim Brennan, 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity' (Northpointe 2016) <http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf > accessed 21 April 2021, 1.

[21] ibid 7.

[22] Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' (FATML conference, New York, November 2016).

[23] ibid 1.

'Checkpoint' programme rather than being charged. HART utilises 34 input variables and employs a random forest algorithm (a type of machine learning) to classify offenders into three risk groups. The model intentionally favours 'cautious' errors, meaning it is more likely to over-predict risk than it is to under-predict.[24] The system was developed in-house by Durham Constabulary and details of the model's construction are publicly available.[25]

While there has been no comprehensive investigation of HART similar to that conducted by ProPublica on COMPAS, HART may nevertheless be discriminatory since it utilises information related to the protected characteristics of race, gender and age. Gender and age are used directly as inputs, and two of the input variables include postcode data[26] which can act as a proxy for race.[27] Since communities of colour in the UK, particularly Black communities, are frequently over-policed,[28] use of such data may generate a 'feedback loop that may perpetuate

[24] Marion Oswald et al., 'Algorithmic risk assessment policing models: lessons from the Durham HART model and "Experimental" proportionality' (2018) 27 Information and Communications Technology Law 223, 228.

[25] See Sheena Urwin, 'Algorithmic Forecasting of Offender Dangerousness for Police Custody Officers: An Assessment of Accuracy for the Durham Constabulary Model' (MSc thesis, University of Cambridge 2016) <www.crim.cam.ac.uk/system/files/documents/sheena-urwin-thesis-12-12-2016.pdf> accessed 5 April 2021, Appendix B.

[26] Oswald et al. (n 24) note that Durham Constabulary intended to remove one of two postcode predictors but it is unclear whether this took place.

[27] See the 2011 census (n 13).

[28] Home Affairs Committee, *The MacPherson Report – 10 Years On* (HC 2008-2009, 427).

or amplify existing patterns of offending…leading to an ever-deepening cycle of increased police attention'.[29]

## *D. System Opacity and Evidential Difficulties*

AIRAS may, however, pose evidential difficulties for litigants seeking to establish discrimination, due to algorithmic opacity. Cobbe differentiates between intentional, illiterate, and intrinsic opacity.[30]

*Intentional* opacity refers to the situation where the workings of a system are deliberately concealed, either because the technology provider is a for-profit entity or because the public authority wishes to prevent individuals from 'gaming' the system. For example, the algorithm used in COMPAS is a trade secret owned by Equivant, and, following a constitutional challenge to the tool, the Wisconsin Supreme Court noted that without access to the source code it could not evaluate 'how the risk scores are determined or how the factors are weighed'.[31]

Even where the algorithm utilised in an AIRAS is publicly available, it may be understandable only to technological experts (*illiterate* opacity) or not understandable to humans at all (*intrinsic* opacity). Like many machine learning systems, HART is intrinsically opaque due to its complexity. Oswald et al. note that the algorithm is 4.2 million 'nested and conditionally-dependent decision points',[32] meaning that any one variable has no direct impact on the result and that discriminatory effects may be so small that they are not noticeable in any particular case, but give

[29] Oswald et al. (n 24) 228.

[30] Jennifer Cobbe, 'Administrative law and the machines of government: judicial review of automated public-sector decision-making' (2019) 39 Legal Studies 636.

[31] *Loomis v Wisconsin* (n 17) [51].

[32] Oswald et al. (n 24) 228.

rise to significant 'cumulative disadvantage' across the system.[33] System opacity may be further compounded by data protection obligations. For example, Veale et al. note that the kind of investigation conducted into COMPAS would not be possible in the UK as the training data used would likely not be released under UK freedom of information law.[34]

The high level of opacity present in AIRAS harms the accountability of public authorities, which may be understood as the obligation of public officials to 'explain and justify their conduct' so that these actions may be subjected to scrutiny and, where appropriate, result in consequences.[35] Accountability between the governing and the governed is disrupted where decision-making occurs through procedures which the public cannot understand. In particular, legal accountability through judicial review may be hindered where decision-making processes become inscrutable or indecipherable. Because public law review mechanisms are 'primarily concerned with decision-making processes',[36] algorithmic opacity may mean that claimants are unable to gather the evidence necessary to establish that a decision is affected by public-law error.

