# 'AI Theory of Justice': Using Rawlsian approaches to better legislate on machine learning in government

GRACE, Jamie <http://orcid.org/0000-0002-8862-0014> and BAMFORD, Roxanne

# 'AI THEORY OF JUSTICE': USING RAWLSIAN APPROACHES TO BETTER LEGISLATE ON MACHINE LEARNING IN GOVERNMENT

JAMIE GRACE* AND ROXANNE BAMFORD**[1]

*Sheffield Hallam University **Tony Blair Institute for Global Change

**Abstract**

Policymaking is increasingly being informed by 'big data' technologies of analytics, machine learning, and artificial intelligence (AI). John Rawls used particular principles of reasoning in his 1971 book *A Theory of Justice* which might help explore known problems of data bias, unfairness, accountability and privacy, in relation to applications of machine learning and AI in government. This paper will investigate how the current assortment of UK governmental policy and regulatory developments around AI in the public sector could be said to meet, or not meet, these Rawlsian principles, and what we might do better by incorporating them when we respond legislatively to this ongoing challenge. This paper uses a case study of data analytics and machine learning regulation as the central means of this exploration of Rawlsian thinking in relation to the re-development of algorithmic governance.

**Key words**

Data, algorithms, machine learning, fairness, Rawls, justice, privacy, bias, transparency, accountability, data protection, human rights

## [A] INTRODUCTION

The difficulty in regulating 'algorithmic justice' according to clear human rights standards forms the issue under discussion in this paper. It uses the legal and moral

[1] The authors would like to thank Ben Archer and Dr. Collette Barry at the Department of Law and Criminology, Sheffield Hallam University, for their advice on an early draft, and colleagues at the Tony Blair Institute for Global Change for their support for this work.

philosophy of John Rawls to re-investigate the need for a purposive approach to regulating algorithmically-assisted decision-making in government, and to regulate that 'algorithmic governance' according to certain Rawlsian principles with regard to equality, liberty and distributive justice. Jennifer Cobbe, amongst a wide range of authors, has recently highlighted that '[m]achine learning systems are known to have various issues relating to bias, unfairness, and discrimination in outputs and decisions, as well as to transparency, explainability, and accountability in terms of oversight, and to data protection, privacy, and other human rights issues', but also that 'the processes and metrics for fair, accountable, and transparent machine learning developed through … research do not always translate easily to legal frameworks' (Cobbe 2018: 4-5). We argue that Rawlsian principles can guide this process of marrying data science approaches to fairness for machine learning and AI to the development of new legal frameworks.

In the words of Alistair Duff (2006: 17), 'The ideas of philosopher John Rawls should be appropriated for the information age.' John Rawls, in his 1971 book *A Theory of Justice*, set out the idea that from behind a 'veil of ignorance', in an 'original position', a human policy maker with no conception of the disparities and inequalities in power, wealth or privilege that come about through the realities of class, race and geopolitics, would contract with other policymakers, also in a similarly ignorant position, to ensure a system of fair and liberal rules to benefit all (Rawls 1999: 11). Duff (2006: 21) argues that for 'neo-Rawlsians, therefore, the response to the digital divide, as to any other inequality, will be to regulate social and economic institutions, including information institutions, so that differentials demonstrably work for the good of all, and especially the worst off.'

Rawls used two principles of reasoning to set out and encapsulate this theory of justice. In 'The Original Position', an essay by Ronald Dworkin, Rawls' critic explained this pair of principles. Firstly, 'every person must have the largest political liberty compatible with a like liberty for all' (Dworkin 1975: 17). First we should note that inequality and discrimination are not new issues brought about by AI. They occur all the time, whenever we are not in the original position. And machines, like humans discriminate. Of course, we might accept that not everybody will be subject to governance or AI governance equally. But decisions should be easy to scrutinize. For Rawls liberty, and equality of challenge, is a public good that should be available

to all. This results in an imperative to create the accessible avenues required for scrutiny, and to enable civilians to challenge those that govern them. Everybody would like to think that if they were unfairly, in their view, 'profiled' by a human or by AI, then it would be easy to challenge the resulting decision. Even if AI was only being used as a tool to advise the subsequent decision of a human it should be easy to understand the steps the AI has taken to reach its output. Members of the public must be able to hold those that govern them to account, this includes the algorithms informing their decisions.

Additionally, Dworkin explains, Rawls develops a second principle that 'inequalities in power, wealth, income and other resources must not exist except in so far as they work to the absolute benefit of the worst-off members of society' (Dworkin 197: 17). This second principle translates into an imperative that 'big data' technologies used to assist decision-making must be used in such a way that they do not re-entrench inequality in power, wealth, income and other resources i.e. that they work to the absolute benefit of the worst-off members of society.

Policymaking is increasingly being informed by 'big data' technologies of analytics, machine learning, and AI. But the application of data science through a general legal framework on data protection (which in the UK differentiates mainly along the lines of law enforcement versus non-law enforcement uses of data), non-binding professional codes of ethics and a body of human rights law that is catching up with the developing practice of data-informed governance. To deliver a sense of the variety and scope of the challenge of regulating the use of data science in government, in its next two sections this paper presents a case study highlighting the issues with 'algorithmic justice' in policing contexts. First it is appropriate to give an overview of the common problems of 'algorithmic justice in government'.

Grace (2019) has previously tried to develop a theoretical account of how the use of machine learning and AI within government, in both policymaking and in the application of policy, could raise concerns over 'algorithmic impropriety'. As Grace (2019) has highlighted, strands of algorithmic impropriety can include 'decisional opacity', leading to an inability to effectively challenge the results of algorithmic justice; 'data inequality', resulting in the embeddedness of inequalities, and arising

from unfairly skewed data sets; and 'accuracy bias', resulting from a risk-averse, and predominantly public protection-oriented approach to defining accuracy in predictions and algorithmic profiling. This piece now looks at these issues using a case study of data analytics in policing, drawing on an approach taken by other studies - notably the ground-breaking piece by Selbst (2017).

