Algorithmic Discrimination: How algorithms expose the limits of the ICCPR's anti-discrimination framework

Joseph Higgitt

Abstract

In recent decades, algorithms have become an unavoidable part of society. They are used to filter job applications, target police resources and identify terror threats. Whilst they undoubtedly carry great benefit, algorithms also raise serious human rights concerns, as improper use can lead to certain groups being disproportionately overlooked by employers, or minorities being over-policed by anti-terror authorities and police.

The International Covenant on Civil and Political Rights (ICCPR) nominally prohibits discrimination. However, algorithmic discrimination continues to challenge the ICCPR's anti-discrimination rules as, unlike other forms of discrimination, it often lacks the 'recognisability' essential to engaging ICCPR rights. This essay argues that the nature and use of algorithms give rise to three types of 'opacity' that make their decision-making process incomprehensible to most people, meaning that any potentially discriminatory decisions they make are unrecognisable and thus difficult to contest and rectify. This essay concludes by suggesting that a procedural 'right to explanation' that operationalises existing ICCPR rights may be necessary to prevent algorithmic discrimination.

Introduction

The emergence of new technologies has exposed the limits of the ICCPR and other similar rights frameworks.[1] State and private actors increasingly use algorithms in 'high stakes'

[1] International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 ('ICCPR').

situations, such as policing and recruitment, where outcomes have been proven to be discriminatory.[2] Although the ICCPR clearly prohibits discrimination, it still struggles to address algorithmic discrimination. This is because it is difficult to recognise algorithmic discrimination due to how incomprehensible algorithms are to the average person. This article will explore three ways in which 'algorithmic opacity' can manifest; namely incidental, deliberate and emergent opacity.

This 'algorithmic opacity' poses a problem for the ICCPR, as recognising that discrimination has occurred is necessary for initiating anti-discrimination procedures and holding relevant parties to account. As such, a procedural 'right to explanation,' that operationalises existing non-discrimination frameworks, is needed.

Discrimination Within Algorithms

Algorithmic decision-making has become ever more common in the modern world, including in 'high stakes' situations, where errors can cause tangible affects to individuals, such as whether a candidate receives an employment offer or whether a prisoner is offered parole. Algorithms often replicate human discriminatory behaviour due to the use of biased data. For example, Amazon used a machine-learning algorithm to review job applications by observing the patterns of successful resumes from previous years.[3] However, the company had a history of disproportionately selecting male candidates, so the algorithm learned to deprioritise resumes that contained the word 'women' (e.g. 'women's college' or 'women's chess club captain'), leading to fewer women being selected for clearly discriminatory reasons.[4]

---

[2] Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1 Nature Machine Intelligence 206, 206.

[3] Jeffrey Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women' (*Reuters*, 11 October 2018) <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/> accessed 28 June 2026.

[4] Dastin (n 3).

Similar problems occur within algorithms that governments use. Court systems in the U.S. use the COMPAS program (Correctional Offender Management Profiling for Alternative Sanctions) to determine whether a criminal is likely to reoffend should they be granted parole.[5] The program considers factors including the length of the offender's sentence.[6] A ProPublica investigation found that the program was far more likely to produce false positives for black offenders, leading to their parole being denied more often.[7] Even if COMPAS was not directly using race to determine recidivism, variables such as sentence length act as a proxy for race, since black offenders tend to receive harsher sentences than white offenders.[8] The ProPublica investigation exposed how algorithms can produce discriminatory results, making them a concern for human rights frameworks; a point supported by the current UN Special Rapporteur on racial discrimination, who has reported that AI algorithms have a 'huge capacity… to perpetuate and deepen systemic racial discrimination.'[9]

The Right to Non-Discrimination

Article 26 of the ICCPR provides individuals with a right to non-discrimination based on enumerated characteristics and places an obligation on states to guarantee 'effective protection' against discrimination.[10] This right is autonomous, meaning it does not require the invocation of any other ICCPR rights to take effect. The United Nations Human Rights Council (UNHRC) specifies that discrimination includes acts with either the purpose (direct discrimination) or effect (indirect discrimination) of impairing the rights of individuals based on the enumerated

---

[5] Rudin (n 2), 209.
[6] Rudin (n 2), 209.
[7] Jeff Larson and others, 'How We Analyzed the COMPAS Recidivism Algorithm' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> accessed 28 June 2026.
[8] Stephen Demuth and Darrell Steffensmeier, 'Ethnicity Effects on Sentence Outcomes in Large Urban Courts: Comparisons among White, Black, and Hispanic Defendants' (2004) 85 Social Science Quarterly 994, 1008.
[9] United Nations General Assembly (UNGA), 'Contemporary forms of racism, racial discrimination, xenophobia and related intolerance' (3 June 2024) UN Doc A/HRC/56/68, para 6.
[10] ICCPR, Article 26.