The urgency of legal reform to facilitate transparency and explainability of algorithmic systems has been recognised in the literature,[37] but this article will not offer a policy exploration of

[33] See Oscar Gandy, 'Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems' (2010) 12 Ethics and Information Technology 29.

[34] Veale et al. (n 6) 21, note 66.

[35] Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 European Law Journal 447.

[36] Cobbe (n 30) 649.

[37] See for example Swee Leng Harris, 'Data Protection Impact Assessments as rule of law governance mechanisms' (2019) 2 Data & Policy 1.

how to remedy opacity in AIRAS or automated systems more generally. It will concern itself with the narrower question of how EA 2010 sections 29 and 149 might be utilised to challenge discriminatory AIRAS, and will consider the impacts of opacity on the prospect of success of challenges under those provisions. The following section will argue that existing legal principles under the EA 2010 can offer significant protection to claimants, and will conclude that opacity is unlikely to frustrate such claims. Instead, opacity may in some instances work in claimants' favour, by allowing them to take advantage of legal principles to lower their evidential burden while simultaneously rendering it more difficult for public authorities to meet the procedural and substantive burdens placed upon them by equality law. Accordingly, judicial review claims under the EA 2010 may not only offer relief to claimants, but could contribute to the broader aim of combating algorithmic opacity.

## *3. Legality Review Under the Equality Act 2010*

There are, in theory, many avenues of judicial review through which claimants may challenge either individual discriminatory decisions made through AIRAS, or policy decisions on the design and use of such tools. This article, however, will consider only review under the EA 2010, which is directly concerned with equality issues and would accordingly provide litigants with the most straightforward route. While the EA 2010 has largely been utilised for civil claims against the state, a decision which fails to comply with the EA 2010 may also be judicially reviewed as *ultra* *vires.*[38] In particular, a decision is unlawful where it constitutes unlawful discrimination under section 29(1), (2) or (6),[39] or fails to comply with the PSED in section 149(1).

## *A. Prohibition on Direct and Indirect Discrimination*

Decisions may be challenged as violating EA 2010 section 29 where they either directly or indirectly discriminate against the claimant.

### *I. Substantive Requirements*

Section 13(1) defines direct discrimination as less favourable treatment 'because of' a protected characteristic as defined in sections 4-12. The Supreme Court held in R *(Coll) v Secretary of State for Justice*[40] that a greater *risk* of a disadvantage which materialises for the claimant is sufficient to make out direct discrimination even if not all members of the protected group suffer the disadvantage; in that case, it sufficed that women were at a higher risk of being required to live in premises further from their home after release from prison due to the lower number of facilities for women. Allen and Masters argue that following *Coll*, it should be possible to argue that risk asymmetry under an automated system gives rise to less favourable treatment where people possessing a protected characteristic are at higher risk of an unfavourable score.[41] This focus on risk dovetails with the

[38] EA 2010 s 113(3)(a).

[39] The prohibition against discrimination in the provision of services in s 29(1)-(2) also includes provision of service in the exercise of a public function: s 31(3).

[40] [2017] UKSC 40.

[41] Robin Allen and Dee Masters, 'In the Matter of Automated Data Processing in Government Decision Making' (*AI Law Hub*, 7

public law doctrine of systemic unfairness developed in recent cases which places emphasis on a 'risk of unfairness, rather than its realisation'.[42] Courts will not be receptive to arguments that actual statistical difference between populations means there is no discrimination, since equality law mandates that 'individuals should be treated as individuals, and not assumed to be like other members of a group'.[43] For example, since the HART algorithm uses age and sex as input variables, claimants should be able to establish direct discrimination if the system disadvantages them on these grounds (though age discrimination can be objectively justified under section 13(2)).[44]

However, a challenge may be more difficult where a protected characteristic is only considered through a proxy variable, such as the inclusion of race through the postcode variable in HART. Lady Hale reaffirmed in *Coll* that where the criterion used by the discriminating actor is a mere proxy for the protected characteristic, there must be 'exact correspondence' between the protected characteristic and the 'disadvantaged class'

September 2019) <www.cloisters.com/wp-content/uploads/2019/10/Open-opinion-pdf-version-1.pdf> accessed 5 June 2021, 14.

