Auditing public sector algorithms for discrimination: a framework for human rights law applied to the CNAF's fraud risk scoring algorithm

Word count: 5000

Candidate number: 1087056

Table of Contents

Introduction

Anaïs is a recently divorced 31-year-old mother-of-two. She has struggled with employment following maternity leave and a divorce. She has been on welfare for over a year, receives disability benefits, and her household's income per member is €500. Despite her situation, Anaïs is very careful to ensure she only claims the child benefits she is entitled to from the State.

Hugo and Marie are married parents of two. Stably employed for several years and earning €140,000 per year between them, they have high disposable income having inherited a house from Hugo's parents. French law nevertheless entitles them to claim child benefits, which Hugo has found he can fraudulently overclaim, defrauding the state of thousands of euros.

Anaïs receives a letter telling her she will be visited by an investigator from the Caisse Nationale des Allocations Familiales (CNAF), the French state agency responsible for the country's social security system. While most of her paperwork is in order, she is told she owes the CNAF €1500 because of financial assistance she received from her sister allowing her to travel to visit their dying mother.[1] In a case of honest mistake, she did not report this gift to the CNAF so that it could be deducted from her welfare benefit allowance. She will struggle to repay this debt given she is already having trouble making ends meet.

Anaïs is ignorant to the fact that the reason she was investigated in the first place had nothing to do with evidence that she was committing fraud. It was because she was a poor, disabled, recently divorced, young single mother who was flagged by a predictive algorithm which claims to calculate a beneficiary's risk of committing fraud. Conversely, despite continually committing fraud, Hugo and Marie are never flagged courtesy of their able-bodiedness, wealth and stable family situation.

The CNAF's is just one of a rapidly growing number of predictive algorithmic risk scoring systems being introduced by governments worldwide, especially in Europe. Investigations have repeatedly found these to infringe on citizens' rights through discriminatory outcomes

[1] Illustrative example based on Sénécat et al. (2023)

and breaches of their right to private life (Davidson et al., 2022; Romain et al., 2023). Such automated probabilistic prediction is by its very nature biased. This is inescapable and part of its technical and mathematical identity, especially when it is built to operate in a world whose reality is the product of centuries of bias. Machine learning algorithms, which most of these are based on, use data about group behaviour to predict the behaviour of an individual. The algorithmic prediction is never truly be tailored to specific individuals within the group, no matter the size or quality of the input dataset. Despite this, these predictive algorithmic models, the CNAF's included, are widely used to make individually focused decisions whose subjects are often completely unaware they are subject to algorithmic decision-making (McGregor et al., 2019).

There is thus an urgent need to ensure legal protection is afforded to citizens such that their right not to be arbitrarily discriminated against extends to public sector algorithmic practices. Human rights law provides a means to achieving this, but the nature of algorithmic practices makes these cases deeply complex. This paper therefore outlines an auditing framework specifically designed to guide the evaluation of predictive public sector algorithms' compliance with human rights law, with a focus on the European Convention on Human Rights' non-discrimination regime. It will be applied to the CNAF algorithm as an illustrative case study.

This paper's focus is heavily European due to the heavy use of welfare algorithms in Europe and because of the depth with which its human rights regime has been developed. Analysis will be centred on the ECHR, and especially Art.14, because it has received little attention compared with EU non-discrimination law. Though the latter is more widely applicable, I argue the ECHR has important potential in the public sector algorithmic context.

The CNAF algorithm

The CNAF's predictive welfare fraud risk scoring algorithm has been subject to extensive campaigning by French digital rights groups (La Quadrature du Net, 2022a). The full source code for previous versions of the algorithm (2010-2014 and 2014-2020) was eventually disclosed in 2023 after successful arguments made in front of the Committee on Access to Administrative Documents (CADA), although the nature of 3 out of 35 variables were

redacted for the 2014-2020 code (La Quadrature du Net, 2023a). While the nature of the redacted variables is not conclusively known, external evidence suggests one of them is the 'socio-economic characteristics of the beneficiary's neighbourhood', suggesting residence in an underprivileged neighbourhood negatively influences the risk score (La Quadrature du Net, 2023b). While further investigation is necessary to evaluate whether it is the case in this context, the ability for neighbourhood to serve as a proxy for race, religion, or national origin is well documented (Pan Ké Shon, 2009; Sambasivan et al., 2021).