## [B] AN OVERVIEW OF THE PROBLEM—AND A CASE STUDY OF DATA ANALYTICS IN POLICING

People will not experience justice evenly, and algorithmic justice is no exception. There is a risk that algorithms entrench existing inequalities. Those who are more reliant on state welfare hand-outs, or who are the object of criminal investigations, are a cost that will be, increasingly over time, algorithmically ranked and assessed for risks posed to the public purse, or to public protection. A Rawlsian approach would demand a high degree of information in the public domain, enabling individuals to challenge decisions on a range of grounds. Assessing a system from the 'original position', requires that citizens be well equipped with the knowledge needed to take on the state if they felt they were subject to informational discrimination. A lack of transparency over the algorithms used to govern us is an innate threat to our equal system of liberties for all.

We can see a recent (and so far rare) example of a successful challenge to a lack of transparency in algorithmic justice in the Systemic Risk Indication (SyRI) judgment from the first instance Hague Divisional Court in the Netherlands. In *Netherlands Committee of Jurists for Human Rights v State of the Netherlands* (2020) there were findings on transparency failures in relation to an algorithmically-assisted benefit fraud prediction tool. Given the importance of proportionality in interferences with the right to respect for private and family life, and the requirement of a 'fair balance' between that right and the public interest in the investigation of benefit fraud, there were problematic shortfalls in transparency over the extent to which members of the public subject to risk reports under the SyRI process were aware of this, or could challenge their profiling as likely fraudsters or otherwise.

In the SyRI judgment, the Hague Divisional Court found (para. 6.49) that the Netherlands authorities had 'not made public the risk model and the indicators that make up the risk model', or 'any objectively verifiable information to the court to enable her to test the State's view of what SyRI is', noting that this less than transparent approach was 'a conscious choice by the State…'. The Hague Divisional Court was dismissive of the state defence that if there were more transparency over the algorithm then citizens could adjust their behaviour accordingly. In terms of the detailed issues over transparency shortcomings, the Hague Divisional Court observed (at para. 6.90) that:

> it is not possible to check how the simple decision tree, which the State speaks about, is created and which steps it consists of. It is thus difficult to see how a data subject can defend himself against the fact that a risk report has been made with regard to him or her. Likewise, it is difficult to see how a data subject whose data has been processed in SyRI but has not led to a risk report can be aware that his or her data has been processed on appropriate grounds. The fact that in the latter situation the data did not lead to a risk notification and, moreover, must have been destroyed no later than four weeks after analysis does not detract from the required transparency with regard to that processing.

SyRI had been based on a piece of legislation which was successfully challenged as non-compatible with the European Convention of Human Rights (ECHR), and the Hague Divisional Court noted (at para 6.54) that

> SyRI law does not provide for an information obligation of those whose data are processed in SyRI so that those involved can reasonably be considered to know that his or her data is or has been used for that processing. Nor does the SyRI legislation provide for an obligation to inform data subjects separately, where appropriate, of the fact that a risk report has been made. There is only a legal obligation in advance to announce the start of a SyRI project by publication in the Government Gazette and afterwards on request access to the register of risk reports. The model letter that can be used in practice… is not based on a legal obligation to inform those involved 'house to house', while the court cannot determine on the basis of the available information whether there is a fixed practice [between] municipalities in the implementation of the law. Those involved are also not automatically informed afterwards. This only happens if there is an audit and investigation in response to a risk report. This is not simply done.

The Hague Divisional Court also picked up on the point that greater transparency in relation to predictive modelling of a profiling system is crucial for those who would be aware of the need to challenge biases and system unfairness or discrimination in a system. As the Court observed (at para. 6.91):

> The importance of transparency, with a view to controllability, is important in part because the use of the risk model and the analysis that is carried out in this context involves the risk that (unintentionally) discriminatory effects will occur.

The Hague Divisional Court judgment in the SyRI case, above, is one of the first cases brought against state authorities in relation to issues of transparency algorithmic profiling, and the first known to be successful in that regard, and on the basis of Article 8 ECHR. In the UK, there has been a (so far unsuccessful) challenge to the use of live facial recognition in the case of *R (Bridges) v South Wales Police* (2019), and a claim for judicial review, as yet unheard by the High Court, brought by the data rights advocacy NGO known as Foxglove, in relation to a visa decision algorithm used by the Home Office (McDonald 2019).

Policing in the UK is prone to complex multi-faceted regulation on any issue, with an interplay in policy terms at all times between the Home Office, the National Police Chiefs' Council, the College of Policing, Her Majesty's Inspectorate of Constabulary, Fire and Rescue Services, the National Crime Agency, and any one, through to all, of the nearly 50 regional or specialist police forces in the UK. The UK police service has a range of explicit and implicit statutory powers and obligations (but not specific statutory basis to use algorithmic or machine learning approaches for intelligence analysis) and a range of common law powers around information retention, analysis and intelligence sharing. In the UK, the European Convention on Human Rights (ECHR) increasingly informs police leadership and occupational culture, and the training of decision makers in senior operational roles (Poolman et al 2019). The UK police service should also develop, pilot and deploy AI tech and data science expertise, whether in-house or through contractors by following the Defence Contract Management Agency (DCMA) Code of Ethics on AI, while there is also a draft code for AI procurement published by the UK Office for AI. The Committee on Standards in Public Life have had their say in a report on AI and standards in public life (2020), discussed below, and there has been a report of a Parliamentary Committee on AI technology implications for civil society in the UK (Lords Select Committee, 2017). Furthermore, the Information Commissioner's Office has published its own consultation on a draft AI auditing framework (2020). The UK Centre for Data Ethics and Innovation is also to undertake a public consultation on a code of practice for policing in the UK with regard to the use of data analytics and

machine learning (Macdonald 2020), following reports from the Royal United Services Institute (Babuta and Oswald 2020) on concerns around bias in predictive policing and other data-led approaches.