[42] Abi Adams-Prassl and Jeremias Adams-Prassl, 'Systemic Unfairness, Access to Justice and Futility: A Framework' (2020) 40 Oxford Journal of Legal Studies 561, 567.

[43] *R(E) v Governing Body of JFS* [2009] UKSC 15 [90].

[44] While sch 3 para 3(1)(c) provides that s 29 does not apply to 'a decision not to commence or continue criminal proceedings' the wording of this section implies that *positive* decisions to charge prosecute are not excluded. It should be noted, however, that a system like COMPAS used in sentencing decisions would likely not be reviewable since most judicial decisions are not subject to judicial review and in any event sch 3 para 3(1)(a) excludes judicial functions from consideration under s 29.

(i.e. the group of people disadvantaged by the provision).[45] It is unlikely that this test would be met in most cases, since proxy variables do not usually track protected characteristics perfectly. As a result, some individuals who do not possess the protected characteristic suffer the disadvantage while some individuals who possess the protected characteristic do not. This may be illustrated by considering HART. Assuming that the tool assigns individuals a less favourable risk score because they reside in a postcode with higher populations of ethnic minority residents, some ethnic minority residents who live in majority-white postcodes will not suffer this disadvantage. Conversely, some white residents of postcodes with higher populations of ethnic minority residents will be disadvantaged even though they do not share the protected characteristic.

In proxy cases, claimants may instead seek a declaration that use of an AIRAS constitutes indirect discrimination contrary to section 19. Following section 19(1), a person or organisation discriminates against an individual indirectly by applying to them a provision, criterion or practice (PCP) which, though apparently neutral, is in practice discriminatory. A PCP is to be construed widely,[46] and so could include the algorithm itself, the policy used to guide its implementation, or its training dataset.[47] A claimant would need to demonstrate that the PCP places them and others sharing their protected characteristic at a particular disadvantage when compared to those who do not share the protected

[45] *Coll* (n 40) [28]-[29].

[46] Equality and Human Rights Commission, 'Services, Public Functions and Associations: Statutory Code of Practice Equality Act 2010' (2011) <www.equalityhumanrights.com/sites/default/files/servicescode_0.pdf> accessed 5 April 2021, 70.

[47] Allen and Masters (n 41) 16.

characteristic.[48] Helpfully for challenges to AIRAS, UK courts have held that a relatively small discrepancy is sufficient to show disadvantage[49] and that disparate impact can be established on the basis of statistical evidence.[50] If it were shown, for example, that ethnic minority individuals were statistically more likely than white individuals to be given a higher risk score by HART and accordingly less likely to be referred to Checkpoint, a claimant could argue the tool (or related policy guidance) to be a *prima facie* discriminatory PCP.

While the two types of discrimination are distinct from one another and 'mutually exclusive',[51] so that a finding of direct discrimination 'rules out' a finding of indirect discrimination,[52] they are frequently pleaded in the alternative. In challenges to AIRAS, which are likely to focus on decision-making *systems* and *risk* of disadvantage evidenced through statistical disparities, the two types of discrimination may be difficult for courts to distinguish in practice. As noted above, they are likely to find the dividing line between the two wrongs to be whether the protected characteristic is an input variable or merely considered through a proxy.

The distinction between direct and indirect indiscrimination is of practical importance, since indirect discrimination can be justified where the PCP is shown to be a proportionate means of achieving a legitimate aim.[53] Helpfully to claimants, it will be difficult for the public authority to justify a

[48] s 19(2).

[49] *London Underground v Edwards (No 2)* [1999] ICR 494 (CA).

[50] *Essop and ors v Home Office* [2017] UKSC 27 [28].

[51] Bob Hepple, *Equality: The Legal Framework* (2nd edn, Hart 2014) 81.

[52] *Coll* (n 40) [43].

[53] EA s 19(2)(d).