The 2014-2020 algorithm scores risk using 35 input variables, each of which is assigned a coefficient value determined using a simple undisclosed logistic-regression-based supervised machine learning model. This non-disclosure makes it impossible to audit the validity of the coefficient values allocated to each variable value (Romain et al., 2023). Despite the availability of the source code, the non-availability of the coefficient-generating algorithm means the most appropriate technical auditing method was not a code audit but a functional audit using different test profiles and varying variable values to examine how coefficients affect output risk scores.

Key findings include that having a high income dramatically reduces risk scores, while having a low income and receiving benefits due to low income increases risk scores substantially. Being younger increases risk scores, as -significantly- does receiving disability benefits. Being a parent, especially a parent of older children, markedly increases risk scores, as does being divorced or having a new partner (Romain et al., 2023). As QdN remarked, this demonstrates how the CNAF's algorithm has thus not been designed to identify suspicious behaviour and use that to generate risk scores. As is generally the case with predictive algorithms, it is not evidence-based but instead uses people's personal characteristics, some of which are discriminatory, to label them as fraud risks purely on the basis of correlation between other factors and past documented cases of welfare fraud (La Quadrature du Net, 2023a).

An outline of the law

Public sector algorithmic practices are subject to an extensive range of legal regimes that do not bind the private sector, notably international human rights law and administrative law.

There is merit in Engstrom & Ho (2020)'s claim that administrative law is 'the body of law that is most likely to negotiate the collision of technology and the administrative state'. However, its potential role is underexplored and in its current state in most jurisdictions is such that it remains ill-equipped and requires serious rethinking to address the challenges posed by the practices documented in this paper. Human rights law, conversely, is better placed to be put into immediate action and can already form the basis of domestic judicial review proceedings.

Under international human rights law, states have a positive legal obligation to take 'adequate' measures to prevent human rights violations (*Al-Skeini and Others v. The United Kingdom*, 2011). Any decision or action taken by a public authority must comply with the rights and freedoms protected by the ECHR (*Airey v. Ireland*, 1979). Furthermore, public authorities have a positive obligation to ensure contracted private entities comply with human rights standards (*Costello-Roberts v. The United Kingdom*, 1993). Public authorities' decisions made relying in whole or in part on an algorithm must therefore comply with the ECHR, including its provisions on discrimination. Even if a rights-abusing public sector algorithm is developed by a private third party, responsibility for its right-compliance thus falls on the state.

Human rights law is flexible and widely applicable by design, allowing it to be applied to the algorithmic context (Zuiderveen Borgesius, 2020). As such, the legality of European states' public sector algorithmic practices can be examined in light of their impact on -among others- the right to fair trial (Art.6 ECHR), the right to privacy (Art.8 ECHR), the right to freedom of expression (Art.10 ECHR), the right to freedom of assembly and association (Art.11 ECHR), the right to an effective remedy (Art.13 ECHR), and the prohibition of discrimination (Art.14 ECHR) (Wagner et al., 2017). As McGregor et al. (2019) argue, human rights law thus "offers a framework through which algorithmic accountability can be situated" by providing "the means to analyse when the use of algorithms in decision-making could contribute to, or result in, harm, even if unintentionally" and by establishing obligations on states in relation to the identification and protection against such harm. It offers a pre-existing legal framework for assessing algorithms and -at least in the European context- provides a number of legal instruments and remedies for citizens to contest their use. While only alleged victims can bring complaints having exhausted domestic remedies in their member state, this paper's

contribution applies equally to the domestic human rights regimes those states have implemented in line with the ECHR.

This paper focuses its attention on the prohibition on discrimination in Article 14 ECHR. While debates about algorithmic fairness, bias and discrimination abound in computer science, ethics, and social sciences, they have not yet been taken up as widely in European law, especially in practice. Indeed, even when the question has been asked in court, it has generally been avoided. In the *SyRI* case, part of an algorithmic scandal which ultimately led to the Dutch national government to resign after admitting its algorithmic practices were discriminatory, the court's decision turned on Article 8, the right to private life, avoiding consideration of the much more conceptually complex question of how it might apply non-discrimination law to algorithmic decision-making under Article 14. This was despite that question being at the centre of the case having explicitly been argued by the plaintiffs (Rachovitsa & Johann, 2022; Van Bekkum & Zuiderveen Borgesius, 2021).

