15

Discriminatory AI and the Law

Legal Standards for Algorithmic Profiling

Antje von Ungern-Sternberg

I. INTRODUCTION

One of the great potentials of Artificial Intelligence (AI) lies in profiling. After sifting through and analysing huge datasets, intelligent algorithms predict the qualities of job candidates, the creditworthiness of potential contractual partners, the preferences of internet users, or the risk of recidivism among convicted criminals. However, recent studies show that building and applying algorithms based on profiling can have discriminatory effects. Hiring algorithms may be biased against women,[1] and credit rating algorithms may disfavour people living in poorer neighbourhoods.[2] Algorithms can set prices or convey information to internet users classified by gender, race, sexual orientation, or disability,[3] and predicting recidivism algorithmically can have a disparate impact on people of colour.[4]

While some observers stress the particular danger posed by discriminatory AI,[5] others hope that it might eventually end discrimination[6]. Before examining the particular challenges of discriminatory AI, one should keep in mind that human decision-making is also affected by prejudices and stereotypes, and that algorithms might help avoid and detect manifest and hidden forms of discrimination. Nevertheless, possible discriminatory effects of AI need to be assessed for several reasons. First, algorithms can perpetuate existing societal inequalities and stereotypes if they are trained with datasets that reflect inequalities and stereotypes. Second, algorithms used

[1] C O'Neil, *Weapons of Math Destruction* (2017) (hereafter 'O'Neil, *Weapons of Math Destruction*') 105 *et seq*; P Kim, 'Data-Driven Discrimination at Work' (2017) 58 *William & Mary Law Review* 857, 869 *et seq*.

[2] O'Neil, *Weapons of Math Destruction* (n 1) 141 *et seq*; J Allen 'The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining' (2019) 46 *Fordham Urban Law Journal* 219.

[3] J Angwin and T Parris, 'Facebook Lets Advertisers Exclude Users by Race' (*ProPublica*, 28 October 2016) www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race; A Kofman and A Tobin, 'Facebook Ads Can Still Discriminate against Women and Older Workers, Despite a Civil Rights Settlement' (*ProPublica*, 13 December 2019) www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement; N Kayser-Bril, 'Automated Discrimination: Facebook Uses Gross Stereotypes to Optimize Ad Delivery' (*AlgorithmWatch*, 18 October 2020) https://algorithmwatch.org/en/story/automated-discrimination-facebook-google/; S Wachter, 'Affinity Profiling and Discrimination by Association in Online Behavioural Advertising' (2020) 35 *Berkeley Technology Law Journal* 367 (hereafter Wachter, 'Affinity Profiling').

[4] J Angwin, J Larson, S Mattu, and L Kirchner, 'Machine Bias' (*ProPublica* 23 May 2016) www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (hereafter Angwin and others, 'Machine Bias').

[5] O'Neil, *Weapons of Math Destruction* (n 1). *Cf*. also K Zweig, *Ein Algorithmus hat kein Taktgefühl* (3rd ed., 2019) 211.

[6] J Kleinberg and others, 'Discrimination in the Age of Algorithms' (2018) 10 *Journal of Legal Analysis* 1 (hereafter Kleinberg and others, *'Discrimination in the Age of Algorithms*').

https://doi.org/10.1017/9781009207898.020 Published online by Cambridge University Press

by large companies or state agencies affect many people. Third, the discriminatory effects of AI have not been easy to detect and to prove until now. What's more, some of the predictions resulting from AI analysis cannot be verified. If a person does not obtain credit, then she can hardly prove creditworthiness; likewise, if an applicant is not hired, there is no way he can prove to be a good employee. Finally, algorithms are often perceived as particularly rational or neutral, which may prevent questioning of its results.

Therefore, this article offers an assessment of the legality of discriminatory AI. It concentrates on the question of material legality, leaving many other important issues aside, namely the crucial question of detecting and proving discrimination.[7] Drawing on legal scholarship showing discriminatory effects of AI,[8] this article analyses existing norms of anti-discrimination law,[9] depicts the role of data protection law,[10] and treats suggested standards such as a right to reasonable inferences[11] or 'bias transforming' fairness metrics that help secure substantive rather than mere formal equality.[12] This chapter shows that existing standards of anti-discrimination law already imply how to assess the legality of discriminatory effects, even though it will be helpful to develop and establish these aspects in more detail. As this assessment involves technical and legal questions, both lawyers as well as data and computer scientists need to cooperate. This article proceeds in three steps. After explaining the legal framework for profiling and automated decision-making (II), the article analyses the different causes for discrimination (III) and develops the relevant aspects of a legality or illegality assessment (IV).

[7] Some of the arguments developed in this chapter can also be found in A von Ungern-Sternberg, 'Diskriminierungsschutz bei algorithmenbasierten Entscheidungen' in A Mangold and M Payandeh (ed), *Handbuch Antidiskriminierungsrecht – Strukturen, Rechtsfiguren und Konzepte* (forthcoming) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3828696.

[8] *Cf.* n 1–6; B Friedman and H Nissenbaum, 'Bias in Computer Systems' (1996) 14 *ACM Transactions on Information Systems* 330(333 *et seq*) (hereafter Friedman and Nissenbaum, 'Bias in Computer Systems'); Calders and I Žliobaitė, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures' in B Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 43, 50 *et seq* (hereafter Calders and Žliobaitė, 'Unbiased Computational Processes'); S Barocas and A Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671, 681 *et seq* (hereafter Barocas and Selbst, 'Big Data's Disparate Impact'); C Orwat, *Diskriminierungsrisiken durch Verwendung von Algorithmen* (*Antidiskriminierungsstelle des Bundes*, 2019) 34 *et seq*, 77 *et seq* www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/Expertisen/studie_diskriminierungsrisiken_durch_verwendung_von_algorithmen.html (hereafter Orwat, *Diskriminierungsrisiken*).

[9] P Hacker, 'Teaching Fairness to Artificial Intelligence' (2018) 55 *Common Market Law Review* 1143; F Zuiderveen Borgesius, 'Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence' (2020) 24 *The International Journal of Human Rights* 1572; J Gerards and F Zuderveen Borgesius, 'Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence' (*SSRN*, 2020) https://ssrn.com/abstract=3723873 (hereafter Gerards and Zuderveen Borgesius, 'Protected Grounds'); Wachter, 'Affinity Profiling' (n 3); S Wachter, B Mittelstadt and C Russell, 'Why Fairness Cannot Be Automated' (*SSRN*, 2020) https://ssrn.com/abstract=3547922 (hereafter Wachter, Mittelstadt and Russell, 'Why Fairness Cannot Be Automated'); M Martini, *Blackbox Algorithmus: Grundfragen einer Regulierung Künstlicher Intelligenz* (2019) 73-91, 230-249.

[10] W Schreurs, M Hildebrandt, E Kindt, and M Vanfleteren, 'Cogitas, Ergo Sum. The Role of Data Protection Law and Non-discrimination Law in Group Profiling in the Private Sector' in M Hildebrandt and S Gutwirth, *Profiling the European Citizen* (2008) 241 (hereafter Schreurs and others, *Profiling*); I Cofone, 'Algorithmic Discrimination Is an Information Problem' (2019) 70 *Hastings Law Journal* 1389, 1416 *et seq* (hereafter Cofone, 'Algorithmic Discrimination'); S Wachter and B Mittelstadt, 'A Right to Reasonable Inferences' (2019) *Columbia Business Law Review*, 494 (hereafter Wachter and Mittelstadt, 'A Right to Reasonable Inferences'); A Tischbirek, 'Artificial Intelligence and Discrimination' in T Wischmeyer and T Rademacher (eds), *Regulation Artificial Intelligence* (2020) 104.

[11] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

[12] S Wachter, B Mittelstadt and C Russell, 'Bias Preservation in Machine Learning' *West Virginia Law Review* (forthcoming) https://ssrn.com/abstract=3792772 (hereafter Wachter, Mittelstadt and Russell, 'Bias Preservation').