PCP where it has failed to comply with the PSED.[54] While courts have not yet applied the *Bank Mellat* proportionality test to an AIRAS,[55] its requirement of 'rational connection to the objective' may be a pressure point for litigants, as some AIRAS have been shown to be only marginally more accurate or even less accurate than manual risk assessments.[56] Claimants may also rely on the importance of the interests affected by a decision—and correspondingly the 'seriousness of the detriment'[57] caused by the discrimination—to argue that the decision did not strike a fair balance between individual rights and community interests. Moreover, the level of scrutiny to which courts subject government justifications may be heightened if judges have regard to the academic criticism of discriminatory AIRAS indicating that such systems risk entrenching systemic injustice as well as disadvantaging individuals.[58] What remains uncertain is how these factors will be weighed against submissions by public authorities emphasising the significant public benefit created by the cost effectiveness and efficiency of AIRAS-assisted decision-making.

## *II. Evidential Challenges*

Algorithmic opacity creates evidential challenges for claimants in establishing that an AIRAS is discriminatory. These challenges are especially acute for claimants alleging direct discrimination, as in the context of a complex or privately owned machine learning AIRAS it will be near-impossible to positively establish that the

[54] 'Statutory Code of Practice' (n 46) 81. See for example *R(Ward) v Hillingdon LBC* [2019] EWCA Civ 692 [69].

[55] *Bank Mellat v HM Treasury (No 2)* [2013] UKSC 39.

[56] Julia Dressel and Hany Farid, 'The accuracy, fairness and limits of predicting recidivism' (2018) 4 Science Advances.

[57] R *(Elias) v Secretary of State for Defence* [2006] IRLR 934 [151].

[58] For example see Oswald et al. (n 24) 228; Veale et al. (n 6) 18.

reason for an individual's less favourable treatment is membership of a protected group.

Tomlinson, Sheridan and Harkens have argued that courts should take a 'distinct approach' to the standard of evidence in judicial review of opaque automated systems, by either adjusting the standard depending on the degree of opacity in a system, or shifting the duty of explaining how a system works to the public authority.[59] However, they concede that principles of evidence in judicial review are difficult to distill, lying in the 'legal subsconscious',[60] and there is no guarantee that courts would choose to adopt such an approach.

However, claimants may be aided by established equality law principles applied to the automated decision-making context. EA 2010 subsections 136(2) and (3) reverse the burden of proof where a claimant successfully establishes facts from which the court could decide that discrimination had occurred. In *ex p Danfoss*,[61] the European Court of Justice (ECJ) held that a lack of transparency may give rise to an inference of direct discrimination where less favourable treatment is established. In that case, the claimants alleged sex-based pay discrimination, relying on a statistical disparity in average pay; the Court found this sufficient to place the burden on the employer, since it was not possible for employees to identify the criteria used to determine pay. Allen and Masters consider this equality law principle to be generally applicable to public functions.[62] While ECJ jurisprudence has never been binding in UK judicial proceedings unless the public

[59] Joe Tomlinson, Katy Sheridan, and Adam Harkens, 'Proving Public Law Error in Automated Decision-Making Systems' (PLP Annual Conference, London, October 2019) 13.

[60] ibid 3.

[61] Case C-109/88 [1991] ICR 74.

[62] Allen and Masters (n 41) 51.

authority was acting to fulfil EU law obligations, and its status was recently further diminished by the end of the Brexit transition period,[63] much of the EA 2010 is nevertheless derived from the UK's EU law obligations. Applying the principle that 'it is inconceivable that Parliament intended the same concepts to be interpreted differently in different contexts',[64] courts adjudicating an AIRAS may draw a similar inference to that in *Danfoss*. It is therefore plausible that a claimant seeking judicial review of an opaque AIRAS could persuade a court to infer direct discrimination so long as they were able to identify a statistical disparity in outcome between members of a protected group and others,[65] placing the burden of proof on the public authority; where a public authority chooses to maintain intentional opacity or where the system is intrinsically opaque, it will be unable to discharge this burden.