In the midst of this regulatory complexity, at the time of writing, many forces within the UK police service use a self-regulation framework in relation to machine learning and data analytics, aimed at police forces that are adopting greater data science approaches in their intelligence analysis processes. Known as 'ALGO-CARE', this regulatory framework is a checklist of key considerations in legal, ethical and data science best practice, to be used by police forces in their innovation and adoption of capabilities around data analytics and machine learning applications.

ALGO-CARE requires police forces to use predictive analytics in an advisory (not determinative) way, with control over their intellectual property in the algorithm concerned, and in a way that is lawful; granular; challengeable; accurate; responsible and explainable. The research developing ALGO-CARE was a co-authored evaluation of the legalities of the 'Harm Assessment Risk Tool' (HART), used currently by Durham Constabulary (Oswald & Ors 2018). The HART tool is a leading application of machine learning technology as used in intelligence analysis and risk management practices by police in the UK. HART was the first such police machine learning project in the UK to be open to early academic scrutiny; and as a result was the first which has led to the development of a model regulatory framework, in the form of ALG-CARE, for algorithmic decision-making in policing.

The National Police Chiefs' Council (NPCC) took the decision in November 2018 to promote the use of ALGO-CARE as a model for best practice in the self-regulation by UK police forces of their development of machine learning/algorithmic tools (Grace 2020). In the summer of 2019, it was confirmed by the NPCC that West Midlands Police (WMP) were incorporating the ALGO-CARE checklist or framework in internal development processes in relation to new intelligence analysis tools. West Midlands Police now host the National Data Analytics Solution (NDAS) for the UK police service as a whole. ALGO-CARE is built into the project initiation process for NDAS, and has been used to provide ethical oversight for data analytics projects concerning identifying risks factors around vulnerability to modern slavery, and the perpetration of knife crime (West Midlands Police and Crime Commissioner, 2020b).

Importantly, Essex Police have also drawn on the ALGO-CARE framework in setting up the oversight processes for their data analytics partnership with Essex County Council (Essex Centre for Data Analytics 2019). This adoption of self-regulation is proof of a respect for the professional ethics in the use of machine learning and data analytics in policing.

However, police force ethics committees might never feel they know enough, as outsiders to policing, about exactly what 'interventions' predictive modelling will underpin, and whether these will exacerbate inequalities of opportunity, and unequal interferences with liberties and rights. For example, in April 2019 the independent ethics committee for data analytics for the Office of the Police and Crime Commissioner for the West Midlands ('the committee'), of whom the first author is vice-Chair, at the time of writing, were asked to consider ethical approval for an 'Integrated Offender Management' (IOM) data analysis tool. The terms of reference of the committee put to the fore their scrutiny of the human rights impacts of algorithmic tools, built either by the West Midlands Police (WMP) Data Lab, or the National Data Analytics Solution (NDAS), based at WMP. However, a fundamental question was even more basic than questions of balancing human rights concerns: what was the real purpose, and what would be the estimated impact of the use, of the IOM tool? Offender managers are already experienced in risk scoring offenders under their supervision, and the minutes of the committee meeting from April 2019 reveal that the aim of the IOM tool was to allow for a data-driven means of doing this in a far more rigorous and reliable way, with the IOM tool forming an advisory profiling tool, in time, for those officers 'providing supportive interventions to those considered to be at high risk of reoffending and transitioning to higher harm crimes' (West Midlands Police and Crime Commissioner 2020b), but the committee had initial questions about what these interventions might be, not to mention concerns about the extent to which the tool, in its development iteration at the time, might be 'trained' on stale data stretching back many years, or which was riddled with disproportionality in relation to stigmatised demographic groups. In time, the WMP Data Lab addressed these issues in an informative dialogue with the committee. At the time of writing, a pilot of the IOM tool has only just been started in two small areas of the area covered by West Midlands Police, and plans are in place to begin public engagement over the use of the IOM tool with offender data.

The IOM tool is a *predictive* model, and a running concern of the West Midlands committee is the extent to which initially explanatory models developed out of large datasets might unintentionally become predictive in the way they might influence officers' investigative behaviour. For example, the WMP Data Lab had developed an explanatory analysis of rape and serious sexual offences investigations ('the RASSO project'). The RASSO project identified that, based on fairly recent WMP data, bar the time spent by a lead investigator working on a case, the biggest single factor on a rape investigation being progressed versus being subject to no further action was the failure to obtain the mobile phone data of rape complainants themselves. There are, most understandably, some distinct privacy concerns around requiring complainants to hand over their mobile phones for data extraction, but the process of disclosure of potentially exculpatory evidence to the defence is something that is mandated by an Act of Parliament, in the form of the provisions of Section 3 of the Criminal Procedure and Investigations Act 1996. The committee sought assurances and commitments from the force as to how this finding would be acted on by WMP before it could advise that the RASSO project could progress to its next pilot phase (West Midlands Police and Crime Commissioner 2020b).

In short, there are a wide range of algorithmic justice techniques, and we cannot possible be as comfortable with them all at once, when some of them raise more questions for the rights of victims of crime, or when some of them might mean more of a risk of stigmatising a community than other tools.

