INTERNATIONAL HUMAN RIGHTS LAW AS A FRAMEWORK FOR ALGORITHMIC ACCOUNTABILITY

Lorna McGregor, Daragh Murray and Vivian Ng*

Abstract Existing approaches to 'algorithmic accountability', such as transparency, provide an important baseline, but are insufficient to address the (potential) harm to human rights caused by the use of algorithms in decision-making. In order to effectively address the impact on human rights, we argue that a framework that sets out a shared understanding and means of assessing harm; is capable of dealing with multiple actors and different forms of responsibility; and applies across the full algorithmic life cycle, from conception to deployment, is needed. While generally overlooked in debates on algorithmic accountability, in this article, we suggest that international human rights law already provides this framework. We apply this framework to illustrate the effect it has on the choices to employ algorithms in decision-making in the first place and the safeguards required. While our analysis indicates that in some circumstances, the use of algorithms may be restricted, we argue that these findings are not 'anti-innovation' but rather appropriate checks and balances to ensure that algorithms contribute to society, while safeguarding against risks.

Keywords: human rights, artificial intelligence, big data, algorithms, accountability.

I. INTRODUCTION

Offering greater efficiency, reduced costs, and new insights into current and predicted behaviour or trends,[1] the use of algorithms to make or support decisions is increasingly central to many areas of public and private life.[2]

* Professor of International Human Rights Law, Essex Law School, Director, Human Rights Centre, University of Essex and Principal Investigator and Director, ESRC Human Rights, Big Data and Technology Project, lmcgreg@essex.ac.uk; Senior Lecturer, School of Law & Human Rights Centre, University of Essex, Deputy Work Stream Lead, ESRC Human Rights, Big Data & Technology Project, d.murray@essex.ac.uk; Senior Research Officer, ESRC Human Rights, Big Data & Technology Project, Human Rights Centre, University of Essex, vivian.ng@essex.ac.uk. This work was supported by the Economic and Social Research Council [grant number ES/M010236/1].

[1] L Rainie and J Anderson, 'Code-Dependent: Pros and Cons of the Algorithm Age' (Pew Research Center, February 2017) 30–1 <http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age>; R Kitchin, 'Thinking Critically About and Researching Algorithms' (2017) 20(1) Information, Communication & Society 14, 18–19.

[2] See eg HJ Wilson, A Alter and P Shukla, 'Companies Are Reimagining Business Process with Algorithms' (Harvard Business Review, 8 February 2016) <https://hbr.org/2016/02/companies-are-reimagining-business-processes-with-algorithms>.

[*ICLQ* vol 68, April 2019 pp 309–343] doi:10.1017/S0020589319000046

However, the use of algorithms is not new. An algorithm, as defined by the Oxford English Dictionary, is simply '[a] process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer'.[3] An early example is the use of handwritten algorithms to count votes and determine a winner in the electoral process. What accounts for the increasing centrality of algorithms in contemporary society is their transformational potential. For example, advances in computational power mean that modern algorithms can execute complex tasks beyond human capability and speed, self-learn to improve performance, and conduct sophisticated analysis to predict likely future outcomes. Modern algorithms are fuelled by easily accessible large and/or diverse datasets that can be aggregated and processed efficiently (often labelled 'big data').[4] These algorithms exist in a complex, interdependent, global data ecosystem whereby algorithmically produced outputs can be used as new input data for other algorithmic processes.[5]

The interaction and interdependence of algorithms, including artificial intelligence (AI) or machine-learning algorithms, and big data have enabled their deployment in many key areas of decision-making, such that many functions traditionally carried out by humans have become increasingly automated. For example, algorithms are used to: assist in sentencing and parole decisions; predict crime 'hotspots' to allocate police resources; personalize search engine results, electronic newsfeeds and advertisements; detect fraud; determine credit ratings; facilitate recruitment; and deliver healthcare and legal services. The advent of self-driving cars underscores the speed at which technology is developing to enable more complex autonomous decision-making.[6]

Given the extent of their societal impact, it is perhaps unsurprising that the use of algorithms in decision-making raises a number of human rights concerns. The risk of discrimination arising from the use of algorithms in a wide range of decisions from credit scoring to recidivism models has already been well documented.[7] The range of contexts in which algorithms are used also generates other less studied threats to human rights. For instance, automated credit scoring can affect employment and housing rights; the increasing use of algorithms to inform decisions on access to social security potentially impacts a range of social rights; the use of algorithms to assist with identifying children at risk may impact upon family life; algorithms

[3] Oxford English Dictionary, 'Definition of algorithm' <https://en.oxforddictionaries.com/definition/algorithm>.

[4] For instance, from metadata, smart technology and the Internet of Things.

[5] JM Balkin, '2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data' (2017) 78(5) OhioStLJ 1217, 1219.

[6] Select Committee on Artificial Intelligence, *Corrected Oral Evidence: Artificial Intelligence, Evidence Session No. 1* (HL 2017–2019), 10 October 2017 Evidence Session <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/oral/71355.pdf> 2, 9.

[7] See discussion Part IIA.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

used to approve or reject medical intervention may affect the right to health; while algorithms used in sentencing decisions affect the right to liberty.[8]

In recent years a multi-disciplinary literature has developed on 'algorithmic accountability'.[9] Proposals for achieving better accountability for decisions made or supported by algorithms have focused either on technical solutions, such as blockchain,[10] or modalities for improving the transparency of algorithmic systems, making their decision-making process more understandable and explainable, and creating rules in algorithmic programmes to prevent or detect unfair outcomes.[11] While each of these approaches constitutes a necessary element of accountability, in our view, they are incomplete due to their focus on specific aspects of the overall algorithmic process. Instead, the complex nature of algorithmic decision-making necessitates that accountability proposals be set within a wider framework, addressing the overall algorithmic life cycle, from the conception and design phase, to actual deployment and use of algorithms in decision-making. In light of the diverse range of actors involved, this framework also needs to take into account the rights and responsibilities of all relevant actors.

This article contributes to the literature on algorithmic accountability by proposing an approach based on international human rights law (IHRL) as a means to address the gaps we identify in current proposals for 'algorithmic accountability'.[12] Under IHRL, States are required to put in place a framework that prevents human rights violations from taking place, establishes monitoring and oversight mechanisms as safeguards, holds those responsible to account, and provides a remedy to individuals and groups who claim that their rights have been violated.[13] These obligations apply directly to State actions or omissions and, through the principle of due diligence, the State

[8] See discussion Parts IIIA, IVA and IVB.

[9] See discussion Part IIB; JA Kroll *et al.*, 'Accountable Algorithms' (2017) 165(3) UPaLRev 633; S Barocas, S Hood and M Ziewitz, 'Governing Algorithms: A Provocation Piece' (Governing Algorithms Conference (New York University, 29 March 2013); M Ananny and K Crawford, 'Seeing Without Knowing: Limitations of The Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20(3) New Media & Society 973; DK Citron and F Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89(1) WashLRev 1; T Zarsky, 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making' (2016) 41(1) Science, Technology & Human Values 118; N Diakopoulos, 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures' (2015) 3(3) Digital Journalism 398; S Wachter, B Mittelstadt and C Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31(2) Harvard Journal of Law & Technology 841.

[10] Discussed further in Part IIB.

[11] These approaches tend to focus on computational methods to achieve some form of statistical parity, which is a narrow view of giving effect to the principle of equality. See discussion Part IIB; Kitchin (n 1) 16.

[12] As shorthand in this article we use the abbreviation 'IHRL' to refer to international human rights law and broader norms.

[13] UN Human Rights Committee, 'General Comment No. 31 The Nature of the Legal Obligation Imposed on States Parties to the Covenant' (26 May 2004) UN Doc CCPR/C/21/Rev.1/Add. 13, paras 3–8; UN Committee on Economic Social and Cultural Rights, 'General Comment No. 3

is also required to protect individuals from harm by third parties, including business enterprises.[14] IHRL also establishes an expectation that business enterprises themselves respect human rights, for instance by undertaking ongoing human rights due diligence 'to identify, prevent, mitigate and account for how they address their impact on human rights'.[15]

Some studies have started to emerge that identify the potential impact of AI on human rights.[16] As part of a wider discussion on regulation of the AI sector, some commentators now also propose human rights as an addition or alternative to ethical principles to address some of the (potential) harm posed by the development and use of AI.[17] However, these studies—and existing literature on algorithmic accountability—have not engaged in a detailed examination of whether and how the international human rights law framework might itself offer a response to the overall risks to human rights posed by algorithms. This is problematic as IHRL applies to big data and new technologies just as in any other area of life and, as argued here, offers a framework through which algorithmic accountability can be situated. This article is one of the first to examine 'algorithmic accountability' from the perspective of IHRL

The Nature of States Parties' Obligations (Art. 2, Para. 1, of the Covenant)' (14 December 1990) UN Doc E/1991/23, paras 2–8.

[14] UN Human Rights Council, 'Report of The Special Representative of The Secretary-General on The Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie, on Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework' (21 March 2011) UN Doc A/HRC/17/31, Principles 1–10 [hereinafter *Ruggie Principles*].

[15] ibid, Principle 15.

[16] Council of Europe Committee of Experts on Internet Intermediaries (MSI-NET), 'Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications' (March 2018) Study DGI(2017)12; UN Human Rights Council, 'Report of the Office of the UN High Commissioner for Human Rights on The Right to Privacy in the Digital Age' (3 August 2018) UN Doc A/HRC/39/29, paras 1, 15; F Raso *et al.*, 'Artificial Intelligence & Human Rights: Opportunities & Risks' (Berkman Klein Center for Internet & Society at Harvard University, 25 September 2018); M Latonero, 'Governing Artificial Intelligence: Upholding Human Rights & Dignity' (Data & Society, 10 October 2018); Access Now, 'Human Rights in the Age of Artificial Intelligence' (8 November 2018); P Molnar and L Gill, 'Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System' (University of Toronto International Human Rights Program and The Citizen Lab, September 2018); UN Human Rights Council, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression on A Human Rights Approach to Platform Content Regulation' (6 April 2018) UN Doc A/HRC/38/35; UN Human Rights Council, 'Report of the Independent Expert on the Enjoyment of All Human Rights by Older Persons on Robots and Rights: The Impact of Automation on the Human Rights of Older Persons' (21 July 2017) UN Doc A/HRC/36/48; UN Special Rapporteur on Extreme Poverty and Human Rights Philip Alston, 'Statement on Visit to the United Kingdom' (London, 16 November 2018) <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23881&LangID=E>; Global Future Council on Human Rights 2016–2018, 'White Paper: How to Prevent Discriminatory Outcomes in Machine Learning' (World Economic Forum, March 2018); D Allison-Hope, 'Artificial Intelligence: A Rights-Based Blueprint for Business, Paper 2: Beyond the Technology Industry' (Business for Social Responsibility, August 2018).

[17] See eg C van Veen and C Cath, 'Artificial Intelligence: What's Human Rights Got to Do with It?' (Data & Society, 14 May 2018) <https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5>.

and to detail how human rights can inform the algorithm design, development and deployment process.