II. LEGAL FRAMEWORK FOR PROFILING AND DECISION-MAKING

Using AI to profile involves different steps for which different legal norms apply. A legal definition of profiling can be found in the General Data Protection Regulation (GDPR). It 'means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.[13] Thus, profiling describes an automated process (as opposed to human instances of profiling, for instance by a police profiler) affecting humans (as opposed to AI optimising machines, for example) which increasingly relies on AI for detecting patterns, establishing correlations, and predicting human characteristics. Without going into detail about different possible definitions of AI,[14] profiling algorithms qualify as 'intelligent' as they can solve a defined problem, in other words, they can make predictions about unknown facts based on an analysis of data and patterns. After obtaining the profiling results on characteristics such as credit risk, job performance, or criminal behaviour, machines or humans may then make decisions on loans, recruiting, or surveillance. Thus, it is helpful to distinguish between (1) profiling and (2) decision-making. One can broadly assume that anti-discrimination law governs decision-making, whereas data protection law governs the input of personal data needed for profiling. A closer look reveals, however, that things are more complex than that.

1. *Profiling*

The process of profiling is comprised of several steps. The first step involves collecting data for training purposes. The second step entails building a model for predicting a certain outcome based on particular predictors (using a training algorithm). The final step applies this model to a particular person (using a screening algorithm).[15] Generally speaking, the first and the third steps are governed by data protection law because they involve the processing of personal data – either for establishing the dataset or for screening and profiling a particular person. The GDPR covers the processing of personal data by state actors and state parties alike, and requires that processing is based on the consent of the data subject or on another legal ground. Legal grounds can include necessary processing for the performance of a contract, compliance with a legal obligation, or for the purposes of legitimate interests.[16] Furthermore, the Law Enforcement Directive (LED) provides that the processing of personal data by law enforcement authorities must be necessary for preventing and prosecuting criminal offences or executing criminal penalties.[17] Thus, data protection law requires a sufficient legal basis for collecting and processing training data, as well as for collecting and processing the data of a specific person being profiled. Public authorities will mostly rely on statutes, while private companies will often rely on the necessity for the performance of a contract or base their activities on legitimate interests

[13] GDPR, Article 4(4).
[14] S Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (4th ed. 2022) 19–23.
[15] Schreurs and others, *Profiling* (n 10) 246; Kleinberg and others, *'Discrimination in the Age of Algorithms'* (n 6) 22.
[16] GDPR, Article 6(1)(a)(b)(c)(e) or (f). According to Article 2(2), the GDPR only applies to the processing of data 'by automated means' or if it forms part of a 'filing system' or is intended to form part of such a system. Thus, algorithmic (i.e. automated) forms of profiling fall under this heading.
[17] Article 8(1) Directive (EU) 2016/680 (LED). The GDPR does not apply to these activities of law enforcement authorities, *cf.* GDPR, Article 2 (2)(d).

or the consent of the data subjects. The processing of special ('sensitive') data, including personal data revealing racial or ethnic origin, political opinion, religious or philosophical beliefs, trade union membership, genetic data, biometric data for the purpose of uniquely identifying a natural person, and data concerning health or data concerning a natural person's sex life or sexual orientation, must comply with additional legality requirements.[18]

Yet, several questions remain. First, the second step, building the profiling model, is not covered by data protection law if the data is anonymised. Data protection law only applies to personal data, i.e. information relating to an identified or identifiable natural person.[19] Since it is not necessary to train a profiling algorithm on personalised data, datasets are regularly anonymised before the second step.[20] Some authors suggest that data subjects whose personal data have been collected during the first step should have the right to object to anonymisation, as this also constitutes a form of data processing.[21] However, even if this right exists for those cases when processing is based on consent, data subjects might not bother to object. Subjects may not bother to object either because they benefit from data collection, as in participating in a supermarket's consumer loyalty programme or internet web page access in exchange for accepting cookies, or because they are not immediately affected by the profiling. It is important to keep in mind that the data subjects providing training data (step one) may be completely different from the data subjects which are later profiled (step three).

Second, even during the first and the third step, it is not always clear whether personal data is being processed. Big data analysis can refer to all kinds of data. In a supermarket, for example, shopping behaviour can correlate not only with the date and time of shopping, but also with the contents and the movements (speed, route) of the shopping trolley. In an online environment, data ranging from online behaviour to keystroke patterns and the use of a certain end device may be linked to characteristics like price-sensitivity or creditworthiness. In this context, singling out a person as an individual, even if the data controller does not know the individual's name, should be enough to consider a person 'identifiable'.[22] Thus, cases where a company can recognise and trace an individual consumer or where a state agency can single out an individual fall under data protection law.

Third, it is disputed how the methodology of profiling and the profiling result (i.e. the profile of a particular person) should be treated in data protection law. It is helpful to distinguish different categories of data, notably collected data, like data submitted by the data subject or observed by the data controller, and data inferred from collected data, such as profiles.[23]

[18] GDPR, Article 9; LED, Article 10.

[19] GDPR, Article 4(1); LED, Article 3(1).

[20] Schreurs and others, *Profiling* (n 10) 248.

[21] Schreurs and others, *Profiling* (n 10) 248–253.

[22] *Cf.* that GDPR, Article 4(1) and LED, Article 3(1) also refer to an 'online identifier'; D Korff, 'New Challenges to Data Protection Study – Working Paper No 2: Data Protection Laws in the EU: The Difficulties in Meeting the Challenges Posed by Global Social and Technical Developments' (*European Commission DG Justice, Freedom and Security Report* 15 January 2010) https://ssrn.com/abstract=1638949, 45–48 (hereafter Korff, 'New Challenges to Data Protection'); Schreurs and others, *Profiling* (n 10) 247; F Zuiderveen Borgesius, 'Singling Out People without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation' (2016) 32 *Computer Law & Security Review* 256; F Zuiderveen Borgesius and J Poort, 'Online Price Discrimination and EU Data Privacy Law' (2017) 40 *Journal of Consumer Policy* 347 (356–358).

[23] Article 29 Data Protection Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling' WP251rev.01 (*Directorate C of the European Commission*, 6 February 2018) 8 https://ec.europa.eu/newsroom/article29/document.cfm?doc_id=49826; Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10) 516; R Broemel and H Trute, 'Alles nur Datenschutz' (2016) 27 *Berliner Debatte Initial* 50 (52).

Even though it is misleading to qualify inferred data as 'economy class' data,[24] inferred data is different from collected data in two regards. First, the methodology of inference varies considerably. Based on collected data, physicians diagnose medical conditions, lawyers assess the legality of acts, professors evaluate exams, journalists judge politicians, economists predict the behaviour of consumers, and internet users rate the service of online-sellers, each according to different scientific or value-based standards. Furthermore, one has to acknowledge that the inference itself is an accomplishment based on effort, values, qualifications, and/or skills. Profiling (i.e. algorithmic inferences about humans), also exhibits these two characteristics. Its distinct methodology is determined by its training and profiling algorithms, and its achievement is legally recognised, for example, by intellectual property protecting profiling algorithms[25] or by other rights like freedom of speech.[26]

This does not imply that predictions about characteristics and qualities of a particular person do not qualify as personal data. The Article 29 Data Protection Working Party, the precursor of today's European Data Protection Board, specified that data related to an individual if the data's content, result, or purpose was sufficiently linked to a particular person.[27] If a person's profile provides information about her (content), if it aims to evaluate her (purpose), and if using the profile will likely have an impact on her rights and interests (result), then the profile must be considered personal data.[28] However, the characteristics of inferred data can have an impact upon the data subject's rights. Notably, the right to rectification of inaccurate personal data[29] only refers to instances of inaccuracy which can be verified (e.g. the attribution of collected or inferred data to the wrong person). But the right generally does not include the appropriate (medical, legal, economic, et cetera) methodology of inferring information, as this is beyond the reach of data protection law.[30] This is the reason why scholars call for a right to reasonable interferences.[31] Yet, one might argue that profiling, as opposed to other methods of inferring data, is indeed, at least partially, regulated by data protection law.[32] In any event, profiling is not an activity privileged by the GDPR. The GDPR clauses promoting data processing for 'statistical purposes'[33] are not intended to facilitate profiling.[34] This follows from the wording of the

[24] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10) 494.

[25] GDPR, Recital 63; *cf.* BGHZ 200, 38 (BGH VI ZR 156/13) on the trade secret of Schufa, the German (private) General Credit Protection Agency, concerning its scoring algorithm.

[26] J Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 *UC Davis Law Review* 1149; note that GDPR, Article 85(1) demands that Member States reconcile data protection with the right to freedom of expression.

[27] Article 29 Data Protection Working Party, 'Opinion 4/2007 on the concept of personal data, 01248/07/EN WP 136' (*European Commission*, 20 June 2007) 9–12 https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf.