## [C] RAWLSIAN APPROACHES TO REGULATION: APPLYING *A THEORY OF JUSTICE* TO MACHINE LEARNING IN PUBLIC INSTITUTIONS

John Rawls used two principles of reasoning to set out and encapsulate his theory of justice which might help explore these problems. In 1975, Norman Daniels highlighted that the First Principle, 'which has priority over the Second, guarantees a

maximal system of equal basic liberties', while the Second Principle 'distributes all social goods, other than liberty, allowing inequalities in them provided they benefit the least advantaged and provided equality of opportunity is present' (Daniels, 1975: xxvii).

*Rawls' first principle*

Rawls' first principle (Rawls 1999: 53) reads: 'Each person is to have an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others.'

As discussed in the introduction, in a real world scenario we might accept that not everybody will be subject to AI governance in equal measure or to an equal extent. But everybody would like to think that if they were unfairly, in their view, profiled by AI then it would be easy to challenge that self-same scoring/ranking/risk prediction. The Rawlsian view is that this liberty that should be available to all, resulting in an imperative to create accessible avenues to enable challenge, and to assist civilians to challenge the decisions of those that use ‘big data’ technologies to govern them.

Holding to account those that govern us requires transparency. It is essential that we understand how decisions affecting the most important aspects of our lives have been arrived at. Rawls himself wrote (1999: 49) that 'in a well-ordered society, one effectively regulated by a shared conception of justice, there is also a public understanding as to what is just and unjust.' Much more recently, David Spiegelhatler has observed that there is 'increasing demand for accountability of algorithms that affect people's lives' (Spiegelhalter 2020: 181), since 'if we do not know how an algorithm is producing its answer, we cannot investigate it for implicit but systematic biases against some members of the community…' (Spiegelhalter 2020: 177).

There is a lack of transparency and understanding of how many algorithmic decisions support tools work. The process by which the calculation is made must be accessible to humans and open to challenge. However, many algorithmic systems, particularly machines learning tools, produce predicted outcomes without being able to show how those predictions have been arrived at. It is this prevalent lack of

auditability that led the UK House of Lords Select Committee on Artificial Intelligence to conclude that (2018: 40):

> it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take.

A failure of proper accountability of algorithmic decision-making to individuals threatens the first Rawlsian principle that we all have a set of basic liberties afforded to us all. As the use of algorithms to inform critical and sometimes life changing decisions becomes more prevalent in our criminal justice system and in other public services, the issue of access to justice is fast becoming a problem of access to *algorithmic* justice.

The General Data Protection Regulation and the Law Enforcement Directive (both now part of 'retained EU law' in the UK as a result of the Brexit process) provide some safeguards against fully-automated decision-making using machine learning, algorithms or AI. Greater signposting is likely to be required, however, where these technologies are used in government in fully-automated ways. There is then the crucial issue of the increasingly large degree to which decisions by public bodies - about policy, but often about individuals in particular personal circumstances, are algorithmically-informed decisions. Here, both greater statutory clarity as to rights of challenge, and greater safeguards involving transparency are required. Challenge requires transparency. Transparency is severely limited in what it can achieve if there are no mechanisms for challenge.

*Rawls' second principle*

Rawls' second principle (Rawls 1999: 53) demands that: 'Social and economic inequalities are to be arranged so that they are both (a) reasonably expected to be to everyone's advantage and (b) attached to positions and offices open to all.' This second principle translates into an imperative that 'big data' technologies used to assist decision-making must be used in such a way that they do not re-entrench inequality in power, wealth, income and other resources i.e. that they work to the overall benefit of the worst-off members of society.

Rawls' notion that 'social and economic inequalities are to be arranged so that they are to the greatest benefit of the least advantaged' is termed his 'difference principle'. Duff (2006: 21) explains that:

> The difference principle … is regarded as Rawls's special contribution to the repertoire of principles of distributive justice in the western tradition. Its genius lies in its balancing of two powerful moral intuitions: that equal shares are fair, at least as an initial benchmark; but also that inequalities can be acceptable if the incentives they allow lead to a greater total cake, thus benefiting everyone, including the worst off. For who wants an equality of misery?

Our moral intuition, to use Duff's phrase, concerning the difference principle in the context of algorithmic justice, is that AI and machine learning can be based on 'training data' which is either known or suspected to be biased, as long as this is a) acknowledged and mitigated when the AI or machine learning tool is developed, and b) such a tool is meaningfully used to redress inequalities, not re-embed them, lest there be an inherent unlawfulness in its use. In essence, the stated and true purpose of algorithmic justice must be more equal justice, or algorithmic justice must be avoided altogether. This approach to applying the difference principle to matters of algorithmic justice would need to be based in primary legislation, in a development of something like the Public Sector Equality Duty (PSED) which already exists in the UK under the provisions of the Equality Act 2010.

*Using the PSED and protected characteristics as a Rawlsian structuring tool*

The PSED is a statutory requirement to be 'properly informed', of the equality implications of decisions made in the course of carrying out public functions, following Elias LJ in *R (Hurley and Moore) v Secretary of State for Business, Innovation and Skills* [2012] EWHC 201 (Admin), at 89. There are equality implications in relation to the impact on protected characteristics, including age, disability, and so forth[2]. Section 149 Equality Act 2010 provides that:

> (1) A public authority must, in the exercise of its functions, have due regard to the need to—
>
> > (a) eliminate discrimination, harassment, victimisation and any other conduct that is prohibited by or under this Act;

[2] 'Protected characteristics': S.149 (7) EA 2010: 'age; disability; gender reassignment; pregnancy and maternity; race; religion or belief; sex; sexual orientation'...

> (b) advance equality of opportunity between persons who share a relevant protected characteristic and persons who do not share it;
>
> (c) foster good relations between persons who share a relevant protected characteristic and persons who do not share it.