This article does not suggest that IHRL offers an exclusive or ready-made, fully developed, solution to the issue of algorithmic accountability. The framework itself has limitations. For example, businesses, particularly large technology companies, are central actors in this area. However, the scope and content of businesses' human rights responsibilities are still in a process of development under IHRL. While States have direct obligations to prevent and protect human rights from third-party harm, including that caused by businesses, the fact that global businesses operate across multiple jurisdictions inevitably gives rise to regulatory and enforcement gaps and inconsistencies.[18] IHRL also only establishes 'expectations' as to how businesses should operate, it does not currently establish direct obligations under international law.[19] Within this context, holding businesses to account for harm caused to human rights and ensuring access to an effective remedy against global businesses, in particular, continues to be a challenge.[20] The IHRL framework also cannot resolve all the challenges related to algorithmic accountability, some of which are addressed by other fields of law such as data protection.

This article does not suggest that IHRL offers a panacea. Rather, our argument is that a human rights-based approach to algorithmic accountability offers an organizing framework for the design, development and deployment of algorithms, and identifies the factors that States and businesses should take into consideration in order to avoid undermining, or violating, human rights. This is a framework which is capable of accommodating other approaches to algorithmic accountability—including technical solutions—and which can grow and be built on as IHRL itself develops, particularly in the field of business and human rights.

[18] UN Human Rights Council, 'Report of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie on Protect, Respect and Remedy: A Framework for Business and Human Rights' (7 April 2008) UN Doc A/HRC/8/5, para 3.

[19] Ruggie Principles (n 14) Principle 11 and accompanying commentary. At the time of writing, the open-ended intergovernmental working group on transnational corporations and other business enterprises with respect to human rights has produced a zero draft of 'a legally binding instrument to regulate, in international human rights law, the activities of transnational corporations and other business enterprises', as mandated by UN Human Rights Council Resolution 26/9. See UN Human Rights Council, 'Legally Binding Instrument to Regulate, In International Human Rights Law, The Activities of Transnational Corporations and Other Business Enterprises' (Zero Draft 16.7.2018).

[20] UN General Assembly, 'Report of the Working Group on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises on Access to Effective Remedies Under the Guiding Principles on Business and Human Rights: Implementing the United Nations Protect, Respect and Remedy Framework' (18 July 2017) A/72/162, para 5; UN Human Rights Council, 'Report of the UN High Commissioner for Human Rights on Improving Accountability and Access to Remedy for Victims of Business-Related Human Rights Abuse' (10 May 2016) A/HRC/32/19, para 2, 4–6.

Some commentators have suggested that the 'vastness, never-ending growth and complexity of algorithmic systems' mean that effective oversight and accountability for their use is not possible.[21] Others have argued that it is too late to develop an effective oversight model, particularly in 'an environment dominated by corporate and government interests'.[22] In our view, space remains to address the existing and potential harm to human rights arising from the use of algorithms in decision-making. However, the pace of technological change and the increasingly prominent and determinative role of algorithms mean that this task is urgent.

Part II examines existing proposals for 'algorithmic accountability'. It does so by first explaining in more detail the nature of algorithms and how they can adversely impact human rights and pose challenges for accountability. The majority of proposals for accountability have focused on addressing the complexity and sophistication of modern algorithms through greater transparency and explainability. We argue that these approaches are necessary but not sufficient to address the overall risks to human rights. Greater focus on the scope and implementation of States' obligations and the expectations placed on businesses in relation to prevention, oversight, accountability, and remedies is needed.

In Part III, we propose that IHRL offers an appropriate framework. It does so by setting out a number of internationally agreed substantive and procedural rights which, if violated, constitute harm. It also provides the means to analyse when the use of algorithms in decision-making could contribute to, or result in, harm, even if unintentionally, and establishes a range of obligations and requirements in relation to the identification of, and protection against, such effects. This framework can apply holistically across the full algorithmic life cycle from conception and design to deployment. By incorporating and building on existing models of accountability it provides a deeper way in which to respond to and protect against risks to human rights.

Part IV analyses the impact of this framework on the use of algorithms in decision-making, reaching three key findings. First, IHRL may rule out the use of algorithms in certain decision-making processes. Second, it may require modifications or the building in of additional safeguards in order to ensure rights compliance and thus may create a delay in deployment. Third, it may shift debates on the unpredictability of algorithms, particularly in the future where greater autonomy is anticipated, from a perceived reduced responsibility to a greater responsibility for actors that deploy algorithms in the knowledge that they cannot predict effects, including to human rights. While these three findings act as restrictions on the use of algorithms, in our view, they constitute appropriate checks and balances. They are not intended to be 'anti-innovation'. Instead algorithmic decision-making is addressed in the same way

[21] See Rainie and Anderson (n 1) 83.

[22] ibid, 83.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

as human decision-making. The objective is to ensure that algorithms contribute to society, while safeguarding against risks.

II. THE NATURE OF ALGORITHMS AND CURRENT ALGORITHMIC ACCOUNTABILITY DEBATES

This part begins by discussing certain characteristics associated with algorithmic decision-making and how these pose challenges when identifying the impact on human rights and for accountability. Existing 'algorithmic accountability' proposals are then examined. Although these proposals constitute a necessary baseline, we identify a number of remaining gaps and challenges.

A. How the Nature of Algorithms Impacts Human Rights

At their simplest, algorithms are formulas designed to calculate a particular result.[23] Today, algorithms are typically understood as either a piece of code or a computer application that can be used to support human decision-making or to take actions independent of human input. There are many different types of algorithms. Relatively straightforward algorithms may be used to perform mathematical calculations to compute an equation; to sort data, which can be useful for finding patterns and connections; or to classify data on the basis of specified criteria. These 'traditional' algorithms run on computer code written by human programmers who understand their logical underpinnings and, if required, can explain how a particular decision was reached by demonstrating the inner workings of the system. However, modern algorithms and the manner in which they are used are becoming increasingly sophisticated.[24]

Modern algorithms are used to support a range of decisions. Some of the most reported examples involve the use of algorithms within decision-making processes that directly affect human rights. The use of algorithmically-produced risk scores in sentencing decisions is one of the most frequently cited examples in this respect,[25] given that the risk score may have a direct bearing on an individual's right to liberty and the prohibition of discrimination. Algorithmic risk assessments are also used in other sectors. For example, an automated algorithmic-based social security system is

[23] The nature of algorithms is presented simplistically here, to encapsulate their essential elements relevant to the present discussion. There are multiple ways of understanding what an algorithm is, its functions, and how it executes those functions. See TH Cormen *et al.*, *Introduction to Algorithms* (3rd edn, MIT Press 2009) 5–10; DE Knuth, *The Art of Computer Programming*, vol 1 (3rd edn, Addison Wesley Longman 1997) 1–9.

[24] T Gillespie, 'The Relevance of Algorithms' in T Gillespie, PJ Boczkowski and KA Foot (eds), *Media Technologies: Essays on Communication, Materiality, and Society* (MIT Press 2014) 167, 192.

[25] The case of *Wisconsin v Eric L. Loomis*, which deals with precisely this issue, is discussed in greater detail in Part IVB.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

currently being implemented in the UK with the aim of streamlining and improving the cost-efficiency of the social security payment system. The system risks discrimination by imposing digital barriers to accessing social security and may therefore exclude individuals with lower levels of digital literacy or without connectivity.[26] The accessibility of the system as well as the use of risk assessments have the potential to affect the human rights of those in vulnerable positions in key areas of life, such as food, housing and work.[27] Predictive analytics may also be used in child safeguarding.[28] For instance, a tool reportedly used by London Councils, in collaboration with private providers, combines data from multiple agencies and applies risk scores to determine the likelihood of neglect or abuse. This raises privacy and data protection concerns as well as issues relating to the right to family life and discrimination.[29] When algorithms are used to support a decision, such as a risk assessment, they may introduce or accentuate existing human rights challenges and pose new issues for accountability.

Considering these examples, the first issue to address is whether an algorithm may be used to make or support a decision in a particular context. Big data-driven algorithms—such as AI or machine-learning algorithms—typically operate on the basis of correlation and statistical probability. Algorithms analyse large volumes of data to identify relationships between particular inputs and a specific output, and make predictions on this basis. In this context, a larger dataset provides a bigger sample size, which can contribute to lower margins of error and a more accurate model. However, the nature of big data-driven algorithms means that they generate results that describe group behaviour, but which are not tailored to specific individuals within that group, irrespective of the size or quality of the input dataset.[30] Yet, big data-driven algorithmic models may be used to make individually-focused decisions. For instance, risk assessment tools, such as COMPAS in the US or HART in the UK, are used to predict factors such as an individual's likely recidivism rate. These algorithms calculate individuals' risk factor using data particular to the individual such as their criminal history and interactions with law enforcement but also variables such as where they live, and their associations with others who have a criminal record.[31] In effect, these tools

[26] UN Special Rapporteur on Extreme Poverty and Human Rights Philip Alston, 'Statement on Visit to the United Kingdom' (n 16).

[27] ibid.

[28] London Councils, 'Keeping Children Safer by Using Predictive Analytics in Social Care Management' <https://www.londoncouncils.gov.uk/our-key-themes/our-projects/london-ventures/current-projects/childrens-safeguarding>.

[29] N McIntyre and D Pegg, 'Councils Use 377,000 People's Data in Efforts to Predict Child Abuse' (*The Guardian*, 16 September 2018) <https://www.theguardian.com/society/2018/sep/16/councils-use-377000-peoples-data-in-efforts-to-predict-child-abuse>.

[30] E Benvenisti, 'Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' (2018) 29(1) EJIL 9, 60.

[31] Northpointe, 'Practitioner's Guide to COMPAS Core' (Northpointe, 19 March 2015) Section 4.2.2 Criminal Associates/Peers and Section 4.2.8 Family Criminality <http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf>; AM Barry-Jester, B

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

make determinations about the likelihood of a particular individual reoffending on the basis of others who share similarities to them. It is foreseeable that these tools could be applied not only to inform, but to actually make decisions in areas such as sentencing, parole or entry into rehabilitation or diversion programmes. Outside the criminal justice context and the social security and social care contexts discussed in the previous paragraph, it is equally foreseeable that algorithms could be used to made decisions regarding an individual's suitability for medical intervention, or for employment. As we discuss further below, the nature of how algorithms work points to risks of arbitrariness, discrimination and a range of human rights issues depending on the context. These types of examples raise questions of whether they could ever be used to make decisions on their own since that decision cannot be individualized.

At the moment, algorithms are typically used to support or inform decision-making, particularly with respect to decisions that explicitly and directly involve human rights, as in the types of examples above. The argument is often made that any shortcomings related to the actual operation of an algorithm may be mitigated by requiring that the algorithm only inform and not make the decision; ie a human 'in the loop' acts a safeguard. However, this gives rise to numerous issues regarding the human 'in the loop's' ability to understand how the algorithm functions and therefore to assign appropriate weight to any recommendation. The degree of deference granted to an automated recommendation is also at issue, as there is a risk that individuals may be reluctant to go against an algorithmic recommendation. This may be because of a perception that an algorithm is neutral or more accurate, or because of the difficulty in explaining why the algorithmic recommendation was overturned. This may render the human 'in the loop' ineffective.