[28] Korff, 'New Challenges to Data Protection' (n 22) 52–53; Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10), 515–521.

[29] GDPR, Article 16; LED, Article 16.

[30] *Cf.* CJEU, Case C-434/16 *Nowak* [2017] n 52–57, on the right to rectification concerning written exams which does not extend to incorrect answers but possibly if examination scripts were mixed up by mistake.

[31] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

[32] See Section IV 3(a).

[33] GDPR, Articles 5(1)(b) and (e), 9(2)(j), 14(5)(b), 17(3)(d), 21(6), 89(1) and (2).

[34] This, however, is suggested by V Mayer-Schönberger and Y Padova, 'Regime Change? Enabling Big Data through Europe's New Data Protection Regulation' (2016) 17 *Columbia Sciences & Technology Law Review* 315 (330).

clauses, from Recital 162[35] and from the purpose of the GDPR, which is regulating profiling in order to control the risks emanating from it.[36]

2. Decision-Making

Anti-discrimination law and data protection law can govern the decisions that follow profiling.

a. Anti-Discrimination Law

Anti-discrimination provisions, grounded in national law, European Union law, and public international law, prohibit direct and (often) indirect forms of discrimination.[37] Some non-discrimination provisions address the state, while others are binding upon state and private actors. Some provisions have a closed list of protected characteristics, while others are more public.[38] Some provisions apply very broadly, covering employment or the supply of goods and services available to the public,[39] while still others have a narrower scope, merely affecting insurance contracts or management of journalistic online content, for example.[40] This chapter does not seek to examine the commonalities or differences of these provisions but rather aims to analyse if and when decision-making based on profiling may be justified.

This analysis is based on some general observations. First, anti-discrimination law applies to human and machine decisions alike. It does not presuppose a human actor. Thus, it is not relevant for anti-discrimination law whether a decision has been made solely by an algorithm, solely by a human being (based on the profile), or by both (i.e. by a human being accepting or not objecting to the decisions suggested by an algorithm). Second, anti-discrimination law distinguishes between direct and indirect discrimination, or between differential treatment and detrimental impact.[41] In EU anti-discrimination law, direct discrimination occurs when one person is treated less favourably than another is treated or would be treated in a comparable situation because of a protected characteristic such as race, gender, age, or religion.[42] Indirect

[35] '[. . .] Statistical purposes mean any operation of collection and the processing of personal data necessary for statistical surveys or for the production of statistical results. Those statistical results may further be used for different purposes, including a scientific research purpose. The statistical purpose implies that the result of processing for statistical purposes is not personal data, but aggregate data, and that this result or the personal data are not used in support of measures or decisions regarding any particular natural person.'

[36] Thus, the statistical privilege is only granted if public agencies conduct statistical surveys and produce statistical results, or if similar activities take place in the public interest (and not in support of profiling a particular natural person), *cf.* J Caspar, 'Article 89' in S Simitis, G Hornung, and I Spiecker gen Döhmann (eds), *Datenschutzrecht* (2019) n 23.

[37] Article 3 German Basic Law, German General Equal Treatment Act (2006); Article 21 EU Charter of Fundamental Rights (CFA), Framework Directive 2000/78/EC, Race Directive 2000/43/EC, Goods and Services Sex Discrimination Directive 2004/113/EC, Equal Treatment Directive 2006/54/EC; Article 14 European Convention on Human Rights.

[38] For an overview see European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European Non-discrimination Law* (2010) https://fra.europa.eu/sites/default/files/fra_uploads/1510-FRA-CASE-LAW-HANDBOOK_EN.pdf; M Connolly, *Discrimination Law* (2nd ed., 2011) 15, 55, 79, 151 (hereafter Connolly, *Discrimination Law*); Gerards and Zuderveen Borgesius, 'Protected Grounds' (n 9).

[39] Article 3(1) Framework Directive 2000/78/EC; Article 3(1)(c) and (h) Race Directive 2000/43/EC; Article 3(1) Goods and Services Sex Discrimination Directive 2004/113/EC; Article 14(1) Equal Treatment Directive 2006/54/EC.

[40] In German law, §19(1) n° 2 German General Equal Treatment Act (2006) contains a specific anti-discrimination norm for private insurance contracts; §94(1) of the new State Treaty on Media (2020) forbids big media platforms to discriminate between journalistic content.

[41] *Cf.* D Schiek, 'Indirect Discrimination' in D Schiek, L Weddington, and M Bell, *Non-Discrimination Law* (2007) 323 (372) (hereafter Schiek, 'Indirect Discrimination'). This is also known as disparate treatment and disparate impact in U.S. terminology.

[42] See e.g. Framework Directive 2000/78/EC, Article 2(2)(a).

https://doi.org/10.1017/9781009207898.020 Published online by Cambridge University Press

discrimination occurs when an apparently neutral provision, criterion, or practice would put members of a protected group at a particular disadvantage compared with other persons, unless this is justified.[43] Note the term 'discrimination' implies illegality in German usage, whereas differential treatment or detrimental effect can be legal if it is justified. However, this article follows the English use of the term 'discrimination' which encompasses illegal and legal forms of differential treatment or detrimental effect. Algorithmic profiling and decision-making can easily avoid direct discrimination if algorithms are prohibited from collecting or considering protected characteristics. However, if algorithms are trained on datasets reflecting societal inequalities and stereotypes (indicating, for instance, that men are better qualified for certain jobs than women), profiling and decision-making might put already disadvantaged groups (like female applicants) at a particular disadvantage. Thus, one can expect indirect discrimination to gain importance in an era of algorithmic profiling and decision-making. As a consequence, corresponding questions like "How can a particular disadvantage be established?"[44] or "What are the reasons for banning indirect discrimination?"[45] will become increasingly relevant.

Third, direct and indirect forms of discrimination, or differential treatment and detrimental effect, can be justified. Generally speaking, indirect discrimination is easier to justify than direct discrimination. In EU anti-discrimination law, indirectly causing a particular disadvantage does not amount to indirect discrimination if it 'is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary'.[46] But differential treatment can also be justified, either on narrow[47] or on broader[48] grounds, provided that it passes a proportionality test. Thus, considerations of proportionality are relevant for all attempts to justify direct and indirect forms of discrimination. This chapter submits that these considerations are significantly shaped by the commonalities of intelligent profiling and automation, as will be explained below.

b. Data Protection Law

Examining the legal framework for automated decision-making would be incomplete without Article 22 GDPR and Article 11 LED. These provisions go beyond a mere regulation of data processing by limiting the possible uses of its results. They apply to a decision 'based solely on automated processing, including profiling, which produces legal effects' concerning the data subject or 'significantly affect[ing] him or her'[49] and generally prohibit such a mode of automated decision-making unless certain conditions are met. Thus, the provisions also cover discriminatory decisions if they are automated. Furthermore, there is an explicit link between

[43] See e.g. Framework Directive 2000/78/EC, Article 2(2)(b).

[44] Wachter, Mittelstadt and Russell, 'Why Fairness Cannot Be Automated' (n 9) para V *et seq*.

[45] A Morris, 'On the Normative Foundations of Indirect Discrimination Law' (1995) 15 *Oxford Journal of Legal Studies* 199 (hereafter Morris, 'On the Normative Foundations'); C Tobler, *Limits and Potential of the Concept of Indirect Discrimination* (2008) 17–35 (hereafter Tobler, *Limits*); Connolly, *Discrimination Law* (n 38) 153–156.

[46] See e.g. Framework Directive 2000/78/EC, Article 2(2)(b)(i); Race Directive 2000/43/EC, Article 2(2)(b); Goods and Services Sex Discrimination Directive 2004/113/EC, Article 2(b); Equal Treatment Directive 2006/54/EC, Article 2(1)(b).

[47] The German Federal Constitutional Court, for example, accepted unequal treatment based on gender permissible only 'if compellingly required to resolve problems, that because of their nature, can occur only in the case of men or women' BVerfGE 85, 191 (BVerfG *1 BvR 1025/82*), Konrad-Adenauer-Stiftung, *70 Years German Basic Law* (3rd ed., 2019), 288.

[48] See e.g. Framework Directive 2000/78/EC, Articles 4 and 6; Goods and Services Sex Discrimination Directive 2004/113/EC, Article 4(5); CFR, Article 52(1) with regard to CFR, Article 21; DJ Harris and others, *Harris, O'Boyle and Warbrick: Law of the European Convention on Human Rights* (4th ed., 2018) 772–776 with regard to Art 14 ECHR (hereafter Harris and others, *European Convention on Human Rights*).