When deciding whether unlawful discrimination has taken place, the Court tests whether there has been (1) a difference in treatment of persons in (2) analogous or relevantly similar situations (3) based on one of the prohibited grounds of discrimination. If so, it examines (4) whether this difference is objectively justified by questioning (a) whether it pursues a legitimate aim, and (b) whether the means employed are reasonably proportionate to the aim pursued (European Court of Human Rights, 2022). Art.14 generally places the initial burden of proof on the claimant who must provide evidence capable of showing that there has been a difference in treatment based on one of the grounds listed in Article 14 taken in conjunction with another Convention provision (*E.b. V. France*, 2008). However, and of important potential relevance to questions involving algorithmic decision-making, the burden of proof *can be reversed* where the events in issue lie wholly, or in large part, within the exclusive knowledge of the public authority and where "it would be extremely difficult in practice for the applicant to prove discrimination" (*Cînța v. Romania*, 2020; *Salman v. Turkey*, 2000). In such cases, the claimant must only make out a prima facie case, or even just an arguable allegation, of discrimination for the burden of proof to rest on the state (European Court of Human Rights, 2022).

Art.14 ECHR is generally ancillary in nature, requiring examination in conjunction with a substantive provision of the Convention. Protocol No.12 explicitly prohibits discrimination

‘by any public authority’, thereby encompassing all exercises of discretionary power by public authorities using algorithms, but it only has 20 ratifications. Nevertheless, Art.14 can be applied to many public sector algorithmic practices without relying on Protocol No.12. This includes the CNAF’s case of social security benefits: where an individual has an assertable right under domestic law to a welfare benefit, the importance of that interest is taken to amounts to a possession for the purposes of Art.1 of Protocol No.1 (*Andrejeva v. Latvia*, 2009). Legislation providing for the payment of a welfare benefit as of right is considered to generate a proprietary interest which can only be denied in a manner compatible with Art.14, that is, free from discrimination (*Stec and Others v. The United Kingdom (dec.)*, 2005).

Non-discrimination law prohibits direct discrimination, where a protected attribute is the basis of differential treatment, but also indirect discrimination, where a practice, rule or decision which appears externally neutral but in fact disproportionally discriminates against a protected class, whether intentionally or not (*Biao v. Denmark*, 2016; *D.h. And Others v. The Czech Republic*, 2007). Indirect discrimination is particularly relevant to the algorithmic context because of the widespread use of non-protected proxies which have been shown to lead to de facto discrimination of protected classes. The focus on practice, not intention, is also pertinent: algorithms’ biased or discriminatory outputs and outcomes most often result from negligence or a lack of care.

While human rights law certainly applies to and addresses a number of the bias and discrimination issues that arise from the public sector use of algorithmic systems in decision-making capacities, the extent to which it is able to address *all* of these is debated. Non-discrimination statutes only apply to protected classes or characteristics, such as ethnicity, gender, or sexual orientation (e.g. European Court of Human Rights, 2022). Though these can be interpreted widely (*Carson and Others v. The United Kingdom*, 2010), limits remain. As Wachter (2019) demonstrates, algorithmic systems can differentiate on the basis of invented classes, sometimes incomprehensible to the human mind, which fall outside the scope of non-discrimination law but have arbitrarily discriminatory effect leading to disparities between non-protected groups (Wachter et al., 2020). While the following discussion does not deny this weakness in current non-discrimination law, the existing law does nevertheless catch a large amount of problematic algorithmic practices which discriminate against protected groups in harmful ways.

This overview demonstrates a simultaneous strength and complexity of European non-discrimination law: its commitment to what has been described as a 'contextual equality' approach (Wachter et al., 2020). Rather than lending itself to systematisation and straightforward certainty, it has chosen contextual agility and the ability to investigate the nuances of each case and context on a case-by-case basis. This paper will demonstrate how, with the assistance of some guidelines, this gives it great potential to evaluate the possible unlawfully discriminatory nature of some public sector algorithmic practices.

The transparency question

Public authorities commonly go to great lengths to maintain opacity about their algorithmic decision-making processes. Information about SyRI and the CNAF's algorithm only made it into the public sphere after hundreds of Freedom of Information (FOI) requests, petitions, and lobbying (Geiger, 2023; La Quadrature du Net, 2022b). These were only achieved through relentless, expensive, and time-consuming efforts undertaken by journalistic organisations and NGOs. This is far from the systematic transparency necessary for effective algorithmic scrutiny.