Second, even if the human 'in the loop' is an effective safeguard and the algorithm is only used to inform decisions about sentencing or children at risk, an issue of potential algorithmic bias arises. If the input data is itself biased, say as a result of over-policing of a specific community, or if the algorithm operates in such a way as to produce biased results, then this may give rise to unlawful discrimination. In this regard, modern algorithms depend on good quality input data, but this may not always be available. If particular input data cannot be quantified or obtained, 'proxies' may be used instead. However, as proxies are not an exact substitute they may be inappropriate, inaccurate, or unreliable, affecting the quality and reliability of the results.[32] One example that illustrates the pitfalls of data-driven algorithms is credit scoring. Traditionally, credit scores were calculated on

Casselman and D Goldstein, 'The New Science of Sentencing' (The Marshall Project, 4 August 2015) <https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing>.

[32] See C O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Penguin 2016) Introduction.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

the basis of defined factors such as credit repayment history.[33] With the advent of 'big data', the availability of data used to inform credit decisions has widened, and evaluation can now include information such as social media activity and online shopping patterns.[34] The argument for including these factors is that they may provide more accurate predictions because of 'fuller' data profiles. However, these proxies for creditworthiness are problematic and their incorporation may result in human rights harm. For instance, these new data points may be linked to race or gender and their use may therefore be discriminatory.[35]

Third, there may be a lack of transparency as to the actual operation of the algorithm. For example, this may prevent an accused from challenging the recommendation or risk assessment produced by an algorithm in a sentencing decision, or may prevent a person whose level of social care or social security is to be reduced on the basis of an algorithmic assessment from appealing. Even if there is transparency and a person affected by an algorithmically-influenced decision wishes to challenge that decision, the nature of the algorithmic process may make that very difficult.

One issue in this respect is that a typical application brings together a (potentially large) number of different algorithms that interact to perform a complex task. For instance, a number of different algorithms may be at play with output data from one algorithm providing input data for another. Tracing the factors that contribute to the final output is therefore complex. This complexity is compounded when the development of an application is distributed, either within an organization or through outsourcing, and when deployments utilize input data that is difficult to replicate in a test environment.[36] This diffuses the ability to comprehensively understand the overall operation of an application and thus identify where and/or how human rights issues arise.[37] Equally, machine-learning algorithms can self-learn, identify patterns, and make predictions unimagined by a human operator, and unexplainable by humans.[38] Machine-learning algorithms used to analyse handwriting and sort letters in a post office provide an example. The algorithm analyses a large number of handwriting samples to learn how to classify certain pen marks, infer rules for recognizing particular letters and

[33] US Executive Office of the President, *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights* (May 2016) 11.

[34] M Hurley and J Adebayo, 'Credit Scoring in the Era of Big Data' (2016) 18(1) Yale Journal of Law & Technology 148, 151–2, 163, 166, 174–5.

[35] K Waddell, 'How Algorithms Can Bring Down Minorities' Credit Scores' (*The Atlantic*, 2 December 2016) <https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/>.

[36] For instance, an application may be sold to a third-party who use their own in-house data, or who purchase data sets from data brokers.

[37] See Kitchin (n 1) 21.

[38] F Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015) 3–14.

digits, and develop its own system for doing so.[39] While a human may try to learn when a particular pattern represents the digit '2' on the basis of the curvature of the strokes, etc, a machine-learning algorithm will analyse markedly different factors, such as the configuration and intensity of the shading of relevant pixels. As the algorithm's learning process does not replicate human logic, this creates challenges in understanding and explaining the process.[40] Machine-learning models may also 'learn' in real-time,[41] meaning that over time similar input data may result in different outputs. These systems can thus be unpredictable and opaque, which makes it challenging to meaningfully scrutinize and assess the impact of their use on human rights and thus to effectively challenge decisions made on the basis of algorithms. This was at issue in *State of Wisconsin v Eric L Loomis*,[42] where the defendant raised concerns regarding his inability to challenge the validity or accuracy of the risk assessment produced by the COMPAS tool, which was used to inform his sentencing decision. Issues raised by the defendant included the problem of looking inside the algorithm to determine what weight was given to particular information and how decisions were reached. Difficulties in effectively challenging the risk assessment were acknowledged by the Court,[43] which noted a number of factors suggesting caution vis-à-vis the tool's accuracy:

> (1) the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined; (2) risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed; (3) some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and (4) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations.[44]

The above characteristics of the algorithmic decision-making process all pose human rights challenges and raise difficulties for accountability efforts. These difficulties are compounded when multiple characteristics are present in the same process, as will often be the case. Further complexities arise when the impact of an algorithm on an individual has knock-on effects for others. For

[39] Y LeCun *et al.*, 'Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition' in J-H Oh, C Kwon and S Cho (eds), *Neural Networks: The Statistical Mechanics Perspective* (World Scientific 1995) 261.

[40] N Bostrom and E Yudkowsky, 'The Ethics of Artificial Intelligence' in K Frankish and W Ramsey (eds), *Cambridge Handbook of Artificial Intelligence* (Cambridge University Press 2014) 316, 316–17.

[41] As distinct from learning on the basis of training data, and then being deployed to a real-world context.

[42] *State of Wisconsin v Eric L. Loomis* 2016 WI 68, 881 N.W.2d 749.

[43] Although they were not held to be decisive in respect to the matter at hand.

[44] *State of Wisconsin v Eric L. Loomis* (n 42) para 66.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

instance, if credit decisions are based not only on data specific to an individual, but are expanded to include data relating to those with whom they interact and maintain relationships, it may amplify the discriminatory effect.[45] A poor credit score for a particular individual may result in a poorer score for those in their neighbourhood or social network.[46] This potential cascade effect is often referred to as 'networked discrimination',[47] which echoes the historically discriminatory practice of 'redlining', whereby entire neighbourhoods of ethnic minorities were denied loans by virtue of where they lived.[48]

To counter the potential adverse effects of the way in which algorithms work on human rights, scholars and practitioners have focused on addressing the way in which algorithms function and their transparency, explainability and understandability, as discussed in the next section. We argue that although these approaches are necessary, in and of themselves they are not sufficient to address the overall risks posed to human rights.

B. Existing Proposals for Algorithmic Accountability and Their Ability to Address the Impact of Algorithms on Human Rights

The pursuit of 'algorithmic transparency' is a key focus of existing approaches to algorithmic accountability. This relates to the disclosure of information regarding how algorithms work and when they are used.[49] To achieve transparency, information must be both accessible and comprehensible.[50] Transparency in this context can relate to information regarding why and how algorithms are developed,[51] the logic of the model or the overall design,[52] the assumptions underpinning the design process, how the performance of the algorithm is monitored,[53] how the algorithm itself has changed over time,[54] and factors relevant to the functioning of the algorithm,

[45] See eg Lenddo, 'Credit Scoring Solution', <https://www.lenddo.com/pdfs/Lenddo_FS_CreditScoring_201705.pdf>, which includes social network data in credit scores.

[46] J Angwin *et al.*, 'Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas with the Same Risk' (ProPublica, 5 April 2017) <https://www.propublica.org/article/minority-neighborhoods-higher-car-insurance-premiums-white-areas-same-risk>; M Kamp, B Körffer and M Meints, 'Profiling of Customers and Consumers – Customer Loyalty Programmes and Scoring Practices' in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008) 201, 207.

[47] D Boyd, K Levy and A Marwick, 'The Networked Nature of Algorithmic Discrimination' (Open Technology Institute, October 2014).

[48] *Conference of Federal Savings & Loan Associations v Stein*, 604 F.2d 1256 (9th Cir. 1979), *aff'd mem.*, 445 U.S. 921 (1980) 1258.

[49] N Diakopoulos and M Koliska, 'Algorithmic Transparency in the News Media' (2017) 5(7) Digital Journalism 809, 811.

[50] BD Mittelstadt *et al.*, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) Big Data & Society 6.

[51] See Diakopoulos and Koliska (n 49) 816.

[52] See Ananny and Crawford (n 9) 977.

[53] D Kehl, P Guo and Samuel Kessler, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' (July 2017) Responsive Communities 32–3.

[54] See Diakopoulos and Koliska (n 49) 816.

such as data inputs (including proxies), and the relative weight attributed to inputs.[55] Transparency can also relate to the level of human involvement,[56] in order 'to disentangle the roles and decisions of humans versus algorithms'.[57] This section discusses why transparency is valuable for accountability, addresses challenges in achieving transparency, and highlights remaining accountability gaps.

1. *The value of transparency*

The focus on transparency is a result of the nature and complexity of modern algorithms and the view that if algorithms cannot be scrutinized, any risks to human rights within decision-making processes will be difficult to identify and to rectify. Transparency is essential for trust, and to ensure that a system operates within appropriate bounds.[58] The ability to predict the behaviour of an algorithm and to explain the process by which it reasons is necessary to control, monitor, and correct the system,[59] and to audit and challenge decisions supported by algorithms.[60] Understanding how an algorithm works can also be useful in anticipating how it could perform if deployed in a different context.[61] Some authors have asserted that transparency should be the policy response for any governmental use of automated decision-making.[62]

2. *Transparency challenges*

Notwithstanding the importance of transparency as a normative objective, some commentators have noted that it may be difficult to achieve in practice,[63] highlighting that of itself transparency may not be meaningful.[64] For example, certain algorithms can 'learn' and modify their operation during

[55] See Kehl, Guo and Kessler (n 53) 28.

[56] N Diakopoulos, 'Accountability in Algorithmic Decision Making' (2016) 59(2) Communications of the ACM 56, 60.

[57] See Diakopoulos and Koliska, n (n 49) 822.

[58] See Diakopoulos (n 56) 58–9, 61; Diakopoulos and Koliska (n 49) 810–12; L Edwards and M Veale, 'Slave to the Algorithm: Why A 'Right to An Explanation' Is Probably Not the Remedy You Are Looking For' (2017) 16(1) Duke Law & Technology Review 18, 39; A Tutt, 'An FDA for Algorithms' (2017) 69(1) Administrative Law Review 83, 110–11.

[59] The Royal Society, 'Machine Learning: The Power and Promise of Computers That Learn by Example' (2017) 93–4; See Tutt (n 58) 101–4.

[60] See Ananny and Crawford (n 9) 975–7; E Ramirez, Chairwoman, Federal Trade Commission, 'Privacy Challenges in the Era of Big Data: A View from the Lifeguard's Chair' (Keynote Address at the Technology Policy Institute Aspen Forum, Aspen, Colorado, 19 August 2013) 8; Kehl, Guo and Kessler (n 53) 32–3.

[61] See The Royal Society (n 59) 93.