Claimants in indirect discrimination cases would not need to establish a causal link between their membership of a protected group and the disadvantage suffered, and so would be able to avail themselves of the section 136 presumption by merely demonstrating a statistical disparity in outcomes. Though these claimants would not need to rely on the *Danfoss* inference, the level of opacity present within a system would nevertheless be relevant to the court's assessment. In order to rebut the presumption of discrimination, the public authority would need to establish either that there was no causal link between the PCP

[63] See for example David Feldman, 'Departing from Retained EU Case Law' (*UK Constitutional Law Association Blog*, 11 January 2021) <https://ukconstitutionallaw.org/2021/01/11/david-feldman-departing-from-retained-eu-case-law/> accessed 5 April 2021.

[64] *Essop* (n 50) [19] (Lady Hale).

[65] Indeed, there has been increased focus by courts on the use of quantitative data to prove error: Tomlinson, Sheridan and Harkens (n 59) 6, referring to *R (UNISON) v Lord Chancellor* [2017] UKSC 51.

and the disadvantage suffered by the individual, or, if it failed to do so, that the measure is justified. It would be very difficult to demonstrate that a PCP satisfies a proportionality test where a public authority itself is unable to comprehend its workings, since, for example, the public authority will be unable to establish that no less discriminatory measures would have satisfied the desired objective. Moreover, complete opacity will, as discussed below, make it difficult for the authority to satisfy its PSED which, in turn, is likely to harm the authority's ability to justify the measure.

## *B. The PSED*

### *I. The Potential of the PSED to Eliminate Opacity*

The PSED, properly harnessed, has significant potential to prevent discriminatory AIRAS decisions and to counter opacity. Under EA 2010 section 149(1), public authorities have a mandatory duty to have 'due regard' to three equality issues when exercising their functions.[66] The duty applies to all 'functions' of public authorities, so that both policy and individual decisions are subject to challenge;[67] in respect of AIRAS, challenges may be made to the procedure followed in making a high-level decision to introduce a tool into decision-making, but also to the procedural element of individual decisions made based on risk scores.

[66] The three issues are the need to: (a) eliminate discrimination and other conduct prohibited by the Act; (b) advance equality of opportunity between those who share a relevant protected characteristic and those who do not; and (c) foster good relations between those who share a relevant protected characteristic and those who do not.

[67] *Pieretti v Enfield LBC* [2010] EWCA Civ 1104 [26].

Though subsections 149(3)-(5) elaborate the terms of these three duties in detail and they are conceptually distinct from one another, the jurisprudence has largely conflated them to focus on the more general question of whether there has been adequate consideration of the negative equality impacts of decisions.[68] While this is generally a question for the court,[69] the *weight* to be given to the relevant section 149(1) 'needs' is a matter for the decision-maker, subject only to rationality review.[70] A successful PSED challenge will render a decision void for illegality and will have the added effect of lending support to a concurrent indirect discrimination claim.

Although the PSED is a procedural rather than substantive obligation and decision-makers are not prevented from simply making the same decision again after it has been quashed, the duty may produce substantive outcomes indirectly due to its potential to combat algorithmic opacity. Hickman notes that because compliance is a fact-specific question, the courts give 'very close scrutiny' to questions of fact in PSED cases.[71] In judicial review proceedings, reliance is ordinarily placed on the public authority's duty of candour, so that proceedings are conducted on the basis that a decision-maker's witness statement about the basis for a decision is accurate. In PSED cases, however, the courts have been willing to closely review primary evidence—including internal documentation—and reject the views of public authorities,[72] resulting in an 'onerous' burden on

[68] Tom Hickman, 'Too hot, too cold or just right? The development of the public sector equality duties in administrative law' [2013] Public Law 325, 326.

[69] R *(JM) v Isle of Wight Council* [2011] EWHC 2911 (Admin) [102].

[70] R *(Harris) v Haringey LBC* [2010] EWCA Civ 703 [40].

[71] Hickman (n 68) 339 notes the example of R *(JL) v Islington LBC* [2009] EWHC 458 (Admin).

[72] Hickman (n 68).

authorities 'to demonstrate that all equality issues have been considered and to close off all arguments [about non-compliance]'.[73] This high evidential burden has effects both at the decision-making stage and at the litigation stage.