characteristics.[11] Article 2(3) of the ICCPR requires states to provide 'effective remedy' in instances where violations of ICCPR rights occur and to properly enforce those remedies.[12]

Articles 2(3) and 26 mandate that states have both a negative and positive duty to prevent and remedy discrimination, which includes discrimination committed by private actors. The UNHRC accepts this understanding, explicitly interpreting Article 26 to include a duty on states to prevent 'acts committed by private persons or entities that would impair the enjoyment of Covenant rights,' including 'in fields affecting basic aspects of ordinary life such as work or housing.'[13] This framework prohibits most discrimination, including the algorithmic discrimination previously discussed, which has the effect of discrimination irrespective of intent.

Assumption of Recognisability

Although Article 26 of the ICCPR imposes these clear positive obligations on states to prevent discrimination, an act of discrimination must first be recognised as such for states to engage the anti-discrimination procedures mandated by Article 26. This is what this paper shall refer to as an 'assumption of recognisability.' It often falls on the victims of discrimination to recognise that they have been discriminated against, and it is usually their responsibility to initiate the proceedings that could lead to the issue being resolved. This is implicit within the ICCPR's Optional Protocol, which refers to the 'individual' who is complaining to the UN Human Rights Committee about violations of their ICCPR rights.[14] Alternatively, states may proactively initiate anti-discrimination procedures as part of their duty to provide 'effective

[11] UN Human Rights Committee (UNHRC), 'CCPR General Comment No. 18: Non-discrimination' (10 November 1989) UN Doc A/45/40 (Annex VI), paras 6-7.
[12] ICCPR, Article 2(3).
[13] UNHRC, 'General Comment No. 31 [80]: The Nature of the General Legal Obligation Imposed on State Parties to the Covenant' (26 May 2004) CCPR/C/21/Rev.1/Add.13, para 8.
[14] Optional Protocol to the International Covenant on Civil and Political Rights (adopted 16 December 1966, entered into force 23 March 1976) 999 UNTS 171 ('OPICCPR').

protection,' but this still relies on the state recognising the potential for discrimination to take remedial action. Simply, if no one recognises that discrimination has occurred, then it cannot be combatted, meaning individuals become incapable of accessing the effective remedies prescribed by Article 2(3).

How Algorithmic Opacity Impedes Rights

Algorithmic discrimination is often not recognisable due to the opacity of algorithmic decision-making. This opacity can mainly be categorised into incidental opacity, deliberate opacity and emergent opacity. Though it is noted that this paper is not the first to categorise opacity, and there are other categories of opacity that this paper does not cover.[15] However, this author has selected these three categories of opacity because they are the most relevant to how algorithms can interfere with rights frameworks like the ICCPR.

The following sections will analyse how these types of opacity create an uneven epistemic playing field between those who employ algorithms and those who are subject to its decisions, placing victims at a disadvantage in recognising algorithmic discrimination. This epistemic gap creates a barrier between individuals and the effective remedies demanded by ICCPR Articles 2(3) and 26, suggesting that algorithmic discrimination is not effectively covered by the ICCPR.

Incidental Opacity

Firstly, this article will explore 'incidental opacity,' which is caused by an individual's technical and statistical illiteracy, leading to their ability to recognise algorithmic discrimination being impeded. To understand algorithmic decision-making and recognise discrimination, one must have at least a basic level of such literacy. Recognising discrimination

[15] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 1, 1.

does not require an individual to fully comprehend algorithms and statistics, but some understanding does allow them to identify whether harm created by algorithms is systematic and how algorithms base that harm on protected characteristics.