[62] AR Lange, 'Digital Decisions: Policy Tools in Automated Decision-Making' (Center for Democracy & Technology, 2016) 11.

[63] See Kroll *et al.* (n 9) 639.

[64] See Kroll *et al.* (n 9) 638, 657–60; BW Goodman, 'A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection' (29th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016) 3–4.

deployment,[65] and so the factors that inform a decision (and the resultant outputs) may vary over time, reducing the utility of transparency-induced disclosure.[66] Equally, transparency as to when an algorithm is deployed may not be meaningful unless it is possible to explain the underlying logic, or to interrogate the input data.

Blockchain, an open distributed ledger system that records transactions,[67] is one technical tool that has been suggested as a potential solution.[68] To date, blockchain has been used to reconcile transactions distributed across various entities within and between organisations. This existing ability to track items and specific financial transactions may be adapted and applied to the use of specific data points throughout an algorithmic decision-making process. For example, other authors have suggested that blockchain may be used to track data provenance and to improve accountability in the use of data, by verifying 'if the data was accessed, used and transferred' in compliance with users' consent.[69] This could facilitate tracing back through a decision to see which data points informed it and the weight they were given.

Nonetheless, the extent to which transparency challenges can be overcome is a live debate, and a number of complicating factors arise. First, businesses have an understandable proprietary interest in the algorithms they develop and so may be unwilling to reveal the underlying code or logic.[70] To overcome this challenge, suggestions have been made that the algorithm does not have to be made publicly transparent but rather could be subject to independent review by an ombud for example.[71] Second, transparency regarding an algorithm's code or underlying logic may be undesirable.[72] This 'inside' knowledge may facilitate the 'gaming' of the system,[73] resulting in abuse, and improper results. The risk is particularly clear in the context of security screening or tax audits.[74] In other situations, such as those involving 'sensitive data',

[65] At a simpler level, algorithms themselves may be modified due to a normal update/development system.

[66] See Kroll *et al.* (n 9) 647–52.

[67] MIT Technology Review Editors, 'Explainer: What is a Blockchain?' (MIT Technology Review, 23 April 2018) <https://www.technologyreview.com/s/610833/explainer-what-is-a-blockchain/>.

[68] M Burgess, 'Holding AI to Account: Will Algorithms Ever Be Free from Bias if They're Created by Humans?' (The Wired, 11 January 2016) <https://www.wired.co.uk/article/creating-transparent-ai-algorithms-machine-learning>.

[69] R Neisse, G Steri and I Nai-Fovino, 'A Blockchain-based Approach for Data Accountability and Provenance Tracking' (12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy, August 2017) 1.

[70] J Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society 3–4.

[71] See Tutt (n 58) 117–18.

[72] Some authors argue that transparency is not simply full informational disclosure, and that notions of transparency against secrecy is a false dichotomy. See Ananny and Crawford (n 9) 979; Diakopoulos (n 56) 58–9.

[73] Science & Technology Committee, *Oral Evidence: Algorithms in Decision-Making* (HC 2017–2019, 351), 12 December 2017 Evidence Session, Q112-113 <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/oral/75798.pdf>.

[74] See Kroll *et al.* (n 9) 639, 658.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

transparency may be legally restricted.[75] Third, the complex interaction between algorithms and human agents is another source of opacity. When algorithms assist human decision-making, it is difficult to determine the influence of the algorithm's results on the final decision, and to identify whether inappropriate deference is given to the algorithm.[76] As such, and irrespective of transparency issues, it is also necessary to evaluate how algorithmic outputs influence human decision-making within the context of the overall process. This is discussed in greater detail in Part IV. Fourth, even if it is possible to fully explain the system's reasoning, an important question arises regarding the resources and expertise required to do so.[77] Addressing this question will involve assessments of the cost of transparency against the reasons for using the algorithm in the first place (which may often relate to competitive pricing).

3. *The gaps remaining in the accountability process*

Transparency is essential to accountability but insufficient of itself. This section identifies five additional factors necessary for effective accountability, many of which have either not been addressed or have not been evaluated holistically in existing debates. First, a clear understanding of what constitutes 'harm' is a prerequisite to, and benchmark for, evaluating risks and effects of the use of algorithms in decision-making. In the absence of an agreed understanding, 'harm' is open to a number of different interpretations, and the understanding adopted by a particular actor may fail to effectively take into account the full human rights impact of their actions. For instance, a business' 'community values' may not fully match IHRL. For example, they could focus on the right to privacy but not incorporate the right to freedom of expression or the prohibition of discrimination. Second, in order to prevent and protect against harm, the overall decision-making process and the full life cycle of an algorithm must be taken into account, and the specific role played by an algorithm in any final decision identified. Design, development and deployment of algorithms are interconnected phases within an overall process and decisions made in one phase may affect human rights compliance in another. For example, it may not be possible to monitor the potential discriminatory impact of an algorithm if this is not built in during the development phase. Equally, the role played by an algorithm in the final decision, such as whether it is used to make or inform that decision, will impact upon the human rights considerations. Third, the obligations and responsibilities of States and businesses respectively need to be ascertained from the outset, noting that these will depend on their specific role at different stages of the overall decision-making process. Fourth, remedies for harm caused

[75] ibid, 639.

[76] See Diakopoulos (n 56) 57, 60; M Wilson, 'Algorithms (and the) Everyday' (2017) 20(1) Information, Communication & Society 137, 141, 143–44, 147.

[77] See Burrell (n 70) 4.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

must be addressed. To date, the concept of remedy has narrowly focused on fixing the operation of the algorithm where bias is identified, but the concept of an effective remedy under IHRL is much broader by focusing on the individual(s) affected as well as taking measures to ensure that the harm is not repeated in the future. Fifth, an overall shift in focus may be required. Existing approaches to accountability tend to focus on after-the-fact accountability. While this is important, it is also crucial that accountability measures are fully incorporated throughout the overall algorithmic life cycle, from conception and design, to development and deployment.[78] Discussion in this regard is emerging, for instance with respect to whether and how 'ethical values' can be built into algorithms in the design phase,[79] and whether algorithms can themselves monitor for ethical concerns. While this is a welcome start, the discussion needs to go further, and the operationalization of the IHRL framework can play a significant role in this regard.

Achieving effective accountability is therefore a complex problem that demands a comprehensive approach. Somewhat surprisingly, IHRL has been neglected in existing discourse until relatively recently.[80] While values such as dignity,[81] legal fairness, and procedural regularity in the design of algorithms[82] are referenced in the literature, neither the range of substantive rights established under IHRL, nor the possibility that IHRL provides a framework capable of underpinning the overall algorithmic accountability process, have received significant attention. This is beginning to change with more actors starting to support a human rights-based approach to the development and use of artificial intelligence. The next part of this article contributes to these developments by clearly setting out the specific contribution that IHRL can make. This article advances an overall framework through which to address the issue of algorithmic accountability, and adds depth to existing discussion.

III. THE CONTRIBUTION OF THE HUMAN RIGHTS FRAMEWORK TO ALGORITHMIC ACCOUNTABILITY

IHRL contributes to the algorithmic accountability discussion in three key ways. First, it fills a gap in existing discourse by providing a means to define

[78] Science & Technology Committee (n 73) Q207.

[79] See Institute of Electrical and Electronics Engineers Global Initiative on Ethics of Autonomous and Intelligent Systems, 'Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, (Version 2) (2017); M Ananny, 'Towards an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness' (2016) 41(1) Science, Technology & Human Values 93, 94–6; L Jaume-Palasí and M Spielkamp, 'Ethics and Algorithmic Processes for Decision Making and Decision Support' (AlgorithmWatch, Working Paper No. 2, 2017) 9–13; Mittelstadt *et al.* (n 50) 10–12.

[80] See discussion Part IIB.

[81] C Cath *et al.*, 'Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach' (2018) 24(2) Science & Engineering Ethics 505, 508.

[82] See Kroll *et al.* (n 9) 637–8, 662–72, 678–9.

and assess harm. Second, it imposes specific obligations on States and expectations on businesses to prevent and protect human rights and sets out the mechanisms and processes required to give effect to or operationalize these obligations and responsibilities. Third, the IHRL framework can map on to the overall algorithmic life cycle and thus provides a means for assessing the distinct responsibilities of different actors across each stage of the process. IHRL therefore establishes a framework capable of capturing the full algorithmic life cycle from conception to deployment. Although we do not suggest that IHRL provides an exclusive approach, it does provide a key lens through which to analyse accountability. As such, it forms an important dimension and organizer for algorithmic accountability that fits together with existing approaches such as transparency, explainability, and technical solutions.[83] Necessarily, the specifics of the approach will need to be further developed and refined in a multi- and interdisciplinary way.

A. IHRL as a Means for Assessing Harm

In the current discourse on 'algorithmic accountability' harm is regularly referred to but often using vague or abstract terms such as unfairness, or by reference to voluntary corporate policies.[84] These terms make it difficult to pinpoint the exact nature of the harm and to assess whether and which legal obligations attach. There is also a risk that the extent of potential harm is underplayed or narrowly construed.[85]

The focus on 'bias' illustrates the risks arising in this regard. Within the literature on algorithmic accountability, the term 'bias' (and less often 'discrimination') is used in a range of different ways, often without clarity on the meaning employed. It is sometimes used to convey a specific technical meaning, for example with reference to statistical bias.[86] In other contexts, it

[83] The utility of this approach has recently been noted by others. See Amnesty International & Access Now, 'The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems' (16 May 2018) <https://www.accessnow.org/cms/assets/uploads/2018/05/Toronto-Declaration-D0V2.pdf>. The Toronto Declaration applies the human rights framework with a focus on the right to equality and non-discrimination. This article proposes a human rights-based framework based on the full range of substantive and procedural rights. See discussion Part IIIA.

[84] See eg Facebook, 'Community Standards', <https://www.facebook.com/communitystandards/>.

[85] In her oral evidence to the House of Commons Science and Technology Committee inquiry on algorithms in decision-making, Sandra Wachter suggested that a more refined harm taxonomy is required to respond to the ethical and real-world problems that may be difficult to predict at the outset and 'new harms and new kinds of discrimination' arising from inferential analytics. See Science & Technology Committee, *Oral Evidence: Algorithms in Decision-Making* (HC 351, 2017–2019) 14 November 2017 Evidence Session, Q55 <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/oral/73859.pdf>. This paper agrees that a robust understanding for harm is necessary, but argues that it is provided by existing definitions in IHRL.

[86] C Dwork *et al.*, 'Fairness through Awareness' (3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, January 2012) 215.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

is employed as a general, 'catch-all' term to mean some form of preference or 'unfairness' (which itself has been criticized as a vague term). When used in such a broad way, actors may develop or gravitate to locally defined understandings as to what constitutes bias or discrimination, giving rise to a variety of meanings. This can also create uncertainty for actors designing, developing and using algorithms in decision-making as to whether a particular instance of bias is unlawful.