The PSED has required a degree of proactivity and culture change in the work of government bodies in the UK, as well as a clear commitment to gather information and to undertake consultation that would inform their work on preventing discrimination. In the words of Lord Boyd, in the recent case of *R (McHattie) v South Ayrshire Council* (2020: 31): 'The duties in the Equality Act 2010 and specifically section 149 are not simply about the prevention of discrimination but the promotion of policies which will help eliminate differences between the protected group and those who do not share that protection.'

The Metropolitan Police have published a document (Metropolitan Police Service 2020) that sets out what they term their legal mandated for Live Facial Recognition (LFR), and this acknowledges the need for compliance with the PSED in deploying the controversial technology in public spaces, but has few details as to how decision-makers in this regard would ensure they were 'properly informed' about the equality issues inherent in deploying LFR in various areas of London with a differing prevalence of people of different ethnicities, for example - when there is a consistent and important concern about poorer accuracy rates for facial recognition technology with regard to the real-time identification of non-white persons (Harwell 2019).

The Met have an interesting pair of issues in their own recently published *guidance* document on live facial recognition technology (as opposed to their purported 'legal mandate' document). First they seem to make a policy commitment to public notification prior to deployments; second they committed to only deploy the technology overtly. These are two commitments to be applauded, from the perspective of Rawls' first principle concerning an equality of liberties for all. But the Met say their watch lists of suspect photographs used in the deployment of LFR are not marked with data about ethnicity, meaning that accuracy rates for 'hits' or 'flags' in each LFR will be harder to determine. This seems to undermine PSED compliance, in either the spirit or the letter of the law. The Met claim it is because they should only process ethnicity data when strictly necessary for policing purposes,

and that this is not a strictly necessary purpose under the terms of Part 3 of the Data Protection Act 2018. But this disregards the self-monitoring the PSED requires. The PSED is a statutory duty, just as the requirement for minimal data processing under the Data Protection Act 2018 is a statutory duty. Perhaps in an evaluation of LFR deployments, 'hits' or matches by ethnicity can be added back in to the watch list data - but if this is the case, the Met's claim that they need to remove ethnicity data from watch lists seems pointless. Efforts to engage with the public over LFR must be more genuine than this sort of dry, data protection-driven detailing in response to valid concerns (Yesburg & Ors 2020).

The operational guidance from the Met concerning LFR would be more reassuring for public confidence on the issue of bias if there was a clearer commitment and explanation as to the overall purpose of the use of LFR by the police in London in meeting their duties under the PSED, and actually reducing bias in street-level policing over time. The force falsely claimed in an equalities impact assessment that the use of the technology was supported by the UK Biometrics Commissioner under current governance arrangement, risking public confidence in the integrity of their use of LFR (Gayle 2020).

The Information Commissioner's Office picked up on a key issue of intersectionality - in this case an increased impact on the protected race and age, and a likely breach of the PSED - when it issued an Enforcement Notice under the Data Protection Act in relation to the Gangs Matrix operated by the Metropolitan Police (ICO 2018a). The Gangs Matrix had not been used in a way that was sufficiently transparent or open to challenge by the disproportionately young black men and teenage males that it 'scored' for gang connections in the London area (MOPAC 2018). The ICO has produced a checklist for compliance for police forces using gang intelligence databases (ICO 2018b) but arguably the best confirmation of the impact of greater scrutiny arising from the Enforcement Notice against this algorithmic (in)justice came when the Mayor's Office for Policing and Crime (MOPAC) in London purported to overhaul the workings of the Gangs Matrix (MOPAC 2020).

Gangs of any type can be statistically modelled as networks with a numbers of nodes representing suspects (or 'nominals', in police intelligence parlance),

victims and witnesses to reported crimes. The WMP Data Lab plan to use the measure of 'network centrality' when building algorithmic models that explain the links between organised crime groups and the individuals that make them up - a tool to be used to better target police operations and investigations aimed at disrupting serious organised crime (West Midlands Police and Crime Commissioner 2020c).The problem with using 'network centrality' in predictive or explanatory modelling is the potential for bias when this 'centrality' is calculated from police intelligence.

The data from police intelligence is always going to be subjective and prone to human bias, especially when a 'network' is partly or wholly a proxy for, or situated within, a demographic group affected by societal inequalities. Humans all behave according to 'assortativity', a tendency to be aligned with or attracted to people who are like ourselves[3]. This is a great underlying influence on calculating an individual's 'eigenvector centrality' - or influence in a network (University of Chieti-Pescara, 2020 online). So if society does not allow for much mobility and is not really progressive, the disadvantaged will form denser 'network nodes' (read: closer human relationships) with other disadvantaged people.

Issues around the uneven spread of poverty in society, on a geographical basis, is a real problem for the use of analytics from the perspective of Rawls' Second Principle. When the WMP Data Lab seek to use postcode data linked to individuals ('nominals') as a reasonable proxy, even though they acknowledge this is not ideal, we see an example of the police building an explanatory model using an analytical approach they know to be biased against individuals who reside in poorer areas of cities or towns (West Midlands Police and Crime Commissioner 2020d). It must be acknowledged that this model is statistically valid as a matter of data science. It is an explanatory model which is not designed to make predictions about individuals and target interventions, though it is built from masses of data about many individual cases. However, the data science considerations are separate from the questions this model raises for in relation to its implications for operational policing and thus Rawls' philosophy.

[3] For an overview of assortativity and other network theory principles please see Niall Ferguson, *The square and the tower: networks, hierarchies, and the struggle for global power*, 2017, Penguin, pp.24-29.