At the point of decision-making, the duty incentivises public authorities to conduct investigations about discriminatory effects of AIRAS and record their findings. Indeed, Aikens LJ noted in *R (Brown) v Secretary of State for Work and Pensions* that it was good practice to keep 'adequate record' showing that the PSED had been 'actually considered'.[74] Moreover, the PSED incentivises public authorities to avoid systems with high levels of intrinsic or illiterate opacity, as they will be unable to demonstrate compliance where they are themselves unable to understand an AIRAS' equality impacts. It has also been suggested by practitioners that courts may be amenable to an argument that public bodies cannot comply with their PSED unless they disclose details of automated systems for an independent assessment.[75]

Moreover, the evidence provided by the public authority at trial in the course of PSED claims aids in creating transparency about a given system, regardless of the ultimate success of the claim. This will not only incentivise decision-makers to eliminate discrimination *ex ante*, but may act as a valuable tool to assist litigants in gathering evidence of discrimination to support subsequent or concurrent judicial review challenges on other grounds such as illegality (under EA 2010 section 29 or the Human Rights Act 1998), or under common law grounds including irrationality and the doctrine of relevant considerations. This may be the case even where permission is refused, as

[73] ibid 340-341.

[74] [2008] EWHC 3158 (Admin) [96].

[75] See comments by Megan Goulding in Irving (n 15).

evidenced by *R (Fawcett Society) v Chancellor of the Exchequer*,[76] in which permission to apply for judicial review of the national budget was refused only after a 'substantial hearing and… detailed judgment'.[77]

## *II. The Decision in Bridges*

The potential of the PSED in combating algorithmic opacity may be illustrated by the decision of the Court of Appeal in *R (Bridges) v Chief Constable of South Wales Police*.[78] The case concerned a challenge to the South Wales Police's (SWP's) use of automated facial recognition (AFR) technology on the grounds that, *inter alia*, the SWP had failed to comply with the PSED as it had not considered the possibility that AFR was indirectly discriminatory in producing a higher rate of false positives for women and Black and minority ethnic people.

The Divisional Court found that there was no violation since there was 'no firm evidence' of discriminatory results[79] and 'no specific reason'[80] for the SWP to believe there might be discriminatory effects when it commenced use of AFR. Strikingly, the Court reached this conclusion despite the statement of the claimant's expert witness, who noted that bias was a common feature of AFR systems[81] but explained he could reach no definite conclusion on SWP's system because he did not have access to the training data—the claimant's request for access was refused on the basis that the data was a trade secret[82]—and that the

[76] [2010] EWHC 3522 (Admin).

[77] Hickman (n 68) 333.

[78] [2020] EWCA Civ 1058.

[79] *R (Bridges) v Chief Constable of South Wales Police* [2019] EWHC 2341 (Admin) [153].

[80] ibid [157].

[81] ibid [155].

[82] See Irving (n 15).

defendant, who also had no access, could not make the evaluation either.

However, the Court of Appeal took a more expansive view of the PSED as mandating a public authority to mitigate algorithmic opacity rather than being itself impeded by it. Allowing the appeal, the Court of Appeal found that the Divisional Court's approach had 'put the cart before the horse', since 'the whole purpose of the positive duty… is to ensure that a public authority does not inadvertently overlook information which it should take into account'.[83] Though the Court conceded that commercial confidentiality 'may be understandable' it could not 'enable a public authority to discharge its own, non-delegable, duty under section 149'.[84] In the Court's assessment, the SWP should have 'sought to satisfy themselves, either directly or by way of independent verification, that the software program in this case does not have an unacceptable bias on grounds of race or sex'.[85] The Court also explicitly recognised the importance of combating opacity for *both* the procedural PSED and the substantive prohibitions against direct and indirect discrimination under the EA 2010.[86]

The Court of Appeal's decision offers some guidance to public authorities on how they may comply with the PSED even where systems are opaque, though it remains to be seen what arrangements with third parties courts will consider sufficient to constitute 'independent verification'. What appears certain, however, is that reliance on a developer's bare assurance that a system is not biased will not meet the threshold. It is hoped that

[83] *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058 [182].
[84] ibid [199].
[85] ibid [199].
[86] ibid [200].

the Court of Appeal's approach will be applied appropriately by first instance courts considering AIRAS, as well as other automated decision-making and decision-assistance systems, and that future appellate decisions will continue to utilise the PSED to facilitate transparency and accountability.