Scholars sometimes indicate that unlawful bias may constitute a narrower category under an overall heading of 'bias' but without concretely explaining how.[87] IHRL can make a central contribution in this regard as a counter to general descriptors of 'bias' or 'discrimination' by providing a method for understanding when bias and discrimination are unlawful. In this regard, IHRL provides a concrete and universally applicable definition of harm that is capable of identifying prohibited and unlawful forms of bias and discrimination.[88] This definition is accompanied by well-developed and sophisticated tests for establishing when the prohibition of discrimination has been violated, including what constitutes direct, indirect or intersectional discrimination as well as structural and unconscious bias. IHRL therefore not only provides a means to determine harm through its interpretation of how rights may be interfered with, it also provides established tests to assess when and how rights may have been violated.

The IHRL framework also offers a deeper and fuller means of analysing the overall effect of the use of algorithms. This moves beyond the current singular and narrow framings of harm which tend to focus on 'bias' or 'privacy' to look at the full impact of algorithms on the rights of individuals and groups. For example, the use of algorithms to aid sentencing and parole decisions have been reported to be 'biased' against certain ethnic minorities.[89] The IHRL framework not only assesses whether such use violates the prohibition of discrimination but also examines the impact from the perspective of the individual's right to a fair trial and to liberty. This broader approach is essential as it captures the overall impact of algorithms and may indicate, for example, that algorithms cannot be used in a particular context, even if discrimination-related concerns are addressed.[90]

The IHRL framework therefore provides a means of categorizing and labelling harm through its establishment of an internationally agreed set of substantive and procedural rights which, if violated, constitute harm.

[87] R Binns, 'Fairness in Machine Learning: Lessons from Political Philosophy' (Conference on Fairness, Accountability and Transparency (New York, 2018) 3–5.

[88] See eg UN Human Rights Committee, 'General Comment No. 18 (Non-discrimination' (10 November 1989) para 6.

[89] J Angwin *et al.* (n 46); F Raso *et al.* (n 16) 21–4; R Caplan *et al.*, 'Algorithmic Accountability: A Primer, Tech Algorithm Briefing: How Algorithms Perpetuate Racial Bias and Inequality' (Data & Society, 18 April 2018) <https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf.

[90] See further Part IVA.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

Incorporating a means to assess (potential) harm is critical to developing an effective accountability framework for the use of algorithms in decision-making. Importantly, the IHRL framework not only describes the nature of harm but triggers an existing framework that attaches to these rights. As discussed in the next section, this framework connects directly to concrete legal obligations imposed on States to prevent and protect against such violations, including with respect to the regulation of business actors, and establishes clear expectations on businesses themselves as regards to the actions necessary to respect human rights. IHRL thus brings clarity regarding the actions that States and businesses are expected to take and the consequences of failing to act.

B. Clearly Defined Obligations and Expectations That Apply Across the Algorithmic Life Cycle

By identifying the range of rights brought into play by the use of algorithms in decision-making, IHRL establishes a clear set of obligations on States and expectations on businesses to prevent and protect human rights across the algorithmic life cycle. Focusing on existing legal obligations is critical as it emphasizes that addressing the (potential) harm caused by the use of algorithms in decision-making is not a voluntary exercise, as often appears to be the implication in existing discourse and debates.[91] This section analyses how the IHRL framework applies across the life cycle of algorithms in order to demonstrate its potential contribution to filling gaps in the current accountability debate.

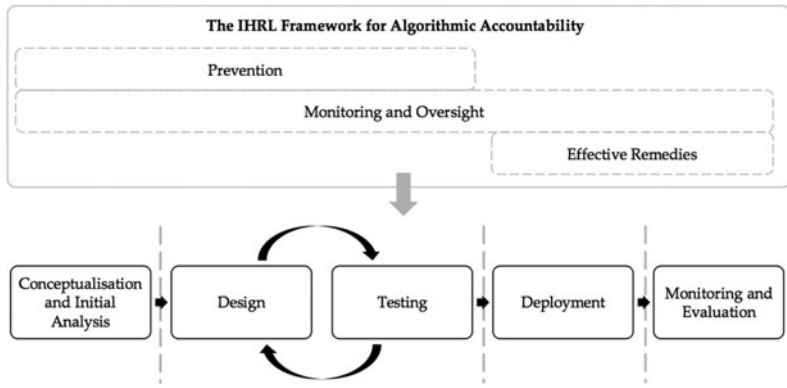
1. Identifying roles and responsibilities attached to different entities across the full algorithmic life cycle

IHRL requires that States put in place an accountability framework that prevents violations from taking place, establishes monitoring and oversight mechanisms as safeguards, and provides a means to access justice for individuals and groups who claim that their rights have been violated.[92] The components of this accountability framework are necessarily interdependent, and apply across the full algorithmic life cycle. This is illustrated in general terms in the diagram below.

[91] It is noted that, even where international law does not impose direct obligations on businesses, States' obligations to protect against human rights violations requires that they take measure at the domestic level to ensure that individuals' human rights are not violated by businesses. States may be held accountable for failure to take appropriate measures in this regard. See Ruggie Principles (n 14) Principle 1.

[92] See UN Human Rights Committee, General Comment No. 31 (n 13) paras 3–8; UN Committee on Economic, Social and Cultural Rights, General Comment No. 3 (n 13) paras 2–8.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press



As IHRL traditionally focuses on State conduct, these obligations apply to the actions or omissions of the State directly. This is important as States are increasingly reported to be integrating algorithms within their decision-making processes across a range of sectors that may have significant consequences for individuals and groups in areas such as policing, sentencing, social security and the identification of children at risk, as noted above.[93] IHRL also addresses business activity, by requiring the State to protect against third-party harm, and by imposing specific expectations directly on businesses. The principle of due diligence requires States to prevent and protect individuals from harm by third parties, including business enterprises. For instance, States are required to devise 'appropriate steps to prevent, investigate, punish and redress private actors' abuse … [t]hey should consider the full range of permissible preventative and remedial measures, including policies, legislation, regulations and adjudication'.[94] Human rights standards and norms also apply directly to business enterprises, as articulated, for example, in the UN Guiding Principles on Business and Human Rights (the

[93] See eg Canadian Institute for Advanced Research, 'CIFAR Pan-Canadian Artificial Intelligence Strategy' <https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy>; E Macron, 'Artificial Intelligence: "Making France a Leader"' (AI for Humanity Conference, Collège de France, 30 March 2018) <https://www.gouvernement.fr/en/artificial-intelligence-making-france-a-leader>; NITI Aayog, 'National Institution for Transforming India (national Strategy for Artificial Intelligence #AIFORALL' (June 2018), <http://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf>; Japan Strategic Council for AI Technology, 'Artificial Intelligence Technology Strategy' (New Energy and Industrial Technology Development Organization, 31 March 2017) <http://www.nedo.go.jp/content/100865202.pdf>; AI Singapore, <https://www.aisingapore.org/>; UK Department for Digital, Culture, Media & Sport, 'Policy Paper: AI Sector Deal' (26 April 2018) <https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal>; US White House, 'Artificial Intelligence for the American People' (10 May 2018) <https://www.whitehouse.gov/briefings-statements/artificial-intelligence-american-people/>.

[94] See Ruggie Principles (n 14) Principle 1.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

Ruggie Principles). These principles establish an expectation that businesses should prevent or mitigate 'adverse human rights impact', establish a means of access to justice where human rights violations are alleged, and provide remedies where rights are found to be breached. For both States and businesses, giving effect to these obligations or expectations requires establishing monitoring and oversight mechanisms that apply throughout the entire algorithmic process.

As discussed in the next sections, IHRL sets out the measures required if States or businesses are to comply with human rights law: IHRL details the actions that different actors should take at each point in the process, from conception to deployment. This facilitates a means of engaging with the overall picture, both in terms of the process of developing and deploying algorithms, and evaluating their impact. This portrayal of how to achieve a holistic approach to accountability is currently absent from the discourse.

2. *Operationalizing the measures necessary to ensure rights compliance*

IHRL provides a range of measures to guide States in the implementation and operationalisation of their obligations to prevent and protect human rights, and to guide businesses regarding the actions they should take to respect human rights. For example, the Office of the UN High Commissioner for Human Rights (OHCHR) defines 'direct prevention' as 'aim[ing] to eliminate risk factors and establish a legal, administrative and policy framework which seeks to prevent violations'.[95] It notes that this obligation comprises a number of different elements, highlighting that some 'provisions point to an obligation of negative result (prevention being successful when there is no violation), while in some cases prevention can be seen as an obligation of positive conduct (taking all necessary steps to adopt concrete and effective measures to prevent violations)'.[96] The prevention of harm is directly linked to the obligation to respect, whereby States must refrain from taking measures that will result in a human rights violation.[97] The UN Guiding Principles on Business and Human Rights apply this same requirement to businesses, in terms of '[t]he responsibility of business enterprises to respect human rights'.[98] As discussed above, in ensuring that they respect human rights, States and businesses need to ensure that policies and practices are in place to identify and assess any actual or potential risks to human rights posed by the use of algorithms in decision-making.

[95] UN Human Rights Council, 'Report of the Office of the UN High Commissioner for Human Rights on 'The Role of Prevention in the Promotion and Protection of Human Rights' (16 July 2015) UN Doc A/HRC/30/20, para 9.

[96] ibid, para 7.

[97] See eg UN Committee on Economic, Social and Cultural Rights, 'General Comment 14 The right to the highest attainable standard of health (article 12 of the International Covenant on Economic, Social and Cultural Rights)' (11 August 2000) UN Doc E/C.12/2000/4, para 33.

[98] Ruggie Principles (n 14) Principle 12.

The IHRL framework provides further guidance as to the type of measures that can operationalize the respect principle. The full life cycle approach allows for existing algorithmic accountability proposals—relating, for example, to auditing or impact assessments—to be situated within a comprehensive process. This facilitates greater clarity and focus by setting out what the objectives underpinning specific measures should be, their scope and depth, what the indicators of effectiveness are, and when measures should be undertaken.

Impact assessments provide an example. As a result of a narrow conceptualization of harm, impact assessments in an algorithmic context have typically focused on issues relating to discrimination and privacy.[99] The IHRL framework contributes in three key ways. First, it clarifies the content of the right to privacy and the prohibition of discrimination. This ensures that all aspects of the rights—including indirect discrimination, for example—are taken into account, while also facilitating consistency across assessments. Second, the use of algorithms in decision-making can potentially affect all rights. The IHRL framework requires that impact assessments encompass the full set of substantive and procedural rights under IHRL, and that analysis not be unduly limited to privacy or discrimination. Third, the IHRL framework underscores the need for risks to be monitored at all stages of the algorithmic life cycle.[100]

Applying this in practice means that impact assessments should be conducted during each phase of the algorithmic life cycle. During the design and development stage, impact assessments should evaluate how an algorithm is likely to work, ensure that it functions as intended and identify any problematic processes or assumptions. This provides an opportunity to modify the design of an algorithm at an early stage, to build in human rights compliance—including monitoring mechanisms—from the outset, or to halt development if human rights concerns cannot be addressed. Impact assessments should also be conducted at the deployment stage, in order to monitor effects during operation. As stated, this requires that, during design and development, the focus should not only be on testing but steps should also be taken to build in effective oversight and monitoring processes that will be able to identify and respond to human rights violations once the algorithm is deployed. This ability to respond to violations is key as IHRL requires that problematic processes must be capable of being reconsidered, revised or adjusted.[101]

The establishment of internal monitoring and oversight bodies can play an important role in coordinating and overseeing the implementation of regular

[99] See eg K Crawford and J Schultz, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms' (2014) 55(1) Boston College Law Review 93, 95; Kroll *et al.* (n 9) 678.