A worrying concern, is how this explanatory model, when used in relation to young people at risk of being drawn into serious violence, will be interpreted by officers. There is a potential that it is misapplied by officers who allow the results of the 'Youth MSV' project to confirm their own assumptions about the affluence or poverty in the places where people live, and the effect this relative deprivation has on them. For this reason, it is to be welcomed that the independent ethics oversight committee for WMP have required that public consultation over the use of the 'Integrated Offender Management' (IOM) tool in development by the Data Lab at the force be augmented by a qualitative evaluation of how the piloting of the tool saw changes in the way that officers worked with offenders and how this related to their decisions on 'interventions' (West Midlands Police and Crime Commissioner 2020b). As Couldechova and Roth have observed (2018: 5), when 'dealing with socio-technical systems, it is also important to understand how algorithms dynamically effect their environment, and the incentives of human actors.'

The impact of the PSED, and the Equality Act 2010, on the regulation of the use of algorithmic governance would be extended even further, in order to meet Rawls' Second Principle, if S.1 of the Equality Act 2010 were brought into effect. Current enacted but not in force in England and Wales, this provision of the 2010 Act would require police forces, and many other public bodies, to have 'due regard' to the need for decisions 'designed to reduce the inequalities of outcome which result from socio-economic disadvantage'. However, as currently enacted this 'due regard' duty for socio-economic disadvantage would apply only to 'strategic decisions' as opposed to all the 'public functions' of a force. A natural step, and one in line with Rawls' Second Principle, would be to make 'relative poverty' or 'low income' a protected characteristic under S.149(7) of the 2010 Act.

# [D] DISCUSSION: HOW CAN WE INCORPORATE RAWLS INTO THE USE OF AI IN PUBLIC INSTITUTIONS?

Increasingly algorithms play a vital role in our lives. In relation to the use of big data technologies in our public institutions there are many areas which require effective

oversight and clearer laws and guidance in order to conform to Rawls' philosophy. The need for greater transparency and accountability is highlighted in a recent report on the use of *Algorithms in the Criminal Justice System* produced by the Law Society, along with issues of privacy, fairness and equality (Law Society 2020).

In the recent case of *Gaughran v UK* (2020), the European Court of Human Rights made an interesting comment on Article 8 ECHR and technology, noting (86):

> the importance of examining compliance with the principles of Article 8 where the powers vested in the state are obscure, creating a risk of arbitrariness especially where the technology available is continually becoming more sophisticated.

Are the powers of UK police forces to use algorithmic technologies obscure, creating that risk of arbitrary use of continually more sophisticated machine learning or AI? The Committee on Standards in Public Life (CSPL), in their report on *Artificial Intelligence and Public Standards*, published in February 2020, made as one of its key recommendations the creation of a duty on public bodies to clearly articulate their legal basis for the use of algorithmically-informed governance, arguing (CSPL 2020: 40) that: 'All public sector organisations should publish a statement on how their use of AI complies with relevant laws and regulations before they are deployed in public service delivery. ' This degree of transparency would be admirable, as it would entail the creation of a statutory duty through a new act of Parliament to apply to law enforcement agencies and bodies, and public bodies more broadly, alike. The CSPL also concluded that on 'AI' including the use of machine learning for predictive policing and for live facial recognition, the current 'regulatory framework is not yet fit for purpose' (CSPL 2020: 40).

Another report in February 2020, by the Royal United Services Institute for the Centre for Data Ethics and Innovation (Babuta and Oswald 2020: ix), recommended that for UK police forces 'investing in new data analytics software as a full operational capability, an integrated impact assessment should be conducted, to establish a clear legal basis and operational guidelines for use of the tool'. Babuta and Oswald argued for a range of requirements to be placed on UK police forces adopting algorithmic justice approaches and practices, recommending the mandated 'integrated impact assessment'. Their RUSI report calls, overall, for the use of combined data protection impact assessments, equality impact assessments, human

rights impact assessments (with a particular focus on positive obligations in relation to protection of the right to life, and protecting individuals from serious violence or abuse), assessments of expected levels of errors in any predictions made by an algorithmic model, and a requirement for independent ethical oversight mechanisms for data analytics or AI projects in police forces (Babuta and Oswald 2020).

With regard to the notion of ethical oversight as valuable, some academic critics have reminded us of the need to maintain the necessary focus on legal reform so as to not drift into using more flexible and ultimately non-binding ethical standards for regulating algorithmic justice. Black and Murray (2020: 7), for example, explain that:

> The wider discourse that is taking place is drawing us away from law, or even traditional models of command and control or co-regulation and governance, towards soft self-regulation and codes of practice. This ethical model … has seen the adoption of codes of practice for general AI and for data-driven health and care technology, among others. However r… ethical standards for such systemic risks are insufficient.

## [E] CONCLUSIONS

To begin our conclusions on an optimistic point, we would agree with Ori Gilboa (2019 online), who has suggested that: 'AI provides us with the unprecedented opportunity to transform our society into one that is more just.' And while it is important to note, as Kalle Eriksson does, that with regard to increasingly algorithmic governance, the approach of '… 'business as usual' is bound to move us towards increased inequalities and decreased possibilities for most individuals to pursue their conception of the good life,' we also agree with Eriksson that 'there are reasons for hopefulness, since we have also seen that this development could be reversed by making the social choice to own and administer the technology jointly' (Eriksson 2018: 40).

Machines like humans can be flawed. However, it can be easier to identify their flaws and correct for them. While artificial intelligence systems today are often opaque and poorly understood if they can be unpacked to show how the output was reached, as is possible in the HART model, algorithms can increase auditability. This is not the same for a solely human decision-making process which is always going to be opaque to some extent. If people exercise judgement with access to auditable

information provided by an algorithm this could increase transparency, accountability and correct for human bias.