Some basic statistical literacy is essential in understanding how capable algorithms are in using data, on top of the general limits of data in making equitable decisions.[16] Some technical literacy is also crucial; algorithms are built out of code and easy explanations of how a particular algorithm functions are often not available, especially for machine-learning algorithms which are more complex than traditional algorithms.[17] At present, the design of algorithms is a 'highly specialised skill' that is beyond the reach of the layperson.[18] The general population's statistical literacy is also limited.[19]

Furthermore, statistical and technical literacy correlate strongly with social factors, such as gender, age and education level.[20] This is concerning as those less likely to be statistically and technically literate, such as women and the non-university educated (who have significantly higher rates of poverty), are more vulnerable to discrimination and have fewer resources to combat it.[21]

This poses a problem for a state's ICCPR obligations as they are required to take positive action to ensure individuals have 'accessible and effective remedies' to rights violations.[22] The technical and statistical knowledge barrier to recognising algorithmic discrimination presents

---

[16] Fernando Marmolejo-Ramos and others, 'Factors Influencing Trust in Algorithmic Decision-Making: An Indirect Scenario-Based Experiment' (2025) 7 Frontiers in Artificial Intelligence 1, 4-5.
[17] Burrell (n 15), 4.
[18] Burrell (n 15), 4.
[19] Office for Statistics Regulation, 'Statistical Literacy: Research' (Office for Statistics Regulation, 9 February 2023) <https://osr.statisticsauthority.gov.uk/publication/statistical-literacy-research/pages/4/> accessed 9 January 2026, para 3.
[20] Office for Statistical Regulation (n 19), para 3; Maximillian Eder and Helle Sjøvaag, 'Artificial Intelligence and the Dawn of an Algorithmic Divide' (2024) 9 Frontiers in Communication 1, 7.
[21] Joseph Rowntree Foundation, 'UK Poverty 2025: The essential guide to understanding poverty in the UK' (Joseph Rowntree Foundation 2025) 140.
[22] UNHRC, 'General Comment No.31' (n 13), para 8.

a clear example of remedies being inaccessible. As such, states must work to resolve this. For example, states could provide education that allows individuals to understand algorithms. However, this is unfeasible for an immediate solution and is far more resource-intensive than the issue requires (though Article 13(1) of the International Covenant on Economic, Social and Cultural Rights (ICESCR) does impose a duty on states to provide education that strengthens 'respect for human rights,' which may apply in this case).[23] Alternatively, states could require explanations for algorithmic decisions in a way that is intelligible to most people, including those with limited technical and statistical knowledge. This constitutes part of the 'right to explanation' that I will explore later.

Deliberate Opacity

Secondly, 'deliberate opacity,' is caused by the purposeful obfuscation of how an algorithm functions by those who employ them. Both states and corporations value secrecy in their algorithms. There are valid reasons why this should be the case; we would not want terrorists to be able to evade detection because they know what factors national security are looking for, and corporations may have a legitimate desire to shield trade secrets from competitors to maintain a market advantage.[24] The proprietary nature of algorithms, such as COMPAS, has already been used as a defence for protecting their code in courts, to prevent competitors from replicating their algorithm.[25] However, there are also less palatable reasons for algorithmic secrecy; states and corporations may wish to use algorithms to intrusively collect data to predict behaviour.

---

[23] International Covenant on Economic, Social and Cultural Rights (adopted 16 December 1966, entered into force 3 January 1976) 993 UNTS 3 ('ICESCR'), para 13(1).

[24] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015), 2-4

[25] *State v Loomis*, 2016 WI 68, 371 Wis. 2d 235, 881 N.W.2d 749, [66].

In either case, such secrecy makes it difficult for individuals to recognise when discrimination is taking place and when they should engage their ICCPR rights. Take the terrorism example – whilst it is true that it would be undesirable for a terrorist to know what factors a 'terrorist-detection algorithm' is using, it is also undesirable for the public not to know what those factors are. How can the public be sure that the algorithm is not using race as a factor, leading to counterterrorist agencies policing and intruding upon the lives of racial minorities? If this were happening, it would certainly constitute an Article 26 violation, as ICCPR Article 4(1) prohibits discriminatory measures. States cannot use 'national security' to justify these measures – even in national emergencies. Yet it is impossible to contest the algorithm when 'national secrets' protect it; this is a violation of the 'effective remedies' Article 2(3) guarantees.

Corporate algorithmic opacity can also take a different shape, related to the emergent opacity mentioned earlier, but in a deliberate form. Rudin explains that corporations sometimes purposefully choose the algorithmic opacity seen in black box machine learning (ML) models over more transparent alternatives. This is because there is currently a greater profit motive in complex black box models, which are proprietary, than in simple transparent algorithms.[26] Multiple jurisdictions across the U.S. use COMPAS and pay hundreds of thousands in yearly licencing fees; the State of Michigan alone paid over $2 million from 2016-2019.[27] Despite these large fees, COMPAS performs equally to a free, open-source code (CORELS) that only considers four variables for recidivism.[28] According to a 2018 study, COMPAS performs similarly to humans with no experience in criminal justice, and has comparable levels of

[26] Rudin (n 2), 209-210.
[27] Michigan Department of Technology, Management and Budget, Contract 7700014 (11 September 2024), 55, 127.
[28] Rudin (n 2), 209.

discriminatory outcomes.[29] However, these two alternatives are not nearly as profitable for corporations, so they instead push for expensive black box models.