[49] GDPR, Article 22(1); LED, Article 11(1).

data protection and anti-discrimination law in Article 11 (3) LED, which prohibits profiling that results in discrimination against natural persons on the basis of special ('sensitive') data. A similar clause is missing in the GDPR, but the recitals indicate that the regulation is also intended to protect against discrimination.[50]

However, the scope and relevance of Article 22 GDPR are much debated. The courts have not yet established what 'a decision based solely on automated processing' means or what amounts to 'significant' effects.[51] Likewise, automated decision-making can still be based on explicit consent, contractual requirements, or a statutory authorisation as long as suitable measures safeguard the data subject's rights and freedoms and legitimate interests,[52] in other words, legal bases can also be understood in a restrictive or permissive way. The same applies to the anti-discrimination provision of Article 11(3) LED, which could extend to all forms, automated and human alike, of decision-making based on profiling (or be confined to automated decision-making) and which is open to different standards of scrutiny if differential treatment or factual disadvantages are justified.

3. *Data Protection and Anti-Discrimination Law*

The brief overview of relevant norms of data protection and anti-discrimination law shows that both areas of law are important in prohibiting and preventing discriminations caused by decision-making based on algorithmic profiling. Data protection law can be characterised not only as an end in and of itself, but also as a means to prevent discrimination based on data processing.[53] Such an understanding of data protection law flows from the recitals referring to discrimination,[54] from the special protection for categories of 'sensitive' data such as race, religion, political opinions, health data, or sexual orientation (which conform to the categories of protected characteristics in anti-discrimination law),[55] and from particular provisions concerning profiling.[56] These provisions do not only limit profiling and automated decision-making, but they also specify corresponding rights and duties, including rights of access ('meaningful information' about the logic of profiling),[57] rights to rectification and erasure,[58] or the duties to ensure data protection by design and by default[59] and to carry out a data protection impact assessment.[60]

[50] Recital 71 in regard to Article 22 GDPR states: '[. . .] In order to ensure fair and transparent processing in respect of the data subject, [. . .] the controller should [. . .] secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. [. . .]'. The prevention of anti-discrimination is also referred to in Recitals 75 and 85.

[51] The Article 29 Data Protection Working Party favours a broad reading of Article 22 GDPR for machine-human interaction, qualifying as automated decision-making if a human 'routinely applies automatically generated profiles to individuals', in other words, if human intervention is reduced to a mere 'token gesture'. It suggests a similarly broad understanding of significant effects, possibly including the refusal of a contract or targeted advertising; Guidelines on Automated individual decision-making (n 23) 10–11.

[52] GDPR, Articles 22(2)–(4).

[53] *Cf.* R Poscher, Chapter 16, in this volume.

[54] *Cf.* n 50 for the GDPR and LED, Recitals 23, 38, 51, and 61.

[55] GDPR, Article 9; LED, Article 10.

[56] GDPR, Article 22; LED, Article 11.

[57] GDPR, Article 15(1)(h); general information rights are granted in Articles 12–15 GDPR, Articles 12–14 LED.

[58] GDPR, Articles 16 and 17; LED, Article 16.

[59] GDPR, Article 25; LED, Article 20.

[60] GDPR, Article 35; LED, Article 27.

III. CAUSES FOR DISCRIMINATION

After examining the legal framework for profiling and decision-making, it is now crucial to ask why discrimination occurs in the context of intelligent profiling. This article suggests that one can distinguish two (partially overlapping) causes of discrimination: (1) the use of statistical correlations and (2) technological and methodological factors, commonly referred to as 'bias'.

1. *Preferences and Statistical Correlations*

American economists were the first to distinguish between taste-based discrimination and statistical discrimination ('discrimination' meaning differentiation, bearing no negative connotation). According to this distinction, discrimination either relies on preferences or implies the rational use of statistical correlations to cope with a lack of information. If, for instance, young age correlates with high productivity, a prospective employer who does not know the individual productivity of two applicants may hire the younger applicant in efforts to increase the productivity of her enterprise. Due to its rational objective, statistical discrimination seems less problematic than enacting ones' irrational preferences, for example not hiring older applicants based on a dislike for older people.[61]

It is evident that direct or indirect discrimination resulting from group profiling[62] also qualifies as statistical discrimination. Group profiling describes the process of predicting characteristics of groups, as opposed to personalised profiling which aims to identify a particular person and to predict her characteristics.[63] Data mining and automation allows for increasingly sophisticated profiles and correlations to be established. Instead of relying on a simple proxy like age, gender, or race, decision-making can now be based on a complex profile. The use of these profiles rests on the assumption that the members of a certain group defined by specific data points also exhibit certain (unknown, but relevant) characteristics. Examples of this practice can be found everywhere as more and more private companies and state agencies use algorithmic group profiles. Companies, for example, rely on group profiles assessing the capabilities of prospective employees, the risks of prospective insurees, or the preferences of online consumers. But state agencies also take group profiles into account, when, for instance, predicting the inclination to commit an offence or the need for social assistance.[64]

Even if contrasted with taste-based discrimination, statistical discrimination is not wholly unproblematic. Sometimes, it implies direct discrimination based on protected characteristics, for example if certain risks allegedly correlate with race, religion, or gender.[65] Furthermore, statistical discrimination means that the predicted characteristic of a group is attributed to its

[61] E Phelps, 'The Statistical Theory of Racism and Sexism' (1972) 62 *The American Economic Review* 659; *cf.* G Britz, *Einzelfallgerechtigkeit versus Generalisierung* (2008) 15 *et seq* (hereafter Britz, *Einzelfallgerechtigkeit*). The term statistical discrimination should not be confused with the statistical proof of (indirect) discrimination.

[62] The term 'profiling' means 'group profiling' unless otherwise noted.

[63] M Hildebrandt, 'Defining Profiling: A New Type of Knowledge' in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen* (2008)17, 20–23 (hereafter Hildebrandt, 'Defining Profiling').

[64] On predictive policing based on group profiles see E Joh, 'The New Surveillance Discretion' (2016) 15 *Harvard Law & Policy Review* 24; A Ferguson, 'Policing Predictive Policing' (2017) 94 *Washington University Law Review* 1109, 1137–1143; examples of European state practice can be found in AlgorithmWatch, 'Automating Society' (*Algorithm Watch*, January 2019) https://algorithmwatch.org/en/automating-society/; e.g. in employment service 43, 108, 121, in children and youth assistance and protection 50, 61, 101, 115, in health care 88–89.

[65] A von Ungern-Sternberg, 'Religious Profiling, Statistical Discrimination and the Fight against Terrorism in Public International Law' in R Uerpmann-Wittzack, E Lagrange and S Oeter (eds), *Religion and International Law* (2018), 191 (hereafter Ungern-Sternberg, 'Religious Profiling').

https://doi.org/10.1017/9781009207898.020 Published online by Cambridge University Press

members, even though there is only a certain probability that a group member shares this characteristic[66] and even though the attributes themselves might be negative (e.g. a correlation of race and delinquency or of age and mental capacity).[67]

Finally, it should be noted that discrimination can be based on a combination of taste and statistical correlations. This is the case, for example, when companies take into account consumer preferences predicted from group profiles. Online platforms respond to presumed user preferences when displaying news, search results, or information on prospective employers, dates, or goods. This can also raise problems. Predicting group preferences might disadvantage certain groups of users, like female or Black jobseekers who are shown less attractive job offers than White male men.[68] Additionally, group preferences might be discriminatory and lead to discriminatory decisions. Google searches for Black Americans might yield ads for criminal record checks, the comments of people of colour or homosexuals might be less visible on online platforms, and dating platform users might be categorised along racial or ethnic lines.[69]

2. *Technological and Methodological Factors*

Discrimination based on correlations can also entail (further) disadvantages and biases stemming from the profiling method. In the literature, this phenomenon is sometimes called 'technical bias'.[70] This term can be misleading, however, as these biases also occur in the context of human profiling.[71] Furthermore, these biases result not only from technical circumstances, but also from deliberate methodological decisions. These decisions involve collecting the training data (step 1), specifying a concrete outcome to predict (including one or several target variables indicating this outcome) (step 2), choosing possible predictor variables that are made available to the training algorithm (step 3), and finally, after the training algorithm has chosen and assessed the relevant predictor variables for the predicting model (i.e. after building the screening algorithm) validating the screening algorithm in another (verification) dataset (step 4).[72] All of these decisions can involve biases.

a. Sampling Bias

A sampling bias may follow from unrepresentative datasets that are used to train (step 1) and to validate (step 4) algorithms.[73] Transferring the result of machine learning to new data rests on the assumption that this new data has similar characteristics as the dataset used to train and

[66] This is why Hildebrandt (in Hildebrandt, 'Defining Profiling' (n 63) 21) considers group profiles 'non-distributive profiles'.