[100] For instance, indirect discrimination may only become visible during the deployment phase.

[101] See OHCHR (n 95) para 31.

impact assessments and ensuring that findings are addressed. Some businesses in the AI sector have started to develop internal committees.[102] The nature and mandate of such committees has been subject to some commentary,[103] and reflects an evolving dimension to legal and policy debates on algorithmic accountability.

Notwithstanding the nature of any internal processes, independent oversight plays an important role in the AI sector as it does in other areas where decision-making by States or businesses has the potential to adversely affect human rights. For instance, independent oversight is a core requirement with respect to State surveillance activities.[104] To date, similar oversight models have not been introduced to the algorithmic decision-making context. However, independent oversight mechanisms may be central to ensuring that States and businesses comply with their respective human rights obligations and responsibilities. They may provide an appropriate means to assess the processes put in place by States and businesses, and may also provide expert input vis-à-vis potential risks. An independent oversight body may also play an essential role in determining whether algorithms can be used in certain contexts, and if so, under what conditions, as discussed in the next part of this article.[105] Independent oversight may take a number of different forms, dependent upon factors such as the public or private function of the algorithm during deployment. For algorithms deployed in a public decision-making context, an independent body, established in legislation, and sufficiently resourced (including with appropriate technical expertise) may be the most appropriate.[106] The newly established UK Centre for Data Ethics and Innovation is an interesting proposition in this regard. This body is intended to strengthen the existing algorithmic governance landscape,[107] but its role is limited to the provision of advice: it is an 'advisory body that will

[102] See eg Microsoft, 'Satya Nadella Email to Employees: Embracing Our Future: Intelligent Cloud and Intelligent Edge' (Microsoft, 29 March 2018) <https://news.microsoft.com/2018/03/29/satya-nadella-email-to-employees-embracing-our-future-intelligent-cloud-and-intelligent-edge/>; S Pichai, 'AI at Google: Our Principles' (Google, 7 June 2018) <https://www.blog.google/technology/ai/ai-principles/>; DeepMind, 'DeepMind Ethics & Society' <https://deepmind.com/applied/deepmind-ethics-society/>.

[103] A Hern, 'DeepMind Announces Ethics Group to Focus on Problems of AI' (*The Guardian*, 4 October 2017) <https://www.theguardian.com/technology/2017/oct/04/google-deepmind-ai-artificial-intelligence-ethics-group-problems>; J Temperton, 'DeepMind's New AI Ethics Unit Is The Company's Next Big Move' (The Wired, 4 October 2017) <https://www.wired.co.uk/article/deepmind-ethics-and-society-artificial-intelligence>; T Simonite, 'Tech Firms Move To Put Ethical Guard Rails Around AI' (The Wired, 16 May 2018) <https://www.wired.com/story/tech-firms-move-to-put-ethical-guard-rails-around-ai/>.

[104] See *Zakharov v Russia*, App No 47143/06 (ECtHR, 4 December 2015) para 233.

[105] See Parts IVA and IVB.

[106] See, by way of analogy to security oversight, Council of Europe Commissioner for Human Rights, 'Issue Paper: Democratic and Effective Oversight of National Security Services' (Council of Europe, 2015) 47.

[107] UK Department for Digital, Culture, Media & Sport, 'Centre for Data Ethics and Innovation: Consultation' (June 2018) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/715760/CDEI_consultation__1_.pdf> 10.

investigate and advise on how [the UK] govern[s] the use of data and data-enabled technologies'.[108] As such, it cannot qualify as an effective oversight body. However, it is conceived, at least in part, as an interim measure to 'allow the government time to test the value and utility of the Centre's functions ahead of the creation of a future statutory advisory body'.[109] Recently, Canada produced a white paper addressing 'Responsible Artificial Intelligence in the Government of Canada', which highlighted the need for an oversight body to review automated decision-making, and to provide advice to ministers during the design of AI systems.[110] Lessons learned from experiences such as these may provide valuable insight going forward. Independent oversight bodies established to monitor State surveillance activity and analysis of their effectiveness may also provide points of reference and comparison.[111] Other models being proposed include dedicated ombuds for the AI sector or the expansion of the mandate of existing ombuds to address these issues as well as industry regulatory bodies.[112]

In the event that violations are found to have occurred, IHRL imposes a number of requirements: measures must be put in place to prevent any reoccurrences, those affected must be provided with effective reparation, and those responsible must be held to account. Within current accountability debates, however, there is often a narrow focus on addressing problems with the use of an algorithm, in order to fix the issue and prevent reoccurrences.[113] This is, of course, an important measure, and one that aligns with IHRL requirements, particularly those relating to the concept of guarantees of non-repetition. However, under the IHRL framework this is just one component of a larger process. In order for individuals and groups to challenge the impact of the use of algorithms in decision-making, IHRL requires that States and businesses provide a means to access justice for those with an arguable claim that their rights have been violated. Given the lack of transparency in this area, broad standing provisions may be necessary, in order to enable individuals to bring claims if they suspect but cannot prove that they have been adversely affected by an algorithmic decision-making process. States

[108] UK Department for Digital, Culture, Media & Sport, 'Centre for Data Ethics and Innovation: Government Response to Consultation' (November 2018) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/757509/Centre_for_Data_Ethics_and_Innovation_-_Government_Response_to_Consultation.pdf> 5.

[109] ibid, 12.

[110] Government of Canada, Treasury Board of Canada Secretariat, 'Responsible Artificial Intelligence in the Government of Canada: Digital Disruption White Paper Series' (Version 2.0, 10 April 2018) <https://docs.google.com/document/d/1Sn-qBZUXEUG4dVk909eSg5qvfbpNlRhzIefWPtBwbxY/edit> 32–3.

[111] See, for instance, the Investigatory Powers Commissioner's Office established to oversee the UK Investigatory Powers Act 2016.

[112] See C Miller, J Ohrvik-Scott and R Coldicutt, 'Regulating for Responsible Technology: Capacity, Evidence and Redress' (Doteveryone, October 2018).

[113] See Kitchin (n 1) 17; M Hardt, E Price and N Srebro, 'Equality of Opportunity in Supervised Learning' 2 (30th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016).

and businesses may develop internal processes for individuals to submit complaints to them directly, although there is currently debate on whether complainants should be required to use such processes, where they are not independent of the entity concerned.[114] What is critical, therefore, is that States establish complaints mechanisms such as ombuds and effective judicial remedies. Determining which entity (or entities) is responsible for a particular harm is equally essential in order to allocate responsibility for providing an effective remedy. The approach to remedy within the algorithmic accountability discussion needs much greater attention, including examination of the possibility that remedies could themselves be driven by algorithms.

3. *Integrating a rigorous accountability framework*

By situating obligations and expectations across the life cycle of an algorithm, what is required of the different actors involved becomes clearer. This is particularly important given the number of different entities that may be involved across the algorithmic life cycle, and the fact that algorithms may be sold and deployed in a variety of different contexts. Indeed, various public-sector organizations have already integrated the use of algorithms into their decision-making processes. For example, in the UK the 'Harm Assessment Risk Tool' (also known as 'HART') is used by Durham Constabulary to determine which individuals are eligible for an out-of-court process, intended to reduce future offending. The tool was developed by statistical experts based at the University of Cambridge in collaboration with Durham Constabulary.[115] Elsewhere in the UK local governments have been using products and services developed by private companies in areas such as child safeguarding, welfare services, and education.[116] This dimension has been raised, but not adequately explored, in existing discourse. The IHRL framework incorporates the diversity of actors involved, and allows for nuance with respect to the obligations or expectations imposed on different actors.

For instance, the obligation/responsibility to respect requires that an entity developing an algorithm identify any potential harm to rights, and take

[114] L McGregor, 'Activating the Third Pillar of the UNGPs on the Right to an Effective Remedy' (EJIL: *Talk!*, 23 November 2018) <https://www.ejiltalk.org/activating-the-third-pillar-of-the-ungps-on-access-to-an-effective-remedy/>.

[115] M Oswald *et al.*, 'Algorithmic Risk Assessment Policing Models: Lessons From the Durham HART Model and 'Experimental' Proportionality' (2018) 27(2) Information & Communications Technology Law 223, 225.

[116] See, for example, London Ventures <https://www.londoncouncils.gov.uk/our-key-themes/london-ventures>; Data Justice Lab, 'Digital Technologies and the Welfare System, Written Submission to the UN Special Rapporteur on Extreme Poverty and Human rights Consultation on the UK' (14 September 2018) <https://www.ohchr.org/Documents/Issues/Epoverty/UnitedKingdom/2018/Academics/DataJusticeLabCardiffUniversity.pdf> 2–3.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

measures to protect against that harm. If the algorithm is to be sold, the developer must also consider future third-party deployments. This may require clarification or elaboration of a number of requirements, such as the intended circumstances of use, the volume and quality of input data required, or the statistical accuracy of the results. This is a means of satisfying the developer's own human rights obligations/responsibilities and facilitating human rights compliance by any subsequent users. If an algorithm is sold, the purchaser's human rights obligations or responsibilities are also brought into play; ie the purchaser must perform their own impact assessment and deploy the algorithm in line with the developer's specifications. If the purchaser subsequently modifies the circumstances of use, they will necessarily have to carry out further impact assessments. The IHRL framework accordingly allows for a division and allocation of responsibilities. For instance, if the original developer fulfils their human rights obligations, then they cannot be held responsible for subsequent third-party misuse, and responsibility will accordingly lie with the purchaser. Equally, if the purchaser deploys the algorithm appropriately (ie in line with their human rights obligations) but a problem arises as a result of the developer's lack of compliance, then responsibility lies with the developer not the purchaser.[117]

To summarize, IHRL defines harm in a universally accepted form and sets out the specific obligations or expectations that apply to the different actors involved across each stage of the algorithmic life cycle. IHRL also details the means necessary to ensure human rights compliance, setting out the different mechanisms that may be employed, and clarifying the objectives underpinning these measures. By looking at the overall algorithmic life cycle, and requiring that human rights obligations/expectations are taken into account from the conception stage, the IHRL framework also facilitates effective accountability and compliance. For instance, it may be difficult if not impossible to detect harm at the deployment stage if oversight mechanisms are not built in during development. Equally, if potential indirect discrimination is not identified by a pre-deployment impact assessment, the consequences for affected individuals or groups may be significant. Ultimately, the comprehensive full life cycle overview approach facilitated by the IHRL framework is essential in order to ensure that technology serves, rather than undermines, human and societal interests.