Civil society is certainly beginning to add momentum toward stricter regulation of algorithmic justice matters. After consulting widely, the UK national human rights body, the Equality and Human Rights Commission (EHRC) has submitted (EHRC 2020: 66) to a UN Committee that it has concerns about algorithmic governance in the UK today. The EHRC submitted that: 'predictive policing replicates and magnifies patterns of discrimination in policing, while lending legitimacy to biased processes. A reliance on 'big data' encompassing large amounts of personal information may also infringe upon privacy rights and result in self-censorship, with a consequent chilling effect on freedom of expression and association.' The EHRC would also 'suspend the use of automated facial recognition and predictive programmes in policing, pending completion of the… independent impact assessments and [a public and parliamentary] consultation process, and the adoption of appropriate mitigating action.' (EHRC 2020: 89)

Zuiderveen Borgesius goes a logical step further, arguing for new legislation aimed at tackling new unfairnesses affecting 'newly invented classes', amongst those subjected to bias in algorithmic governance, explaining that

> Non-discrimination law and data protection law are the most relevant legal instruments to fight illegal discrimination by algorithmic systems… But some types of algorithmic decisions evade current laws, while they can lead to unfair differentiation or discrimination. For instance, many non-discrimination statutes only apply to discrimination on the basis of certain protected grounds, such as ethnic origin. Such statutes do not apply if organisations differentiate on the basis of newly invented classes that do not correlate with protected grounds. Such differentiation could still be unfair, however, for instance when it reinforces social inequality. We probably need additional regulation to protect fairness and human rights in the area of algorithmic decision-making (Zuiderveen Borgesius 2020: 15).

In relation to the PSED, the Law Society has recommended (2020: 7) that with respect to the growing use of algorithmic governance in the criminal justice system, and the 'importance of countering discrimination within algorithmic systems, Equality Impact Assessments should be formalised as a requirement before deploying any consequential algorithmic system in the public sector and these should be made

proactively, publicly available.' The Law Society also recommended (2020: 7) that given 'algorithmic systems' high potential for socioeconomic discrimination, the Government should commence the socioeconomic equality duty in the Equality Act 2010 s1 in England and Wales, at least with regard to algorithmic decision-support systems.'

Our overall conclusion is that in order to gain maximum value and help for the vulnerable, and in doing so by applying Rawlsian thinking to the regulation of algorithmic governance in the UK, there needs to be a political commitment to a rolling programme of sector-by-sector legal reform, in order to legislate more deeply for a culture of algorithmic justice. As Zuiderveen Borgesius has also concluded (2020: 15),

> it is probably not useful to adopt rules for algorithmic decision-making in general. Just like we did not, and could not, adopt one statute to regulate the industrial revolution, we cannot adopt one statute to regulate algorithmic decision-making. To mitigate problems caused by the industrial revolution, we needed different laws for work safety, consumer protection, the environment, etc. In different sectors, the risks are different, and different norms and values are at stake. Therefore, new rules for algorithmic decision-making should be sector-specific.

Mechanisms for oversight such as ethics committees and regulators need to be bolstered by the law. At the time of writing, in April 2020, the Committee of Ministers of the Council of Europe have just published a set of Recommendations concerning 'human rights impacts of algorithmic systems' (Council of Europe, 2020a), 'calling on governments to ensure that they do not breach human rights through their own use, development or procurement of algorithmic systems', and explaining that, '…as regulators, [governments] should establish effective and predictable legislative, regulatory and supervisory frameworks that prevent, detect, prohibit and remedy human rights violations, whether stemming from public or private actors' (Council of Europe, 2020b). The preamble of the recent Recommendation demands that 'the rule of law standards that govern public and private relations, such as legality, transparency, predictability, accountability and oversight, must also be maintained in the context of algorithmic systems' (Council of Europe 2020a). This of course accords with Rawls' First Principle. The Recommendation also follows the notion of Rawls' Second Principle, purporting to mandate that Member States of the Council of Europe, like the UK, put data bias, and equality concerns to the fore in developing

legal standards in relation to algorithmic systems used in government. The Recommendation sets out how:

> In the design, development, ongoing deployment and procurement of algorithmic systems for or by them, States should carefully assess what human rights and non-discrimination rules may be affected as a result of the quality of data that are being put into and extracted from an algorithmic system, as these often contain bias and may stand in as a proxy for classifiers such as gender, race, religion, political opinion or social origin. The provenance and possible shortcomings of the dataset, the possibility of its inappropriate or decontextualised use, the negative externalities resulting from these shortcomings and inappropriate uses as well as the environments within which the dataset will be or could possibly be used, should also be assessed carefully (Council of Europe 2020a).

In summary, it remains to be seen how the UK government will choose to combine data protection, equality law approaches and human rights standards in developing new legislation to meet emerging challenges of algorithmic justice in data-driven governance. In our view, laws and guidance for the use of artificial intelligence in our criminal justice system and in other public institutes must ensure that the data, the technology and the process by which the technology is used reflect Rawls' principles. *A Theory of Justice* provides a blueprint for our democracy and it remains highly relevant today as we grapple with the ethics and regulation of 'big data' technologies.

## References


Babuta, Alexander and Marion Oswald (2020) 'Data Analytics and Algorithms in Policing in England and Wales: Towards A New Policy Framework', London: Royal United Services Institute.

Cobbe, Jennifer (2018) 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making'. A pre-review version of a paper in *Legal Studies*, available at: https://ssrn.com/abstract=3226913

CSPL (Committee on Standards in Public Life) (2020), 'Artificial Intelligence and Public Standards'.

Chouldechova, Alexandra and Aaron Roth (2018) 'The frontiers of fairness in machine learning' arXiv preprint, available at https://arxiv.org/abs/1810.08810

Council of Europe (2020a) Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, available at https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154

Council of Europe (2020b) 'Algorithms and automation: new guidelines to prevent human rights breaches', available at https://www.coe.int/en/web/portal/-/algorithms-and-automation-new-guidelines-to-prevent-human-rights-breaches

Daniels, Norman (ed) (1975) *Reading Rawls: Critical Studies of a Theory of Justice* Oxford: Basil Blackwell.