## *Conclusion*

Despite the relative novelty of AIRAS and the challenges of opacity, it is evident upon closer examination that the existing framework of the EA 2010 may prove to be an unexpectedly effective avenue for challenging discriminatory systems. This may extend to influencing government decision-making before litigation takes place, as is evident from the Home Office's discontinuation of the use of its rules-based visa 'streaming' tool after the The Joint Council for the Welfare of Immigrants and the charity Foxglove obtained permission to apply for judicial review of its use.[87] The claimants alternatively submitted that the use of nationality[88] as a criterion in streaming individuals into risk categories[89] was directly discriminatory, that other criteria used

[87] David Bolt, 'An inspection of entry clearance processing operations in Croydon and Istanbul – November 2016-March 2017' (2017) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/631520/An-inspection-of-entry-clearance-processing-operations-in-Croydon-and-Istanbul1.pdf> accessed 5 April 2021.

[88] See EA 2010 s 9(1)(b).

[89] See Rafe Jennings, 'Government Scraps Immigration 'Streaming Tool' before Judicial Review' (*UK Human Rights Blog*, 6 August 2020) < https://ukhumanrightsblog.com/2020/08/06/government-scraps-

were directly or indirectly discriminatory, and that the Home Office had violated its PSED.[90] The Home Office subsequently agreed to pause use of the tool while it was redesigned,[91] an outcome which further evidences the effectiveness of the EA 2010 in rising to the equality challenges posed by novel forms of technology-facilitated bias.

However, this article does not seek to make the argument that piecemeal development of principles through *ex-post* review under the EA 2010, or Government reversals caused by the threat of litigation, is sufficient to counter the challenges of algorithmic bias in AIRAS. This is firstly because the EA 2010 can offer protection only against forms of bias which constitute discrimination under its narrow definition; it does not, for example, prohibit bias on the basis of socio-economic class, and the procedural socio-economic duty in section 1 has not been commenced. Other grounds of review, such as those noted above at Part 3.B.I, also do not cover every instance of algorithmic bias. More importantly, bias in AIRAS is not merely a challenge for administrative law, but also raises important policy questions which should be addressed by democratic decision-making. Decisions need to be made as to the types of decision, if any, that AIRAS are suitable for, the degree of acceptable bias (and the groups against which bias is acceptable), as well as the need for

immigration-streaming-tool-before-judicial-review/> accessed 5 April 2021.

[90] As noted in the letter before claim sent to the Home Office.

[91] Home Office statement quoted in Henry McDonald, 'Home Office to scrap 'racist algorithm' for UK visa claimants' *The Guardian* (London, 4 August 2020) <www.theguardian.com/uk-news/2020/aug/04/home-office-to-scrap-racist-algorithm-for-uk-visa-applicants> accessed 5 April 2021.

independent assessment or monitoring systems.[92] Politically driven reform could also offer a more comprehensive framework to promote *ex-ante* accountability, such as through the creation of a regulatory body or, as proposed by Cartwright in this Journal, a regulatory 'toolkit'.[93]

However, unless the Government demonstrates appetite for reform in this area,[94] judicial review will likely remain the main mechanism for combating discrimination in AIRAS for some time. Accordingly, it is worth continuing to follow developments on how courts will choose to apply established equality law principles to emerging technologies.

[92] For an example of political consideration of bias in AIRAS, see the European Commission's recent proposal for a Regulation introducing a legal framework on artificial intelligence, available at: <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence> accessed 21 April 2021. The draft Regulation subjects certain types of 'high-risk' artificial intelligence systems to special requirements in respect of training data, including requiring 'examination in view of possible biases': article 10(2)(f). 'High-risk' systems include those 'intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending': Annex III.

[93] Benjamin Cartwright, 'Regulating the Robot: A Toolkit for Public Sector Automated Decision-Making' (2021) 10 OUULJ 21.

[94] The Government could choose to take forward reform internally or to commission work or to refer the issue for consideration by an independent body. For example, the Law Commission has recently invited comment on the potential inclusion of a project on automated decision-making in its 14th Programme of Law Reform—see <https://www.lawcom.gov.uk/14th-programme-kite-flying-document/> accessed 5 April 2021.