Deliberate opacity raises clear ICCPR concerns about a state's obligation to provide an 'effective remedy' for discrimination and to regulate the behaviour of private actors. It also raises concerns about existing power structures; secrecy perpetuates inequality in who can access their rights by allowing powerful actors, such as states and corporations, to relinquish their responsibility to remedy discrimination. This undermines the purpose of ICCPR rights entirely, whose purpose is to guarantee protections from these powerful actors.

Emergent Opacity

Thirdly, there is 'emergent opacity.' This is opacity caused by the innate incomprehensibility of algorithms built on machine learning (ML) models. Such algorithms have become 'black boxes' of information, with some having 'internal states… of millions of interdependent values,' becoming impossible for any human to understand why it has made any decision.[30] This contributes to the non-recognisability of algorithmic discrimination, as there is no way to know if a machine learning algorithm has made a decision based on discriminatory factors.

For example, take the candidate-selection algorithm from earlier; a traditional rule-based algorithm (where a designer selects variables) would be easy to track how it came to any decision by looking at the variables it has used. Thus, if the algorithm had used any directly (such as gender) or indirectly (such as postcode) discriminatory factors, it would be easy to recognise, contest and remedy. But if it were an ML algorithm, a human could not trace a decision back to any individual causal factors, only complex interactions between internal

---

[29] Julia Dressel and Hany Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4 Science Advances 1, 3.

[30] Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI' [2019] Proceedings of the Conference on Fairness, Accountability, and Transparency 279.

representations of variables that the algorithm derives from its training data. This makes it difficult to determine if a decision was discriminatory.

There are indirect methods of interpreting the decision-making process algorithms built on ML models, but these are flawed. 'Explainable' ML models apply a second model to the original 'black box,' replicating most of its behaviour but in a human-intelligible format.[31] There are several criticisms leviable against explainable models, particularly the unfaithful replication of data.

Explainable models cannot entirely replicate the data of the original black box (or the original model would be redundant) and so interpretation errors are possible.[32] Even low inaccuracy rates are unacceptable for use in 'high-stake' situations – even small errors could, for example, lead to bail/parole being denied to a protected class of people.[33] The ProPublica analysis of COMPAS shows how explainable models can fail. ProPublica used an explainable model on COMPAS to determine what factors it uses to produce results, determining race is an important factor.[34] However, it is possible that ProPublica's model unfaithfully replicated COMPAS' data, deriving race from proxy variables that correlate.[35]

It is not that any individual accusation of discrimination produced by an explainable model is or is not true, as explainability errors do not negate the possibility of discrimination. Instead, there is no way of verifying that potential truth in a way that satisfies the threshold for reliability required of high-stake, rights-sensitive situations. Inherently interpretable models are mostly

[31] Rudin (n 2), 206-207.
[32] Rudin (n 2), 207-208.
[33] Rudin (n 2), 206, 207.
[34] Larson (n 7).
[35] Rudin (n 2), 208.

preferable to explainable models, and until a change to such models occurs, emergent opacity will remain a problem for discrimination recognisability.[36]

Emergent opacity puts strain on ICCPR rights in a way that is unique to this type of opacity. Claims of Article 26 breaches rely on proving that differential treatment is based on prohibited grounds such as protected characteristics, but emergent opacity obscures the causal chain back to such grounds, and even explainable models cannot reliably eliminate this obscurity. Any claims of discrimination by ML algorithms then struggle to meet the standard legal burdens of proof. This differs from the other two forms of opacity, in which the causal chain between discriminatory factors and outcomes is clearly identifiable, at least to the operators of the algorithms, or someone with sufficient technical wherewithal.

The limits of this technology's explainability do not absolve states of their Article 26 duties. A UN report on AI systems affirmed states' duty to combat algorithmic discrimination by improving access to remedies.[37] Since access to remedies cannot exist without discrimination being recognisable, states therefore also have a duty to make sure discrimination within ML algorithms is recognisable. This falls within the States' Article 2(3) positive obligations to provide effective remedies, including regulating private actors who use ML algorithms.[38] This may mean regulating algorithms used in high-stakes decisions to ensure they are transparent, or, if this proves unfeasible, ceasing their use until a transparent solution can be found.