Duff, Alistair (2006) 'Neo-Rawlsian Co-ordinates: Notes on A Theory of Justice for the Information Age', 6 *International Review of Information Ethics* 17-22.

Dworkin, Ronald (1975) 'The Original Position', in N. Daniels (ed) (1975) *Reading Rawls: Critical Studies of a Theory of Justice* Oxford: Basil Blackwell.

ECDA (Essex Centre for Data Analytics) (2020) 'Be part of the conversation: the ethics of data analytics' Essex Partnership https://www.essexfuture.org.uk/news-and-events/join-ecdas-ethics-committee/

ECDA (Essex Centre for Data Analytics) (2019) 'Transparency and trust' Essex Partnership https://www.essexfuture.org.uk/ecda/collaborative-learning/transparency-and-trust/

Gayle, Damien (2020) 'Watchdog rejects Met's claim that he supported facial recognition', *The Guardian* 12 February 2020

Gilboa, Ori (2019) 'Rawls' ghost in the machine' *The Varsity* 7 May 2019

Grace, Jamie (2019) 'Algorithmic impropriety in UK policing?' 3(1) *Journal of Information Rights, Policy and Practice* available at http://doi.org/10.21039/irpandp.v3i1.57

Grace, Jamie (2020) 'Fresh, fair, and smart: data reliability in predictive policing', about:intel, available at https://aboutintel.eu/predictive-policing-data-reliability/

Harwell, Drew (2019) 'Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use', *Washington Post* 19 December 2019

ICO (Information Commissioner's Office) (2018a) *Metropolitan Police Service Gangs Matrix Enforcement Notice*, Wilmslow: Information Commissioner's Office.

ICO (Information Commissioner's Office) (2018b) *Processing gangs information: A checklist for police forces*, Wilmslow: Information Commissioner's Office.

ICO (Information Commissioner's Office) (2020) *ICO consultation on the draft AI auditing framework guidance for organisations*, Wilmslow: Information Commissioner's Office.

The Law Society (2020) *Algorithms in the Criminal Justice System*, London: The Law Society.

Lords Select Committee on Artificial Intelligence (2018) *Artificial Intelligence in the UK: Ready, Willing and Able?*, HL Paper 100

McDonald, Henry 'AI system for granting UK visas is biased, rights groups claim', *The Guardian*, 29 October 2019

Macdonald, Lara (2020) 'What next for police technology and ethics?', Centre for Data Ethics and Innovation https://cdei.blog.gov.uk/2020/02/26/what-next-for-police-technology-and-ethics/

Metropolitan Police Service (2020) *Live Facial Recognition: Legal Mandate*, London: Metropolitan Police Service.

MOPAC (Mayor's Office for Police and Crime) (2018) *Review of the MPS Gangs Matrix*, London: MOPAC.

MOPAC (Mayor's Office for Police and Crime) (2020), 'Mayor's intervention results in overhaul of Met's Gangs Matrix', MOPAC, available at https://www.london.gov.uk//press-releases/mayoral/mayors-intervention-of-met-gangs-matrix

Oswald, Marion & Ors (2018) 'Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality', 27:2 *Information & Communications Technology Law*, 223-250

Rawls, John (1971) *A Theory of Justice*, Boston: Harvard University Press

Rawls, John (1999) *A Theory of Justice: Revised Edition*, Boston: Harvard University Press.

Selbst, Andrew D. (2017) 'Disparate impact in big data policing' 52 *Georgia Law Review* 109-195.

Spiegelhalter, David (2020) *The Art of Stats: Learning from Data*, London: Penguin Random House.

University of Chieti-Pescara (2020) 'Eigenvector Centrality', available at https://www.sci.unich.it/~francesc/teaching/network/eigenvector.html

WMPCC (West Midlands Police and Crime Commissioner) (2020a), 'Information: Ethics Committee', Office of the Police and Crime Commissioner for the West Midlands https://www.westmidlands-pcc.gov.uk/ethics-committee/

WMPCC (West Midlands Police and Crime Commissioner) (2020b), 'Ethics Committee minutes' (January 2020) Birmingham: Office of the Police and Crime Commissioner for the West Midlands

WMPCC (West Midlands Police and Crime Commissioner) (2020c) 'Serious Organised Crime (SOC)' paper (January 2020) Birmingham: Office of the Police and Crime Commissioner for the West Midlands

WMPCC (West Midlands Police and Crime Commissioner) (2020d) 'Youth Most Serious Violence (MSV)' paper (January 2020) Birmingham: Office of the Police and Crime Commissioner for the West Midlands

Yesburg, Julia & Ors (2020) 'Public support for Live Facial Recognition and implications for COVID-19 policing', London School of Economics Politics and Policy Blog, available at https://blogs.lse.ac.uk/politicsandpolicy/covid-19-lfr/

Zuiderveen Borgesius, Frederik J. (2020) 'Strengthening legal protection against discrimination by algorithms and artificial intelligence' *The International Journal of Human Rights*, DOI: 10.1080/13642987.2020.1743976

## Legislation Cited

Criminal Procedure and Investigations Act 1996

Data Protection Act 2018

Equality Act 2010

European Convention on Human Rights (1950)

Human Rights Act 1998

## Cases Cited

*R (Bridges) v South Wales Police* [2019] EWHC 2341 (Admin)

*Gaughran v UK* (2020) (Application no. 45245/15)

*Netherlands Committee of Jurists for Human Rights v State of the Netherlands* (2020) ECLI: NL: RBDHA: 2020: 865

*R (McHattie) v South Ayrshire Council* [2020] CSOH 4