The IHRL framework clearly sets out the measures that all actors should take—and which States must take—in order to ensure that the design, development and deployment of algorithms is undertaken in a human-rights-compliant manner. It is a clear expectation of the international community

[117] This example is provided for illustrative purposes. It is not intended to be an absolute guide to responsibility: this is something that must be evaluated on a case-by-case basis.

that businesses fulfil their responsibility to respect human rights.[118] The framework elaborated above provides a road map in this regard.

IV. THE EFFECT OF APPLYING THE IHRL FRAMEWORK TO THE USE OF ALGORITHMS IN DECISION-MAKING

This part analyses how the application of the IHRL framework may affect decisions regarding the development and deployment of algorithms. To reiterate, the application of the IHRL framework is intended to ensure that the potential inherent in technology can be realized, while at the same time ensuring that technological developments serve society. As such, the increasing centrality of algorithms in public and private life should be underpinned by a framework that attends to human rights. This is not intended to be anti-technology or anti-innovation, it is directed at human-rights-compliant, socially beneficial, innovation.

A. Are There Red Lines That Prohibit the Use of Algorithms in Certain Instances?

Most of the debates on algorithmic accountability proceed from the assumption that the use of algorithms in decision-making is permissible. However, as noted at the outset of this article, a number of situations arise wherein the use of algorithms in decision-making may be prohibited. The IHRL framework assists in determining what those situations might be, and whether a prohibition on the use of algorithms in a particular decision-making context is absolute, or temporary, ie until certain deficiencies are remedied. This question should first be addressed during the conception phase, before actual design and development is undertaken, but should also be revisited as the algorithm develops through the design and testing phase and into deployment. In this regard, we envisage a number of scenarios in which the use of algorithms in decision-making would be contrary to IHRL.

1. Prohibition of the use of algorithms to circumvent IHRL

First, and most straightforwardly, IHRL prohibits the use of an algorithm in decision-making if the purpose or effect of its use would circumvent IHRL. This may occur if the intent of using an algorithm is to unlawfully discriminate against a particular group or if the effect of an algorithm is such that it results in indirect discrimination, even if unintentionally. In such scenarios, the use of the algorithm in decision-making would be prohibited as long as discriminatory effects exist. However, as discussed above, these effects could be overcome, if identified in the conceptualization and design phase or

[118] Ruggie Principles (n 14) Principles 11, 12.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

through internal and/or external oversight processes, and the algorithm modified and refined to remove any discriminatory bias, although a remedy would still be required to any individuals adversely affected.

A recent study conducted by researchers at Stanford University exemplifies this point. This study focused on how deep neural networks can extract facial features. The authors hypothesized that neural networks are better at detecting, interpreting, and perceiving facial cues than the human brain. The test involved comparing how deep neural networks performed compared to humans in determining sexual orientation from a set of facial images stored on a dating site. The study assumed that the sexual orientation of the individual could be inferred from the gender of the partners they were looking for on their dating profile. They concluded that 'deep neural networks are more accurate than humans at detecting sexual orientation from facial images'.[119] The authors argued that they carried out the research to generate public awareness about the risks that technology could be used in this way.[120]

The study was heavily criticized in the media, and by academics and civil society.[121] A key point was whether technology should be used for the purpose of determining a person's sexual orientation, particularly as it could result in individuals and communities being targeted for abuse, and possibly put their lives at risk.[122] In general, human determination regarding the sexual orientation of another person is prohibited since IHRL emphasizes self-identification regarding sexual orientation as integral to one's personality, and fundamental to self-determination, dignity and freedom.[123] If technology is deployed to carry out a task which would be prohibited if carried out by a human, it follows that the deployment of the technology would also be prohibited. Of course, in cases where discriminatory effects are identified and the algorithm can be modified accordingly to remove those effects, it is possible that the algorithm may then be deployed. Any affected individuals remain entitled to a remedy.

[119] Y Wang and M Kosinki, 'Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation from Facial Images' (2018) 114(2) Journal of Personality & Social Psychology 246, 254.

[120] H Murphy, 'Why Stanford Researchers Tried to Create a "Gaydar' Machine" (*New York Times*, 9 October 2017) <https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html?_r=0>.

[121] S Levin, 'LGBT Groups Denounce 'Dangerous' AI that Uses Your Face to Guess Sexuality' (*The Guardian*, 9 September 2017) <https://www.theguardian.com/world/2017/sep/08/ai-gay-gaydar-algorithm-facial-recognition-criticism-stanford>; D Anderson, 'GLAAD and HRC Call on Stanford University & Responsible Media to Debunk Dangerous & Flawed Report Claiming to Identify LGBTQ People through Facial Recognition Technology' (GLAAD Blog, 8 September 2017) <https://www.glaad.org/blog/glaad-and-hrc-call-stanford-university-responsible-media-debunk-dangerous-flawed-report>.

[122] See Anderson (n 121).

[123] The Yogyakarta Principles: Principles on the Application of International Human Rights Law in Relation to Sexual Orientation and Gender Identity (March 2007) Principle 3.

2. Prohibition of the exclusive use of algorithms to make certain decisions

Second, IHRL may prohibit certain decisions that are made exclusively on the basis of an algorithm, without the possibility of human intervention. In cases where an individual's rights are interfered with by a decision involving algorithms, the underlying reasoning must be made on the basis of factors specific and relevant to that individual. This derives from the prohibition of arbitrary rights interference as a core principle underpinning IHRL and is therefore relevant to all decisions that have the potential to interfere with particular rights.[124]

Modern algorithms raise issues in the context of arbitrariness as, given the nature of big data-driven algorithms, (a) decisions may be based on group-level characteristics, ie x members of a group are likely to behave in a particular way, as opposed to individually-focused characteristics, ie a specific individual is likely to act in a particular way because of factors specific to that individual, and (b) decisions are often based on correlation and not causation. These two factors are interrelated. They indicate that analysis vis-à-vis likely future behaviour is valid only at the group and not at the individual level, and that predictions are not determinative as to how a specific individual will act.[125] These models fail to account for individual agency, and the relevance of individual choice. This raises concerns that algorithmic decisions applied to individuals may, in certain cases, be inherently inconsistent with the prohibition of arbitrary interference with rights.[126]

These characteristics suggest that while algorithms may be used as a piece of evidence within a decision, they cannot provide the sole basis for a decision that directly affects an individual's rights: some form of human involvement or oversight is necessary. For instance, the application of IHRL indicates that a sentencing, bail or parole decision can never be decided exclusively by an algorithm. These decisions directly affect an individual's right to liberty, a central component of which is the prohibition of the arbitrary deprivation of liberty.[127] This requires, amongst other factors, that detention, or continued detention, be based upon reasons specific to the individual in question.[128] The nature of algorithmic decision-making inherently precludes this possibility as analysis is conducted on the basis of group behaviour, and

[124] See eg UN Human Rights Committee, 'General Comment No. 34, Article 19: Freedoms of Opinion and Expression' (12 September 2011) UN Doc CCPR/C/GC/34, paras 21–22, 24–30, 33–35; UN Human Rights Council, 'Report of the UN High Commissioner for Human Rights on The Right to Privacy in the Digital Age' (n 16) para 10; *Zakharov v Russia* (n 104) para 230; *Khan v The United Kingdom* App No 35394/97 (ECtHR, 12 May 2000) para 26; *Kroon and Others v The Netherlands* App No 18535/91 (ECtHR, 27 October 1994) para 31.

[125] The role of human agency is also an important consideration, noting that individuals may change their behaviour in unexpected—and unpredictable—ways.

[126] See discussion Part IIA.

[127] UN Human Rights Committee, 'General Comment No. 35, Article 9: Liberty and Security of Person' (16 December 2014) UN Doc CCPR/C/GC/35, para 10.

[128] ibid, paras 12, 15, 18, 22, 25, 36.

correlation not causation.[129] Exclusive reliance on algorithmic decision-making in this context must be considered arbitrary and therefore prohibited.[130]

B. Safeguards Required to Permit the Use of Algorithms

In other situations, whether the use of an algorithm within a decision is compatible with IHRL may depend upon the safeguards embedded within the process, including the level and type of human involvement. These safeguards are centred on ensuring that an algorithm operates effectively, within acceptable parameters. For instance, social media companies are currently engaged in efforts to moderate content in light of the alleged use of their platforms to promote terrorism or propagate hate speech.[131] This is a difficult task that directly brings the right to freedom of expression into play; there is a danger that the removal of posts may be inconsistent with IHRL—a post may be offensive but not illegal—and therefore violate the poster's right to freedom of expression.[132] Given the complexity of rights at issue, algorithms are used to flag, filter and classify certain content, but teams of human moderators ultimately decide how content or a particular user account should be managed. In this case, human input acts as a form of safeguard intended to ensure that content is not wrongly restricted or removed by algorithms.

Another key consideration is not only whether safeguards are in place but also whether they are able to operate effectively. This issue may be demonstrated by considering the role of the 'human in the loop'. A human decision-maker is considered to be 'in the loop' when they are involved in the decision-making process, for instance when they make a decision informed by an algorithmic output or when they provide oversight in relation to the algorithmic decision-making process.[133] However, the presence of a human operator does not guarantee that safeguards are effective. For example, questions arise about the operator's ability to meaningfully understand the algorithmic decision-making process, their capacity to determine whether and how any human rights have been affected, and the extent to which they automatically or subconsciously defer to the algorithmic decision. Deference may arise, for example, due to perceptions of the neutrality and accuracy of technology and concerns about going against

[129] As discussed in Part IIA.

[130] This appears to be the conclusion reached by the Wisconsin Supreme Court in *Wisconsin v Eric L. Loomis*, discussed further in Part IVB.

[131] UN Human Rights Council, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression on Freedom of Expression, States and the Private Sector in the Digital Age' (11 May 2016) UN Doc A/HRC/32/38, paras 35–37.

[132] S Cope, JC York and J Gillula, 'Industry Efforts to Censor Pro-Terrorism Online Content Pose Risks to Free Speech' (Electronic Frontier Foundation, 12 July 2017) <https://www.eff.org/deeplinks/2017/07/industry-efforts-censor-pro-terrorism-online-content-pose-risks-free-speech>.

[133] See Part IIA.

the findings of such technology.[134] This creates a risk that algorithms become the *de facto* sole decision-maker, even if there is apparently some human input.