A Right to Explanation

The ICCPR establishes both positive and negative obligations, as affirmed by the UNHRC; Article 2(3) requires states to take measures that give effect to their obligation to investigate

[36] Rudin (n 2), 211-214.
[37] UNGA, 'Human Rights and International Solidarity' (17 July 2024) UN Doc A/79/170, para 17.
[38] UNHRC, 'General Comment No. 31' (n 13), para 8.

allegations of rights violations.[39] This is the issue in algorithmic discrimination cases as the epistemic gap created by algorithmic opacity obfuscates when states should be engaging their obligation to investigate violations, rendering these obligations ineffective. However, the ICCPR currently lacks a framework that outlines what measures states should be taking to bridge the epistemic chasm. A right to explanation would rectify some of these problems. This right derives from the non-discrimination and effective remedies clauses of the ICCPR, and is not strictly a 'new right,' though the ICCPR does not currently explicitly outline it.

This section explores how a 'right to explanation,' as a positive obligation on state actors (grounded in ICCPR Article 2.3), could combat the issues presented by algorithmic discrimination. This 'right' would not be a standalone right, but auxiliary to the ICCPR's existing non-discrimination framework. It would not eliminate bias, but it would shift some of the burdens created by the epistemic gap inherent to algorithms away from individuals and towards state and private actors.

While several definitions exist, for the purpose of this paper, a 'right to explanation' is a procedural right requiring states to ensure those subject to algorithmic decision-making receive sufficiently intelligible information about how an algorithm reached a decision, to allow meaningful contestation and access to remedy in cases of Article 26 breaches. It is not a right to full transparency and will not provide access to source code or proprietary models as national security or market competition can provide legitimate grounds for states or corporations to want to keep their algorithms secret from the public. However, explanations must be as detailed as needed to satisfy legal thresholds for proof in discrimination cases. A right to explanation would make algorithmic decisions more intelligible to those who lack the technical and

[39] UNHRC, 'General Comment No. 31' (n 13), para 15.

algorithmic literacy necessary to recognise cases of discrimination against them. It would also eliminate some arbitrary deliberate opacity of algorithms.

However, it would not eliminate all discriminatory outcomes – these are more often the result of biased data and the structural inequalities it contains; it would only lift the veil that often hides the discrimination occurring, allowing remedial mechanisms to be engaged. Importantly for this paper, a right to explanation would not solve many of the issues with emergent opacity. A right to explanation requires states and corporations to provide intelligible reasons for why an algorithm has made a decision, but black box ML algorithms are incapable of producing such reasons; even well-designed explainable ML models are not fully reliable and lack methods of verifying their results, suggesting ML algorithms are not compatible with non-discrimination frameworks.

A full right to explanation also may not always be feasible, especially in information-sensitive domains, such as counterterrorism. In these instances, a right to explanation may need to be weighed proportionally – independent auditors for sensitive algorithms could be a satisfactory alternative when full explanation could be dangerous. A right to explanation does not fully solve the problems algorithms pose to existing rights frameworks, but it is necessary to restore the recognisability of discrimination in algorithms, which existing frameworks have thus far failed to do.

A right to explanation is not a substantive right; it is a procedural measure designed to support the effective enforcement of other procedural and substantive ICCPR rights, such as Article 2(3) and 26 respectively. In this sense, it is not a ‘new’ right to prevent discrimination, as no normative gap exists in the ICCPR that allows any sort of discrimination.

Conclusion

The purpose of this paper was to discover whether new technologies expose the limits of existing human rights frameworks, necessitating new rights. This paper used algorithms as an example of a technology that could expose the limits of rights frameworks, specifically the ICCPR's non-discrimination protections in Article 26 and its demand for effective remedies in Article 2(3). It concluded that algorithms do, in fact, expose the limits of the non-discrimination protections by undermining ICCPR's reliance on victims being able to recognise and contest discrimination. Furthermore, this paper found that a right to explanation would help to address some of the issues presented by algorithms by operationalising pre-existing ICCPR obligations rather than creating new rights.

There is much ground this paper did not cover – other emerging technologies, such as mass surveillance and neurotechnology, similarly expose deficiencies in rights frameworks, suggesting that these frameworks must operationally adapt across a range of areas. Ultimately, this analysis suggests that focus should be on supplementing and clarifying existing obligations rather than creating a new generation of rights.