Such deliberately opaque practices are antithetical to state accountability, threaten democratic legitimacy, and undermine the rights of citizens (Fenster, 2015; Fox, 2007). Without insight into an algorithm's code, how it was designed, what data it was designed on, what results it outputs, and how these are used, an effective audit is impossible, as the Dutch court found in *SyRI*. Because European human rights law is claim-based, relying on individuals to bring claims against states to protect their human rights, effective human rights protection relies on citizens realising they may have been subject to discrimination. In the context of algorithmic decision-making, which is far more impersonal and which many are unaware is even taking place, the intuition that one has been discriminated against is far less likely to arise than in other contexts (Wachter et al., 2020).

Transparency by design is thus essential, as is reasonable notification that one has been subjected to a decision made in part or in full based on an algorithm. The United Kingdom's Algorithmic Transparency Recording Standard, which will soon require all central

government departments to provide information on algorithmic tools and algorithm-assisted decisions illustrates how this might be partially achieved (United Kingdom Secretary of State for Science, Innovation and Technology, 2024).

Public sector algorithmic transparency is often further challenged because the development of public sector algorithms is widely outsourced to private companies who vigorously defend purported trade secrets (Angwin et al., 2016). This adds a layer of complexity because most FOIA regimes include an exemption for information concerning technical details of products which would objectively be of major economic importance to the private company (Olsen et al., 2024). As Zuiderveen Borgesius (2020) suggests, it may be necessary for the law to improve research exemptions, requiring private organisations to disclose certain information relating to algorithms they have developed for public sector applications to researchers (or courts or regulatory institutions conducting audits) upon request. Where there are justified reasons for keeping an algorithm's workings outside the public domain, it should nevertheless be possible for an independent, regulatory, or court-ordered auditor to access the necessary information to conduct an audit whose findings are made public.

Auditing algorithms for unlawful discrimination

European non-discrimination law's contextual approach, combined with the wide range of areas different types of public sector decision-making can be incorporated in, means any auditing framework for the purposes of non-discrimination law must necessarily be similarly broad and flexible. This section sets out a framework to structure and guide algorithmic audits for the purpose of applying Art.14 ECHR. It pays particular attention to the role of statistical fairness tests in this process while underlining the importance of taking a holistic auditing approach that looks beyond the technical features of the algorithm and includes an evaluation of the wider algorithmic practice.

Comparability

The test for whether two situations are relevantly similar or analogous for the purposes of Art.14 is whether they are 'comparable' in all relevant respects. It is both specific and contextual, with the two situations considered as a whole and avoiding artificially marginal aspects (*Fábián v. Hungary*, 2017). The situations must be similar in terms of the facts and circumstances surrounding the treatment, the purpose of the treatment, and the impact of the treatment on the individuals involved (*D.h. And Others v. The Czech Republic*, 2007). The difference between the groups must be made out along the lines of a prohibited ground.

Comparability claims can be bolstered through analysis of input variables. In the CNAF case, where wealth appears to have a significant influence, an appropriate comparator group might be determinable by looking at people with similar inputs for: 'age', 'months since last email sent to CNAF', 'months since registration of internal information', 'family situation', 'months since affiliation with the CNAF', 'web connection within the last 18 months', 'presence of children aged 19 and over', and 'presence of children aged between 12 and 18' (Romain et al., 2023). People with identical or globally similar inputs for these variables but significantly different ones for income-related or -correlated inputs can form objectively comparable groups. As such, while variable examination must not be the only way comparability is assessed in algorithmic cases, it can simplify the court's task of judging group comparability.

Variables or combinations of variables can amount to proxies for characteristics, however. An analysis which takes a protected characteristic as a starting point (e.g. national origin) and examines how that characteristic affects output is therefore also relevant. In such a situation, controlling for some variables can nevertheless also help make the case for comparability. The test comparability test remains contextual, but these features of algorithmic decision making may be used as evidence in favour or against comparability.

Difference in treatment based on a prohibited grounds of discrimination

Predictive algorithms pose unique conceptual challenges to the identification of differential treatment because discrimination is their signature technical characteristic by design. Our

societies are built on historical patterns of bias and discrimination and different subgroups have different lived experiences resulting in statistical group-level behavioural differences. Useful and effective predictive algorithms identify these differences and make predictions accordingly. Wherever such a difference is present, the demographic parity fairness metric is not achieved, and a form of differential treatment occurs at the algorithmic level. In the CNAF case, poor and disabled people are more likely either to make a mistake or commit a fraudulent act in relation to welfare they receive. Accordingly, the predictive fraud risk algorithm will treat them differently *because they are poor or disabled*, because being poor or disabled is correlated with overclaiming welfare benefits. Identifying demographic disparity is where the auditing of many predictive public sector algorithms, including Lighthouse Reports' on CNAF's, stops (Romain et al., 2023). Non-discrimination law requires more.