A recent decision of the Wisconsin Supreme Court in the USA regarding the use of algorithmic risk assessments in sentencing decisions demonstrates this point. In *State of Wisconsin v Eric L. Loomis*, the Court had to assess whether the use of an algorithmic risk assessment tool to determine if the defendant could be supervised within the community rather than detained violated the defendant's right to due process.[135] As the Court noted,

> risk scores are intended to predict the general likelihood that those with a similar history of offending are either less likely or more likely to commit another crime following release from custody. However, the COMPAS risk assessment does not predict the specific likelihood that an individual offender will reoffend. Instead, it provides a prediction based on a comparison of information about the individual to a similar data group.[136]

The defendant challenged the use of the risk assessment tool on the basis that the proprietary interest in the algorithm meant that he could not challenge its 'scientific validity'[137] or 'accuracy' because Northpointe (the company that owned the algorithm) 'does not disclose how the risk scores are determined or how the factors are weighed'.[138] He argued that this denied him 'an individualized sentence', and 'it improperly uses gendered assessments'.[139]

In this case, the Court found that a risk assessment tool could be used to inform a decision, but it could not be determinative.[140] On its face, this appears to be compatible with IHRL and the 'red line' test outlined above. However, the level of scrutiny applied when analysing how the algorithm reached its conclusions must also be addressed to determine if the safeguards were in fact effective.[141] As noted above, this relates to both the nature of the data inputs and how the algorithm uses that data.

The first key point is that the use of risk assessment tools was considered by the Court as part of a move towards evidence-based sentencing.[142] Framing risk assessment tools as evidence-based already presents the use of algorithms in decision-making as something more objective than other types of judgments, such as a judge's intuition or a correctional officer's standard practice.[143] Without entering into a wider discussion on the merits of this approach, it raises the perceived objectivity of algorithmic outcomes which is relevant

[134] See Kroll *et al.* (n 9) 680 fn 136; DK Citron, 'Technological Due Process' (2008) 85(6) WashLRev 1249, 1283–4.
[135] *State of Wisconsin v Eric L. Loomis* (n 42) para 7.
[136] ibid, para 15.
[137] ibid, para 6.
[138] ibid, para 51.
[139] ibid, para 34.
[140] ibid, para 88.
[141] See Science & Technology Committee (n 73) Q132 (Martin Wattenberg, Google, evidence arguing for a 'very high level of scrutiny' to the use of algorithms in the criminal justice system).
[142] *State of Wisconsin v Eric L. Loomis* (n 42) para 3.
[143] ibid, para 40.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

when considering the risk that judges and other decision-makers might defer to algorithms because they are technologically produced.[144]

Second, the Court acknowledged that a proprietary interest prevented the defendant understanding how the algorithm weighed and analysed the input data.[145] However, it found that the opportunity to challenge the risk score itself, as well as the input data (as relevant to him) was sufficient,[146] provided certain safeguards were in place such as information being provided to the Court on whether (1) a cross-validation study had been conducted, (2) the scores 'raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism', and (3) the tools are 'monitored and re-normed for accuracy due to changing populations and subpopulations'.[147] The Court asserted that this would enable courts to 'better assess the accuracy of the assessment and the appropriate weight to be given to the risk score'.[148]

The information sought by the Court about potential bias appears to introduce levels of scrutiny to how the risk scores will be weighted. However, it provides no indication regarding how this will be conducted in a way that ensures objectivity. It thus potentially undermines the evidence-based approach the Court seeks to achieve.[149] Moreover, the safeguards only introduce the potential of less weight being given to the risk score rather than excluding it as evidence. They still do not enable the defendant—or another body—to assess how information about him and other persons with similar profiles were weighed and what inferences were made to produce an eventual risk score. Therefore, the Court appeared to consider only the input data and the overall outcome relevant to due process. However, as argued above, the way in which the algorithm itself functions can lead to violations of human rights, and must be addressed.

Thus, even if not determinative, the algorithmic decision in this case had a significant bearing on one of the most important human rights, the right to liberty. This case demonstrates the risk that courts and other bodies may pay an unwarranted level of deference to algorithms, meaning that while a human may be 'in the loop', their role in the loop may actually be minimal.[150] In areas such as sentencing and bail applications, there may be a greater deference to algorithms where actors are concerned about going against the findings of an algorithm, in case they are then blamed if a person released goes on to

[144] See eg O'Neil (n 32) (critiquing the assumption that big data algorithms are objective and fair). [145] See *State v Loomis* (n 42) para 54. [146] ibid, para 53, 55–6. [147] ibid, para 66. [148] ibid.

[149] See *Finogenov & Others v Russia* App Nos 18299/03 and 27311/03 (ECtHR, 4 June 2012), para 270, the court stated that 'the materials and conclusions of the investigation should be sufficiently accessible'.

[150] This is contrary to the minimum standards of thoroughness and effectiveness for investigations that demands adequate and rigorous analysis of all relevant elements by competent relevant professionals. See ibid, para 271; UN Human Rights Committee, General Comment No. 31 (n 13) para 15.

commit a crime. On one level, this is understandable, as the human decision-maker is unlikely to want to 'go against the computer' and then explain their reasoning for doing so should they 'get it wrong'. However, it is precisely these situations that involve fundamental rights, such as the right to liberty, and that therefore require particular protection.[151]

C. *Responsibilities in Areas in Which Algorithmic Effect Is Not Possible to Predict*

Finally, the increasing complexity of algorithms and their possible future autonomy may mean that it is difficult for humans to predict the impact they will have. For example, others have asked, '[w]hat happens when algorithms write algorithms? Algorithms in the past have been created by a programmer … In the future they will likely be evolved by intelligent/learning machines. We may not even understand where they came from'.[152] This has resulted in debate over whether humans should have reduced responsibility for the actions of algorithms if humans cannot predict what algorithms will do.[153] At the conceptualization stage, therefore, actors may claim that they cannot predict whether or not the algorithm will result in human rights interferences or violations.

In these circumstances, the actor is taking the decision to use an algorithm in full knowledge that it cannot predict the effect it will have. From an IHRL perspective, this does not automatically reduce the level of responsibility. This is because human entities (primarily in the form of States or businesses) make the decision to design and deploy algorithms. These actors remain subject to an obligation to ensure that their use of algorithms does not result in human rights violations, even if unintended. Thus, under IHRL, blanket assertions of reduced responsibility would be rejected; if the specific outcome of an algorithmic decision cannot be predicted, the parameters within which a decision is made should nonetheless be clear. For example, if in the conception or design phase actors claim that they cannot predict how the algorithm might perform or anticipate risks that might arise because of the complexity and sophistication of the particular algorithm, this would not reduce responsibility. Rather, if the actor decided to proceed with the use of the algorithm, such a decision might actually result in heightened responsibility. Thus, the IHRL framework pushes back on the developing discourse on reduced human responsibility or distributed responsibility between humans

[151] See *Finogenov & Others v Russia* (n 149) para 271 and *Husayn (Abu Zubadyh) v Poland* App No 7511/13 (ECtHR, 24 July 2014) para 480—in both cases the Court stated that "a requirement of a "thorough investigation" means that the authorities must always make a serious attempt to find out what happened and should not rely on hasty or ill-founded conclusions to close their investigation or as the basis of their decisions'; *Paul & Audrey Edwards v UK* App No 46477/99 (ECtHR, 14 March 2002) para 71; *Mapiripán Massacre v Colombia*, Judgment, Inter-American Court of Human Rights Series C No 134 (15 September 2005) para 224.

[152] See Rainie and Anderson (n 1) 55.

[153] See Mittelstadt *et al.* (n 50) 11–12.

https://doi.org/10.1017/S0020589319000046 Published online by Cambridge University Press

and machines as the starting point.[154] IHRL responds to this unpredictability by requiring actors to build in human rights protections so that where an algorithm acts unpredictably, safeguards are in place. This includes the proposals by some scientists to explore whether human rights and other ethical principles could be 'baked into' the algorithmic process so that the algorithm would act according to these norms and also be capable of alerting the human supervisor to problems.[155]

V. CONCLUSION

Existing approaches to algorithmic accountability are important, but of themselves do not address the full complexity of the issue. IHRL can provide an appropriate framework that takes into account the overall algorithmic life cycle, as well as the differentiated responsibility of all the actors involved. Adopting an IHRL framework can: take advantage of both current and future approaches to prevention, safeguards, monitoring and oversight, and remedy; incorporate broadly accepted understandings as to the conduct that constitutes 'harm'; and provide guidance with respect to the circumstances in which algorithmic decision-making may be employed. Mapping the algorithmic life cycle against the human rights framework provides clear red lines where algorithms cannot be used, as well as necessary safeguards for ensuring compatibility with human rights. Overall, it strengthens the protections for individuals who are caught in a power imbalance against entities that rely on technologically advanced algorithmic decision-making tools, as it ensures that responsibility is exercised and not deferred.

As stated at the outset, this article is intended to facilitate a discussion as to the role of IHRL in relation to the design, development, and deployment of algorithms, and to provide guidance as to how the IHRL framework can substantively inform this process. Although IHRL does not currently establish binding obligations on business enterprises, it requires States to address third-party harm and establishes clear expectations of businesses, as set out in the UN Guiding Principles on Business and Human Rights, and this area of IHRL continues to evolve. It is these measures that businesses should apply if they are to comply with human rights, and in order to ensure that they 'do no harm'.

[154] See L Floridi and JW Sanders, 'On the Morality of Artificial Agents' (2004) 14(3) Minds & Machines 349, 351; GD Crnkovic and B Çürüklü, 'Robots: Ethical by Design' (2012) 14(1) Ethics & Information Technology 61, 62–3; A Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata' (2004) 6(3) Ethics & Information Technology 175, 177.

[155] For example, in some industries where the use of automated technology is more developed such as the use of autopilot programmes in aviation, errors in operation can still give rise to responsibility of the pilot and/or liability of the company.

This article has focused on presenting an approach capable of informing the decision-making process – as it relates to the entirety of the algorithmic life cycle – and on providing guidance as to the steps that States and businesses should take to avoid human rights violations. As such, a discussion regarding the role of established human rights mechanisms has been outside the scope of this article. However, these mechanisms can play a critical role in operationalizing existing IHRL so that it makes a contribution and can help shape and address current gaps in algorithmic accountability. Of particular importance at this nascent stage of the discussion are those mechanisms that can help to inform the development of a human rights-based approach and facilitate its incorporation into mainstream discussions. Key in this regard are the Office of the UN High Commissioner for Human Rights and the Special Procedures established by the Human Rights Council. As algorithmic decision-making potentially affects all human rights, a joint statement by a number of UN Treaty Bodies may also be appropriate.[156] Regional human rights mechanisms and Treaty Bodies will also play an important role in addressing suspected violations arising in this regard, and their case law can assist as regards a deeper and more day-to-day understanding of the human rights law framework. This focus on international mechanisms should not, however, distract from the essential role played by national bodies, such as the independent oversight mechanisms discussed in Part IIIB.

Ultimately, much work remains to be done both to operationalize the IHRL framework and to then ensure that it is applied.

[156] See, for instance, a joint statement issued by the UN Human Rights Committee and the UN Committee on Economic, Social and Cultural Rights, addressing 50 years of the Covenants. Joint Statement by the UN Human Rights Committee and the UN Committee on Economic, Social and Cultural Rights, 'The International Covenants on Human Rights: 50 Years On' (17 November 2016) UN Doc CCPR/C/2016/1-E/C.12/2016/3.