[67] On this see Britz, *Einzelfallgerechtigkeit* (n 61) 23.

[68] T Speicher and others, 'Potential for Discrimination in Online Targeted Advertising' (2018) 81 *Proceedings of Machine Learning research* 1.

[69] L Sweeney, 'Discrimination in Online Ad Delivery' (2013) 56 *Communications of the ACM* 44; N Kayser-Bril, 'Automated Moderation Tool from Google Rates People of Color and Gays as "Toxic"' (*Algorithmwatch*, 19 May 2020) https://algorithmwatch.org/en/story/automated-moderation-perspective-bias/; J Hutson and others, 'Debiasing Desire: Addressing Bias & Discrimination on Intimate Platforms' (2018) 2 *Proceedings of the ACM on Human-Computer Interaction* 1.

[70] There does not seem to be an established terminology yet, *cf*. Friedman and Nissenbaum, 'Bias in Computer Systems' (n 8) 333; Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 50; Barocas and Selbst, 'Big Data's Disparate Impact' (n 8) 681.

[71] Britz, *Einzelfallgerechtigkeit* (n 61) 18–22.

[72] Kleinberg and others, '*Discrimination in the Age of Algorithms*' (n 6) 22.

[73] Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 51; Barocas and Selbst, 'Big Data's Disparate Impact (n 8) 684; Orwat, *Diskriminierungsrisiken* (n 8) 79–82.

validate the algorithm.[74] Image recognition illustrates this point. If the training data does not contain images representing future uses, like images with different kinds of backgrounds, this can lead to recognition errors.[75] Bias does not only result from underrepresentation, where, for instance, image recognition training data contains fewer images of Black people or if training data for recruiting purposes includes few examples of successful female employees. Overrepresentation can also cause bias. 'Racial profiling', for example police stops targeting people of colour, typically lead to a much higher detection rate for people of colour than for the White population, which then suggests a – biased – statistical correlation between race and crime rate.[76]

Several factors might lead to the use of unrepresentative datasets. Representative datasets are often unavailable in contemporary societies shaped by inequalities. Moreover, existing datasets might be outdated,[77] designers might simply not realise that data is unrepresentative, or designers might be influenced by stereotypes or discriminatory preferences. If statistical assumptions cannot be properly reassessed, this might also lead to unrepresentative data, like when predictions concerning creditworthiness can only be verified with regard to the credits granted (not the credits that were denied) or if predictions concerning recidivism can only be controlled with regard to the decisions granting parole (not the decisions refusing parole).

b. Labelling Bias

Labelling, or the attribution of characteristics influenced by stereotypes or discriminatory preferences, can also induce bias.[78] Data not only refers to objective facts (e.g. the punctual discharge of financial obligations, high sales results), but also to subjective assessments (e.g. made on an evaluation platform or in job references). As a consequence, target variables (step 2), but also training and validation data (steps 1 and 4) and the predictor variables used in the predicting model (step 3), can relate either to objective facts or to subjective assessments. These assessments may reflect discriminatory prejudices and stereotypes as was shown for legal exams[79] or the evaluation of teachers.[80] In addition to that, discriminatory assessments might also result in – biased – facts, for example if the police stops or arrests members of minority groups at a disproportionately high level.

c. Feature Selection Bias

Feature selection bias means that relevant characteristics are not sufficiently taken into account.[81] Algorithms consider all data available when establishing correlations used for predictions (steps 1, 2, 4). Car insurance companies, for example, traditionally rely on specific data

[74] Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 46.

[75] *Cf.* the recognition of wolves and huskies M Ribeiro, S Singh, and C Guestrin, 'Why Should I Trust You?' in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016) 1135 (1142).

[76] F Schauer, *Profiles, Probabilities, and Stereotypes* (2003) 194; B Harcourt, *Against Prediction. Profiling, Policing and Punishing in an Actuarial Age* (2007) 145 (hereafter Harcourt, *Against Prediction*).

[77] Kleinberg and others, *Discrimination in the Age of Algorithms* (n 6), 41 ('zombie predictions').

[78] Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 50–51; Barocas and Selbst, 'Big Data's Disparate Impact' (n 8) 681; Orwat, *Diskriminierungsrisiken* (n 8) 77–78.

[79] Female and immigrant students receive lower grades E Towfigh, C Traxler, and A Glöckner, 'Geschlechts- und Herkunftseffekte bei der Benotung juristischer Staatsprüfungen' (2018) 5 *Zeitschrift für Didaktik der Rechtswissenschaften* 115.

[80] A Özgümüs and others, 'Gender Bias in the Evaluation of Teaching Materials' (2020) 11 *Frontiers in Psychology* 1074.

[81] *Cf.* Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 52–53; Barocas and Selbst, Big Data's Disparate Impact' (n 8) 688.

concerning the vehicle (car type, engine power) and the driver(s) (age, address, driving experience, crash history; in the past also gender[82]) to specify the risk of a traffic accident. One can assume, however, that other types of data like an aggressive or defensive driving style correlate much stronger with the risk of accident than age (or gender).[83] Instead of imposing particularly high insurance premiums upon young (male) novice drivers, insurance companies could define categories of premiums according to the driving style and thus avoid discrimination based on age (or gender). Similarly, assessing the credit default risk could be based on meaningful features like income and consumer behaviour instead of relying on the borrower's address, which disadvantages the residents of poorer quarters ('redlining').[84]

d. Error Rates

Finally, statistical predictions also generate errors. Therefore, one has to accept certain error rates, such as false positives (e.g. predicting a high risk of recidivism where the offender does not reoffend) and false negatives (predicting a low risk of recidivism where the offender actually reoffends). It is now a matter of normative assessment which error rates seem acceptable for which kinds of decisions, for example for denying a credit or adding someone to the no-fly list. Moreover, when defining the target of profiling (step 2), the designers of algorithms must also decide how to allocate different error rates among different societal groups. If the relevant risks are not distributed evenly among different societal groups (say, if women have a higher risk of being genetic carriers of a disease than men or if men have a higher risk of recidivism than women), it is mathematically impossible to allocate similar error rates to all the affected groups, either overall for women and men, or for women and men within the group of false negatives or false positives respectively.[85] This problem was first detected and discussed in the context of predicted recidivism, where differing error rates manifested for Black versus White criminal offenders.[86] It follows from the trade-off that algorithms' designers can influence the allocation of error rates, and that regulators could shape this decision through legal rules.

IV. JUSTIFYING DIRECT AND INDIRECT FORMS OF DISCRIMINATORY AI: NORMATIVE AND TECHNOLOGICAL STANDARDS

The previous section highlighted different causes for discrimination in decision-making based on profiling. This section now turns to the question of justification, and argues that these causes are a relevant factor for the proportionality of direct or indirect discrimination. After specifying the proportionality framework (1), this section develops general considerations concerning statistical discrimination or group profiling (2) and examines the methodology of automated profiling (3) before turning to the difference between direct and indirect discrimination (4).

[82] This practice has been banned by the CJEU, Case C-236/09 *Test-Achats* [2011].

[83] On this example *cf*. Calders and Žliobaitė, 'Unbiased Computational Processes' (n 8) 52–53.

[84] Barocas and Selbst, 'Big Data's Disparate Impact' (n 8) 689.

[85] J Kleinberg, S Mullainathan, and M Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' in C Papadimitrou (ed), *8th Innovations in Theoretical Computer Science Conference* (ITCS 2017) 43:1 (hereafter Kleinberg, Mullainathan, and Raghavan, 'Inherent Trade-Offs'); K Zweig and T Krafft, 'Fairness und Qualität Algorithmischer Entscheidungen' in M Kar, B Thapa, and P Parycek (eds), *(Un)Berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (2018) 204 (213–218) (hereafter Zweig and Krafft, 'Fairness und Qualität').