Demographic parity is just one of many ways algorithmic treatment can be statistically evaluated. Predictive accuracy is another: there, equal treatment is achieved when algorithms are equally accurate across protected groups (e.g. 99% accuracy for blacks and whites). However, while a test being more accurate for one group than another *could* result in differential treatment, it is more important to evaluate *how the inaccuracy operates*, and whether it *does* lead to factual differential treatment of a protected group (Hellman, 2020). A non-discrimination audit must recognise that algorithms themselves will always treat different groups differently, but that different ones will do so across different dimensions. Different fairness metrics show how this is always the case, and how analytical focus must be on how (and whether) differential algorithmic treatment translates to differential treatment by the public authority.

An array of other useful and conceptually sound 'fairness' metrics have been developed to evaluate different dimensions of algorithmic (un)fairness at both group and individual levels (see Wachter et al., 2021 for a non-exhaustive list). Several studies have demonstrated that no real-world algorithm can satisfy all (or even a majority) of fairness metrics, and that optimising for one will often worsen outcomes for another (Berk et al., 2017; Corbett-Davies et al., 2023; Garg et al., 2020). While this means no single test can be used to assess the nature of an algorithm's differential treatment, it plays into the strengths of European non-discrimination law. Audits for the legal evaluation of algorithms require a nuanced and context-sensitive analysis, incorporating the results of several fairness metrics to evaluate whether and more importantly *how* treatment is differential (Kleinberg et al., 2016).

The error rate balance metric is especially important because it provides a more granular evaluation of treatment by examining the difference between the true positive rate (TPR) and (separately) the false positive rate (FPR) across the chosen groups (Chouldechova, 2016; Hardt et al., 2016; Hellman, 2020). Error ratio disparity suggests more profound differential treatment by highlighting how the burden of false negatives and false positives falls differently on different groups: the higher the disparity, the greater the differential treatment. The issue with predictive algorithms, however, is that differential treatment of this type is present wherever there are cross-group differences in base rates of the predicted variable: higher proportions of errors arise for the group(s) with higher rates (Corbett-Davies & Goel, 2018).

In the COMPAS case, even though there was predictive parity across groups, false positives outweighed false negatives for blacks while false negatives outweighed false positives for whites (Angwin et al., 2016). The inaccuracy's burden depends not only on degree but also on its nature: false positives might impact the individual by leading to investigation or the refusal of bail, while false negatives impacting society at large through cases of recidivism or continued welfare fraud. Both of these establish the possibility for substantially differential treatment to occur. In the CNAF case, demonstration that the poor or disabled are subject to more false positives and fewer false negatives than the wealthy and the able-bodied would imply a different dimension of differential algorithmic treatment on the basis of these protected attributes.

But differential algorithmic treatment alone is insufficient for the purposes of Art.14 unless its alleged violation is in conjunction with Art.8, which seems unlikely in our CNAF case. To be unlawful, it must lead to differential treatment by a public authority that affects an individual's enjoyment of the rights and freedoms guaranteed by the Convention. This is where the audit must step back from focusing narrowly on the technical features of the algorithm to take a holistic approach, examining the algorithmic practice as a whole including how it is incorporated into the public authority's decision-making process and how this influences the treatment of an individual or group (McGregor et al., 2019).

Relevant questions to be asked by the auditor include the nature of the decision being made, the role the algorithm plays in that decision, the extent to which the algorithm is

determinative of a person's treatment by the public authority, and whether safeguards have been put in place to minimise the chance of biased algorithmic output leading to differential treatment. Again, case-by-case analysis is necessary here. The more directly the algorithm's output influences the decision, the more the outcome of fairness metric evaluation will evidence difference in treatment. The presence of a human 'in the loop' by no means necessarily safeguards against carryover of the differential treatments identified in multi-metric fairness analysis, however, and the effectiveness of those safeguards must always form part of an algorithmic audit for non-discrimination law's purposes (McGregor et al., 2019).