[86] Critically Angwin and others, 'Machine Bias' (n 4); on the problem Kleinberg, Mullainathan, and Raghavan, 'Inherent Trade-Offs' (n 85); Zweig and Krafft, 'Fairness und Qualität' (n 86); Cofone, 'Algorithmic Discrimination' (n 10) 1433–1436.

1. Proportionality Framework

The justification of discriminatory measures regularly includes proportionality.[87] EU law, for example, speaks of 'appropriate and necessary' means[88] of 'proportionate' genuine and determining occupational requirements[89] or, in the general limitation clause of Article 52 (1) Charter of Fundamental Rights, of 'the principle of proportionality'. Different legal systems vary in how they define and assess proportionality. The European Court of Human Rights applies an open 'balancing' test with respect to Article 14 ECHR,[90] and the European Court of Justice normally proceeds in two steps, analysing the suitability (appropriateness) and the necessity of the measure at stake.[91] In German constitutional law and elsewhere,[92] a three-step test has been established. According to this test, proportionality means that a (discriminatory) measure is suitable to achieve a legitimate aim (step 1), necessary to achieve this aim, meaning that the aim cannot be achieved by less onerous means (step 2), and appropriate in the specific case, where the legal interest pursued by a discriminatory measure outweighs the conflicting legal interest of non-discrimination (step 3). This three-step test will be used as an analytical tool to flesh out arguments that are relevant for justifying differential treatment or detrimental effect as a result of profiling and decision-making. Before this analysis, some aspects merit clarification.

a. Proportionality as a Standard for Equality and Anti-Discrimination

Some legal scholars claim that the notion of proportionality is only useful for assessing the violation of freedoms, not of equality rights. According to this view, an interference with a freedom, such as limits on the freedom of speech, constitutes a harm that needs to be justified with respect to a conflicting interest, such as protection of minors. In contrast, unequal treatment is omnipresent. It does not constitute prima facie harm (e.g. different laws for press and media platforms), and it typically does not pursue conflicting objectives. Rather, it reflects existing differences. To illustrate, different rules on youth protection for the press and for media platforms are not necessarily in conflict with youth protection. Rather, they result from different risks emanating from the press and media platforms.[93] Thus, in order to justify differential treatment one has to show that this differentiation follows 'acceptable standards of justice'

[87] On justification norms *cf.* n 46–48.

[88] E.g. with respect to direct discrimination Article 4(5) Goods and Services Sex Discrimination Directive 2004/113/EC; with respect to indirect discrimination e.g. Article 2(2)(b)(i) Framework Directive 2000/78/EC; Article 2(2)(b) Race Directive 2000/43/EC; Article 2(b) Goods and Services Sex Discrimination Directive 2004/113/EC; Article 2(1)(b) Equal Treatment Directive 2006/54/EC.

[89] E.g. with respect to direct discrimination Article 4(1) Framework Directive 2000/78/EC; Article 4 Race Directive 2000/43/EC; Article 14(2) Equal Treatment Directive 2006/54/EC.

[90] Harris and others, *European Convention on Human Rights* (n 48) 774; B Rainey and others, *The European Convention on Human Rights* (7th ed. 2017) 646–647.

[91] T Tridimas, 'The Principle of Proportionality' in R Schütze and T Tridimas (eds), *Oxford Principles of European Union Law*, Vol 1, 243, 247 (hereafter Tridimas, 'The Principle of Proportionality'); see, for example, CJEU, Case C-555/07 *Kücükdeveci* [2010] para 37–41; CJEU, Case C-528/13 *Léger* [2015] para 58–68; CJEU, Case C-157/15 *Achbita* [2017] para 40–43; CJEU, Case C-914/19 *GN* [2021] para 41–50; but note also the three-prong test including proportionality in the narrower sense, for example, in CJEU, Case C-83/14 *CHEZ* [2015] para 123–127.

[92] R Poscher in M Herdegen and others (eds), *Handbook on Constitutional Law* (2021) § 3 (forthcoming) (hereafter Poscher in 'Handbook on Constitutional Law'); N Petersen, *Verhältnismäßigkeit als Rationalitätskontrolle* (2015) (hereafter Petersen, *Verhältnismäßigkeit*); on the spread of this concept A Stone Sweet and J Mathews, 'Proportionality Balancing and Global Constitutionalism' (2008) 47 *Columbia Journal of Transnational Law* 72.

[93] The example is mine. The proportionality test is criticised by U Kischel, 'Art. 3 GG' in V Epping and C Hillgruber (eds), *BeckOK Grundgesetz* (47th ed. 2021) para 34–38a (hereafter Kischel, 'Art. 3 GG'), with further references.

reflecting 'relevant' differences,[94] or that the objective reasons outweigh the inequality impairment.[95] Only if differential treatment is meant to promote an 'external' objective unrelated to existing differences[96] should a proportionality assessment be made, according to some scholars.[97]

Nevertheless, the proportionality framework remains useful for the task of justifying discriminatory AI. The aforementioned proportionality scepticism seems partly motivated by the concern that equality rights and justification requirements must not expand uncontrollably. However, this valid point only applies to general equality rights in the context of which this concern was voiced, not to anti-discrimination law. Favouring men over women and *vice versa* does constitute *prima facie* harm, and justifying this differential treatment requires strict scrutiny and the consideration of less harmful alternative measures. In part, proportionality seems to be rejected as a justification standard because its criteria are too unclear. However, the proportionality assessment is flexible enough to take into account the characteristics of discriminatory measures. Thus, the proportionality test should evaluate whether using a particular differentiation criterion (like gender) is suitable, necessary, and appropriate for reaching the differentiation aim (e.g. setting appropriate insurance premiums, stopping tax evasion). For differential treatment based on profiling, this indeed implies that the differentiation criterion and the differentiation aim are not in conflict with each other as the decision-making responds to the different risks predicted as a result of profiling. A proportionality assessment now allows for strict scrutiny of both decision-making and profiling. This advantage of the proportionality test becomes increasingly important as profiling replaces older methods of differentiating between people. Moreover, a second advantage of the proportionality approach is its dual use for both direct and indirect discrimination. The detrimental effect of a facially neutral measure must not be justified with reference to existing differences. Quite the contrary, it must be justified with reference to an 'external' objective and proportionate means to achieve this objective.[98] Thus, apart from the fact that the law calls for proportionality, there are good reasons to stick to this standard, particularly for an assessment of profiling.

b. Three Steps: Suitability, Necessity, Appropriateness

In a nutshell, the proportionality test entails three simple questions: first, do the measures work, that is, does profiling and decision-making promote the (legitimate) aim (suitability)? Second, are there alternative, less onerous means of profiling and decision-making to achieve this aim (necessity)? Third, is the harm caused by profiling and decision-making outweighed by other interests (appropriateness)? If questions one and three can be answered in the affirmative and if question two can be answered negatively, the measure is proportionate and justified.

Note that this counting method does not include the preceding step of verifying that a measure pursues a legitimate aim, nor does it comprise the rarer consideration that the means

[94] S Huster, 'Art. 3' in KH Friauf and W Höfling (eds), *Berliner Kommentar zum Grundgesetz* (50th supplement 2016) para 89 (hereafter Huster, 'Art. 3').

[95] Kischel, 'Art. 3 GG' (n 93) para 37.

[96] S Huster, 'Gleichheit und Verhältnismäßigkeit' (1994) 49 *Juristenzeitung* 541, 543 (hereafter Huster, 'Gleichheit und Verhältnismäßigkeit') gives the examples of (1) different taxation based on different income which he qualifies as reflecting existing inequalities ('internal objective') and (2) different taxation aimed at stimulating the construction industry, providing tax relief for builders, which he qualifies as 'external objective'.

[97] Huster, 'Gleichheit und Verhältnismäßigkeit' (n 96) 549; Huster, 'Art. 3' (n 94) para 75–86, with further references.

[98] One can draw a parallel between direct and indirect discrimination on the one hand and *Huster's* idea of 'internal' and 'external' objectives in equality cases on the other hand (n 94 and 96).

used for pursuing this aim is itself legitimate.[99] It can be assumed that the aims pursued by decision-making based on profiling pursue legitimate aims, such as finding and hiring the most qualified applicant or monitor persons inclined to commit a crime. This article will also neglect the possibility that the means itself is prohibited. Profiling might be prohibited per se, for example, if past human actions are assessed individually. An individual criminal conviction or student performance grade cannot be based on statistical predictions concerning recidivism among certain groups of offenders or based on certain schools' performance.[100]

Turning to the 3-step test, it should be emphasised that it refers to profiling and decision-making, this means to two interrelated, but different acts. It is the decision that needs to be justified under non-discrimination law for involving different treatment or for causing detrimental effect. However, as far as this decision is based on a prediction resulting from profiling, profiling as an instrument of prediction must also be proportionate. Profiling is proportionate if it generates valid predictions (suitability, step 1), if alternative profiling methods that generate equally good predictions at lower costs do not exist (necessity, step 2), and if the harm of profiling is outweighed by its benefits (appropriateness, step 3). In addition, other aspects of the discriminatory decision also come under scrutiny, notably the harm of a decision (for example a police control involves a different sort of harm than a flight ban).[101]

Some proportionality scholars doubt that steps 2 and 3 can be meaningfully separated.[102] The European Court of Justice (ECJ), which typically applies a 2-step test comprising suitability and necessity, sometimes includes elements of balancing in its reasoning at the second step,[103] but increasingly also resorts to the 3-step test.[104] This chapter submits that it is helpful to separate steps 2 and 3. In step 2, the measure in question is compared to alternative measures which are equally effective in achieving a particular aim, for example, different profiling methods equally good at predicting a risk. If an alternative means generates more costs or curtails other rights, the conditions 'equally suitable' and 'less burdensome' are not met.[105] This means comparing both normative and factual burdens for different groups of people: the persons affected by the measure under review, third parties that might be affected by alternative measures, and the decision-maker. An alternative profiling method, for example, could place a different burden on the persons affected by the measure under review (e.g. by using more personal data and thus limiting privacy). An alternative profiling method could also place a burden on third parties (e.g. if the alternative method yields negative profiling results followed by disadvantageous decisions). Finally, an alternative profiling method could also burden the decision-maker because the method requires more resources such as time or money. These considerations involve value

[99] *Cf.* Poscher in 'Handbook on Constitutional Law' (n 92).

[100] In the UK, it was planned to use an A-level algorithm predicting grades in 2020 as the A-level exams were cancelled due to COVID-19. The algorithm was meant to take into account the teachers' assessment of individual pupils and the performance of the respective school in past A-level exams in order to combat inflation in grades. The algorithm would have had disadvantaged good pupils from state-run schools with ethnic minorities. The project was cancelled after public protest. *Cf.* Wachter, Mittelstadt, and Russell, 'Bias Preservation' (n 12) 1–6.

[101] On these points *cf.* Sub-sections IV 2 and 3.

[102] Moreover, it is disputed that rational criteria exist for the balancing exercise of step 3. *Cf.* T Kingreen and R Poscher, *Grundrechte Staatsrecht II* (36th ed. 2020) § 6 para 340–347; for an in-depth analysis on the criticism of balancing and its underlying, see N Petersen, 'How to Compare the Length of Lines to the Weight of Stones: Balancing and the Resolution of Value Conflicts in Constitutional Law' (2013) 14 *German Law Journal* 1387.

[103] Tridimas, 'The Principle of Proportionality' (n 91); *cf.* also G de Burca, 'The Principle of Proportionality and Its Application in EC Law' (1993) 13 *Yearbook of European Law* 105, 113–114.

[104] B Oreschnik, *Verhältnismäßigkeit und Kontrolldichte* (2018) 158–178, 219–227.

[105] Poscher in 'Handbook on Constitutional Law' (n 92) paras 63–67.

assessments, as different burdens have to be identified and weighed. It is not surprising that some legal systems prefer to see these considerations as part of the balancing test (step 3), whereas other legal systems address reasonable alternative measures under the heading of necessity only (step 2).[106] It is nevertheless a useful analytical tool to distinguish between less onerous alternative means (step 2) and other alternative means (step 3).

Finally, it should be emphasised that by treating proportionality as a general issue, this article does not mean to downplay the particularities of specific justification provisions or to conceal the different harms caused by different forms of discrimination. Particularly severe forms of direct discrimination will hardly be justifiable at all (like direct discrimination on grounds of race) or merit very strict scrutiny (for example direct discrimination on grounds of gender which can be justified based on biological differences), other forms might be much easier to justify depending on the circumstances. Furthermore, a distinction must also be drawn between decisions made by the state and by private actors. Even if anti-discrimination law covers both, the state is directly bound by fundamental rights including equality and non-discrimination. By contrast, the choices and actions of private actors are protected by fundamental freedoms such as freedom of contract or freedom to conduct a business, leading to a stricter burden of justification for state actors than for private actors. The point of this article is to elaborate on the commonalities of discriminatory decision-making based on profiling, and to show the relevant aspects for assessing its legality.

2. *General Considerations Concerning Statistical Discrimination/Group Profiling*

In the context of discriminatory profiling and decision-making, it is useful to distinguish general aspects of proportionality that are known from non-automated forms of statistical discrimination (this section), and specific aspects of automated group profiling (IV.3.). Note that the terms 'statistical discrimination' and decision-making based on 'group profiling' designate the same phenomena.[107] The first term is long-established, while the term 'group profiling' is mainly used in the context of automated profiling. Both refer to differential treatment or detrimental effect that results from statistical predictions and affects groups defined by sensitive characteristics or its members. Before looking at specific issues of the methodology of profiling in the next section, this section will highlight some arguments relevant for the proportionality test.

a. Different Harms: Decision Harm, Error Harm, Attribution Harm

As a starting point, one can distinguish different harms stemming from profiling and decision-making.[108] The decision itself contains negative consequences corresponding to a varying degree of 'decision harm': a denial of goods (no credit), bad contract terms (high insurance premiums), a denial of chances (no job interview), or investigations (a police control). 'Decision harms' arise in human and automated decisions alike. But some forms of 'decision harm' are typical of decisions based on profiling. Profiling is meant to overcome an information deficit (Who is a qualified employee? Which person is about to commit a crime?). Therefore, many decisions tend to be part of an information gathering process: Some job applicants are chosen for a job interview, while others are refused right away. Some taxpayers are singled out for an audit, while other filers' tax declarations are accepted without further review. It is important to recognise that

[106] Petersen, *Verhältnismäßigkeit* (n 92), 144–147, 258–262, for example, argues comprehensively that it might be easier for well-established, powerful courts to openly apply a balancing test than for other courts.

[107] See Sub-section III 1.

[108] See also Britz, *Einzelfallgerechtigkeit* (n 61) 120–136, albeit with different classifications.

https://doi.org/10.1017/9781009207898.020 Published online by Cambridge University Press

these decisions involve a harm of their own. They attribute opportunities and risks which can be very relevant for the individual person, but they can also lead to the deepening of existing stereotypes and inequalities.

Other harms relate to profiling. Statistical predictions generated by profiling have a certain error rate, which means that false positives (like honest taxpayers flagged for the risk of fraud) or false negatives (as creditworthy consumers with a low credit score) suffer from the negative consequences of a decision. This sort of 'error harm' is already known as 'generalisation harm' in jurisprudence. Legal systems are based on legal rules which, by definition, apply in a general manner, as opposed to decisions based on specific issues targeting specific individuals. A general rule will often be overinclusive. For example an age limit for pilots addresses pilots' statistically decreasing flying ability with age, but it also applies to persons who are still perfectly fit to fly.[109] This sort of 'generalisation harm' can be quantified in the process of automated profiling as error rates. Finally, group profiles also carry the risk of 'attribution harm' if they associate all members of a group with a negative characteristic, e.g. Black people with higher criminality or women with lower performance. The degree of 'attribution harm' can also vary: some characteristics predicted by profiling can be embarrassing or humiliating (like crime, low work performance, confidential health data), while others are not problematic (e.g. high purchasing power). Some of these negative attributions are visible to others (such as police disproportionately stopping or searching Black people), while others remain hidden in the algorithm. Some attributions confirm and reinforce existing stereotypes, while others run counter to existing prejudices (for example a good driving record for women). Some attributions can be corrected in the individual case (e.g. if a police check does not yield a result), while others remain unrefuted.