In our CNAF example, the differential algorithmic outcomes for poor and disabled people means a greater proportion of them are investigated by the CNAF and a greater proportion of these investigations are due to false positives. While profiling for investigation based on a group attribute been considered differential treatment in conjunction with Art.8, this has only so far been the case in conjunction with Art.8 and in relation to race, which the ECtHR has emphasised is treated with 'special vigilance' (*Basu v. Germany*, 2022). This question therefore needs to be addressed by the court. The fact that the CNAF's algorithmic output leads to investigation, not the automated withdrawal of welfare provision, is relevant here. Where an individual's rights are interfered with (here the right to receive welfare), the underlying reasoning must be made on the basis of factors specific and relevant to that individual, but that is the case for the CNAF's process. The investigations do not appear take the fact of having been flagged as a factor in their findings and decisions, unlike COMPAS outputs which were a direct factor in judges' decisions. This makes the 'human-in-the-loop' component of the CNAF's algorithm relatively prominent, acting as a safeguard against differential treatment in terms of the impact of false positive algorithmic outputs. Algorithmic input does not inform the CNAF's ultimate decision so much as it initiates the CNAF's decision-making process.

In our CNAF case, the nature of the wider process, which relies entirely on investigation findings to determine welfare fraud (or mistakes) means the only way the algorithm's differential treatment of groups materialises as substantive differential treatment by the public authority beyond profiling for investigation is with respect to false negatives which mean the rich and able-bodied are more likely to get away with mistakes or fraud.

Objective and reasonable justification

Art.14 does not prohibit differences in treatment which are "founded on an objective assessment of essentially different factual circumstances and which, being based on the public interest, strike a fair balance between the protection of the interests of the community" and Convention rights (*Zarb Adami v. Malta*, 2006). Given the use of predictive algorithms in decision-making will in a majority of cases involve a degree of difference in treatment, a case-by-case application of the objective justification proportionality test will almost always be necessary.

While there is no specific threshold for the degree of differential treatment required to establish a violation of Article 14, the degree and nature of the differential treatment is implicitly central to the proportionality test. As such, fairness metric analysis will be highly relevant at this stage. If the CNAF's algorithm for instance returns 30% of false positives for disabled people but only 5% for able-bodied people in otherwise comparable situations, it will be more difficult for the respondent to establish that the measure is necessary in a democratic society. The proportionality test should incorporate evaluation of the disparate nature of error burdens and other dimensions of fairness highlighted by fairness metric analysis to ascertain whether the benefit to society is legitimate and proportionate to the aim pursued.

The wider algorithmic process is also relevant at this stage. The proportionality test will be harder to pass if safeguards are weaker or the nature of the wider algorithmic practice is such that false negatives lead to further and more substantial differential treatment. That the CNAF uses risk scores to profile for investigation, for instance, may require less justification than it would if they also served as evidence in an investigatory decision leading to the withdrawal of welfare assistance.

Conclusion

The focus of this paper has been the legal side of fairness and accountability in algorithmic practices in the public sector. Its findings centre specifically on how audits might determine whether *unlawful* discrimination (or unlawful unfairness) is present. Importantly, algorithmic practices that are not found to be unlawfully discriminatory through this framework can still be unethical and deeply problematic. If anything, an evaluation of them may help identify areas human rights or other legal regimes may need to address in the algorithmic context.

Evaluating the case of the CNAF's predictive fraud risk scoring algorithm, this paper has shown how the ECHR's non-discrimination regime is particularly relevant and applicable to public sector algorithmic practices. As a majority of predictive public sector algorithms do by design, the CNAF's treats different protected groups -disabled and poor people in this case- differently across several statistical metrics. However, the way it is incorporated into the CNAF's fraud investigation process means the extent to which it leads to differential treatment by the public authority is not as important as some other algorithms. Accordingly, the CNAF is likely to be able to defend its algorithmic practice through objective and reasonable justification because the State's margin of appreciation is relatively wide when it comes to determining legitimacy and proportionality on social or economic grounds (*Andrejeva v. Latvia*, 2009).

Bibliography

Airey v. Ireland, 78103/14 (ECtHR 9 October 1979). https://hudoc.echr.coe.int/eng?i=001-57420

Al-Skeini and Others v. the United Kingdom, 55721/07 (ECtHR [GC] 7 July 2011). https://hudoc.echr.coe.int/fre?i=001-105606

Andrejeva v. Latvia, 55707/00 (ECtHR [GC] 18 February 2009). https://hudoc.echr.coe.int/fre?i=001-91388

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Basu v. Germany, 215/19 (ECtHR 18 October 2022). https://hudoc.echr.coe.int/eng?i=001-220007

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). *Fairness in Criminal Justice Risk Assessments: The State of the Art* (arXiv:1703.09207). arXiv. http://arxiv.org/abs/1703.09207

Biao v. Denmark, 38590/10 (ECtHR [GC] 24 May 2016). https://hudoc.echr.coe.int/fre?i=001-163115

Carson and Others v. the United Kingdom, 42184/05 (ECtHR [GC] 16 March 2010). https://hudoc.echr.coe.int/eng?i=001-97704

Chouldechova, A. (2016). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments* (arXiv:1610.07524). arXiv. http://arxiv.org/abs/1610.07524

Cînța v. Romania, 3891/19 (ECtHR 18 February 2020). https://hudoc.echr.coe.int/eng?i=001-201533

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). *The Measure and Mismeasure of Fairness* (arXiv:1808.00023). arXiv. http://arxiv.org/abs/1808.00023

Corbett-Davies, S., & Goel, S. (2018). *The Measure and Mismeasure of Fairness* (arXiv:1808.00023v2). arXiv. http://arxiv.org/abs/1808.00023v2.pdf

Costello-Roberts v. the United Kingdom, 13134/87 (ECtHR 25 March 1993). https://hudoc.echr.coe.int/eng?i=001-57804

Davidson, D., Geiger, G., Schot, E., Hijink, M., Adriaens, S., Bulman, M., Konijn, J., Woude, A. van der, Hekman, L., & Howden, D. (2022, December 20). The Algorithm Addiction. *Lighthouse Reports*. https://www.lighthousereports.com/investigation/the-algorithm-addiction/

D.h. and Others v. the Czech Republic, 57325/00 (ECtHR [GC] 13 November 2007). https://hudoc.echr.coe.int/fre?i=001-83256

E.b. v. France, 43546/02 (ECtHR [GC] 22 January 2008). https://hudoc.echr.coe.int/eng?i=001-84571

Engstrom, D. F., & Ho, D. E. (2020). Algorithmic Accountability in the Administrative State. *Yale Journal on Regulation*, *37*.

European Court of Human Rights. (2022). *Guide on Article 14 of the Convention (prohibition of discrimination) and on Article 1 of Protocol No. 12 (general prohibition of discrimination)*. https://www.echr.coe.int/documents/d/echr/Guide_Art_14_Art_1_Protocol_12_ENG#:~:text="The%20enjoyment%20of%20the%20rights,%2C%20birth%20or%20other%20status."

Fábián v. Hungary, 78117/13 (ECtHR [GC] 5 September 2017). https://hudoc.echr.coe.int/eng?i=001-176769

Fenster, M. (2015). Transparency in search of a theory. *European Journal of Social Theory*, *18*(2), 150–167. https://doi.org/10.1177/1368431014555257

Fox, J. (2007). The uncertain relationship between transparency and accountability. *Development in Practice*, *17*(4–5), 663–671. https://doi.org/10.1080/09614520701469955

Garg, P., Villasenor, J., & Foggo, V. (2020). *Fairness Metrics: A Comparative Analysis* (arXiv:2001.07864). arXiv. http://arxiv.org/abs/2001.07864

Geiger, G. (2023, July 17). How We Did It: Unlocking Europe's Welfare Fraud Algorithms. *Pulitzer Center*. https://pulitzercenter.org/how-we-did-it-unlocking-europes-welfare-fraud-algorithms

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of Opportunity in Supervised Learning* (arXiv:1610.02413). arXiv. http://arxiv.org/abs/1610.02413

Hellman, D. (2020). Measuring algorithmic fairness. *Virginia Law Review*, *106*(1). https://virginialawreview.org/wp-content/uploads/2020/06/Hellman_Book.pdf

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores* (arXiv:1609.05807). arXiv. http://arxiv.org/abs/1609.05807

La Quadrature du Net. (2022a, October 19). *CAF: Le numérique au service de l'exclusion et du harcèlement des plus précaires*. La Quadrature du Net. https://www.laquadrature.net/2022/10/19/caf-le-numerique-au-service-de-lexclusion-et-du-harcelement-des-plus-precaires/

La Quadrature du Net. (2022b, December 23). *Notation des allocataires: Fébrile, la CAF s'enferme dans l'opacité*. La Quadrature du Net. https://www.laquadrature.net/2022/12/23/notation-des-allocataires-febrile-la-caf-senferme-dans-lopacite/

La Quadrature du Net. (2023a, November 27). *Notation des allocataires: L'indécence des pratiques de la CAF désormais indéniable*. La Quadrature du Net. https://www.laquadrature.net/2023/11/27/notation-des-allocataires-lindecence-des-pratiques-de-la-caf-desormais-indeniable/

La Quadrature du Net. (2023b, December 12). *La Quadrature du Net / Algorithmes et Contrôle Social / caf · GitLab*. GitLab. https://git.laquadrature.net/la-quadrature-du-net/algo-et-controle/caf

McGregor, L., Murray, D., & Ng, V. (2019). International Human Rights Law as a Framework for Algorithmic Accountability. *International and Comparative Law Quarterly*, *68*(2), 309–343. https://doi.org/10.1017/S0020589319000046

Olsen, H. P., Hildebrandt, T. T., Wiesener, C., Larsen, M. S., & Flügge, A. W. A. (2024). The Right to Transparency in Public Governance: Freedom of Information and the Use of Artificial Intelligence by Public Agencies. *Digital Government: Research and Practice*, *5*(1), 1–15. https://doi.org/10.1145/3632753

Pan Ké Shon, J.-L. (2009). Ségrégation ethnique et ségrégation sociale en quartiers sensibles. L'apport des mobilités résidentielles. *Revue française de sociologie*, *50*(3), 451–487. https://doi.org/10.3917/rfs.503.0451

Rachovitsa, A., & Johann, N. (2022). The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch *SyRI* Case. *Human Rights Law Review*, *22*(2), ngac010. https://doi.org/10.1093/hrlr/ngac010

Romain, M., Senecat, A., Pénicaud, S., Geiger, G., & Braun, J.-C. (2023, December 4). *How We Investigated France's Mass Profiling Machine*. Lighthouse Reports. https://www.lighthousereports.com/methodology/how-we-investigated-frances-mass-profiling-machine/

Salman v. Turkey, 21986/93 (ECtHR [GC] 27 June 2000). https://hudoc.echr.coe.int/eng?i=001-58735

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining Algorithmic Fairness in India and Beyond. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 315–328. https://doi.org/10.1145/3442188.3445896

Sénécat, A., Geiger, G., & Pénicaud, S. (2023, December 4). Dans la vie de Juliette, mère isolée, précaire et cible de l'algorithme des CAF. *Le Monde*. https://www.lemonde.fr/les-decodeurs/article/2023/12/04/dans-la-vie-de-juliette-mere-isolee-precaire-et-cible-de-l-algorithme-des-caf_6203803_4355770.html

Stec and Others v. the United Kingdom (Dec.), 65731/01, 65900/01 (ECtHR [GC] 6 July 2005). https://hudoc.echr.coe.int/eng?i=001-70087

United Kingdom Secretary of State for Science, Innovation and Technology. (2024). *A pro-innovation approach to AI regulation – Government response to consultation*. Department for Science, Innovation and Technology. https://assets.publishing.service.gov.uk/media/65c1e399c43191000d1a45f4/a-pro-innovation-approach-to-ai-regulation-amended-governement-response-web-ready.pdf

Van Bekkum, M., & Zuiderveen Borgesius, F. (2021). Digital welfare fraud detection and the Dutch *SyRI* judgment. *European Journal of Social Security*, *23*(4), 323–340. https://doi.org/10.1177/13882627211031257

Wachter, S. (2019). *Affinity Profiling and Discrimination by Association in Online Behavioural Advertising* (SSRN Scholarly Paper 3388639). https://doi.org/10.2139/ssrn.3388639

Wachter, S., Mittelstadt, B., & Russell, C. (2020). *Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI* (SSRN Scholarly Paper 3547922). https://doi.org/10.2139/ssrn.3547922

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3792772

Wagner et al. (2017). *Algorithms and Human Rights* (DGI(2017)12; Committee of Experts on Internet Intermediaries (MSI-NET)). Council of Europe. https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

Zarb Adami v. Malta, 17209/02 (ECtHR 20 June 2006). https://hudoc.echr.coe.int/fre?i=001-75934

Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. *The International Journal of Human Rights*, *24*(10), 1572–1593. https://doi.org/10.1080/13642987.2020.1743976