Under the proportionality test, these harms, the varying degrees of harm evoked in particular instances, are relevant for steps 2 and 3, that is, for assessing whether alternative means are less onerous (evoke less harm) than the measure at hand (necessity, step 2), and for balancing the conflicting interests (appropriateness, step 3).

b. Alternative Means: Profiling Granularity and Information Gathering

After defining the distinct harms of profiling and decision-making, we can now turn to concrete strategies to better reconcile conflicting interests. This is again either a matter of necessity (step 2) or appropriateness (step 3). The measure at issue is not necessary if an alternative means is equally suitable to reach a particular aim without imposing the same burden, and the measure is not appropriate if it is reasonable to resort to an alternative measure that better reconciles the conflicting interests.

This chapter outlines two possible alternative means for decisions based on profiling. The first concerns the granularity of the profiles. Sophisticated profiles obtained from a wealth of data are more accurate than simple profiles based on a few data points only. If decisions are based on simple profiles, then the above-mentioned 'generalisation harm' can result from both profiling and decision-making, as larger groups of people count among the false positives and false negatives[110] and larger groups also suffer the negative effect of a decision. Blood donation, for example, should not lead to the transmission of HIV. In order to reduce this risk, one could exclude several groups from blood donation: homosexuals, male homosexuals, only sexually active male homosexuals, or only sexually active male homosexuals engaging in behaviour

[109] *Cf.* CJEU, Case C-190/16 *Fries* [2017].
[110] On error rates see also Sub-section III 2(d).

which puts them at a high risk of acquiring HIV. The more the group is defined, the smaller the number of people affected by a prohibition of blood donation.[111] As a consequence, the higher accuracy of fine-granular group profiles must, therefore, be weighed against the advantages of simple group profiles such as data minimisation or simplicity. The need for granular profiles is expressed, for example, in the German implementation of the European Passenger Name Record (PNR) system. The EU PNR Directive provides that air passengers are assessed with respect to possible involvement in terrorism or other serious crime. This is done by comparing passenger data against relevant databases and pre-determined criteria (i.e. by profiling), and these criteria need to be 'targeted, proportionate and specific'.[112] The German Air Passenger Data Act implementing this provision stipulates that the relevant features (i.e. factors providing ground for suspicion, as well as exonerating factors) must be combined 'such that the number of persons matching the pattern is as small as possible'.[113]

Second, as profiling helps address information deficits, alternative means of coping with these deficits can also be a relevant aspect of the proportionality test. If information is particularly important, fully clarifying the facts can be preferable to profiling, provided that this is feasible and that the resources are available. Take the example of airport security screening. Screening of air passengers and their luggage items is not confined to a certain sample of 'high risk' passengers but extends to all passengers. Regarding the blood donation example, systematically screening all blood donations for HIV could be an alternative means to refusing sexually active male homosexuals to donate blood.[114] Similar forms of full fact-finding are also conceivable in the context of automation, although they create costs and they entail the large-scale processing of personal data. Another method of reconciling the need for information and non-discrimination is randomisation, this means gathering information at random. If only a fraction of tax returns can be scrutinised by the fiscal authorities, these tax returns can be chosen at random or based on the profile of a tax evader. Using risk profiles might seem to allocate resources more efficiently, but randomisation has other advantages: it burdens all taxpayers equally and prevents discriminatory effects.[115] In addition, it might also be more efficient and less susceptible to manipulation because taxpayers cannot game the algorithm.[116]

3. *Methodology of Automated Profiling: A Right to Reasonable Inferences*

This section turns to the methodology of automated profiling, which has a decisive impact on the possible harms of discriminatory AI.[117] It looks at legal sources for explicit and implicit methodology standards and links them to the elements of the proportionality test. As a result, this section claims that a 'right to reasonable inferences'[118] already exists in the context of discriminatory AI.

[111] CJEU, Case C-528/13 *Léger* [2015] para 67.
[112] Article 6(4) Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime.
[113] Section 4(3) Passenger Name Record Act of 6 June 2017 *Bundesgesetzblatt* I 1484, as amended by Article 2 of the Act of 6 June 2017 *Bundesgesetzblatt* I 1484.
[114] CJEU, Case C-528/13 *Léger* [2015] para 64.
[115] Harcourt, *Against Prediction* (n 76) 237.
[116] The German automated risk management system which selects tax returns for human review is complemented by randomised human tax reviews, Section 88(5)(1) German Fiscal Code of 1 October 2002 *Bundesgesetzblatt* I 3866, last amended by Article 17 of the Act of 17 July 2017 *Bundesgesetzblatt* I 2541.
[117] See Sub-section III 2.
[118] Called for by Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

a. Explicit and Implicit Methodology Standards

As opposed to other activities, such as operating a nuclear power plant or selling pharmaceuticals, developing and using profiling algorithms does not require a permission issued by a state agency. Operators of nuclear power plants in Germany, for example, must show that 'necessary precautions have been taken in accordance with the state of the art in science and technology against damage caused by the construction and operation of the installation' before obtaining a licence,[119] and pharmaceutical companies need to prove that pharmaceuticals have been sufficiently tested and possess therapeutic efficacy 'in accordance with the confirmed state of scientific knowledge'[120] before obtaining the necessary marketing authorisation. The referral to the 'state of the art in sciences and technology' or the 'confirmed state of scientific knowledge' implies that methodology standards developed outside the law, for example in safety engineering or pharmaceutics, are incorporated into the law. Currently, there is no similar *ex ante* control of profiling algorithms, which means that algorithms are not measured against any methodological standards in order to qualify for a permission. This situation might change, of course. The German Data Ethics Commission, for example, suggests that algorithmic systems with regular or serious potential for harm should be covered by a licensing procedure or preliminary checks.[121]

But the lack of a licensing procedure does not mean that methodology standards for algorithmic profiling do not exist. Some legal norms explicitly refer to methodology, and implicit methodological standards can also be found in the general justification test for discrimination. These standards may be enforced – *ex post* – by affected individuals who bring civil or administrative proceedings, or by public agencies like data protection authorities or anti-discrimination bodies who control actors and fine offenders.[122]

Legal norms that explicitly state methodology requirements for profiling and decision-making exist. The German Federal Data Protection Act, for example, regulates some aspects of scoring, such as the use of a probability value for certain future action by a natural person and, hence, a particular form of profiling. The statute stipulates that 'the data used to calculate the probability value are demonstrably essential for calculating the probability of the action on the basis of a scientifically recognised mathematic-statistical procedure'.[123] Similar requirements can be found in insurance law. The Goods and Services Sex Discrimination ('Unisex') Directive 2004/113/EC contains an optional clause enabling states to permit the use of sex as a factor in insurance premium calculation and benefits 'where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data'.[124] After the ECJ declared this clause invalid due to sex discrimination,[125] the methodology requirement remains nevertheless relevant for old insurance contracts and provides an inspiration for national standards such as the German General Act on Equal Treatment. This statute, which implements EU anti-

[119] Section 7(2)(3) German Atomic Energy Act of 15 July 1985 *Bundesgesetzblatt* I 1565, as last amended by Article 3 of the Act of 20 May 2021 *Bundesgesetzblatt* I 1194.

[120] Section 25(2)(2 and 4) German Medicinal Products Act of 12 December 2005 Bundesgesetzblatt I 3394, as last amended by Article 11 of the Act of 6 May 2019 *Bundesgesetzblatt* I 646. Emphasis by author.

[121] German Data Ethics Commission, *Opinion of the Data Ethics Commission* (2019) 195 (hereafter German Data Ethics Commission, *Opinion*).

[122] *Cf.* the broad powers of the data protection authorities under Articles 58, 70, 83–84 GDPR.

[123] Section 31(1)(2) German Federal Data Protection Act of 30 June 2017 *Bundesgesetzblatt* I 2097, as last amended by Article 12 of the Act of 20 November 2019 *Bundesgesetzblatt* I 1626; a similar provision can also be found in Section 10 (2)(1) Bancing Act (Kreditwesengesetz). Note that it is disputed whether Section 31 Federal Data Protection Act is in conformity with the GDPR, (i.e. whether it is covered by one of its opening clauses).

[124] Article 5(2) Unisex Directive 2004/113/EC.

[125] CJEU, Case C-236/09 *Test-Achats* [2011].

discrimination law and establishes additional national standards of anti-discrimination law, also contains a methodology requirement for calculating insurance premiums and benefits: 'Differences of treatment on the ground of religion, disability, age or sexual orientation [...] shall be permissible only where these are based on recognised principles of risk-adequate calculations, in particular on an assessment of risk based on actuarial calculations which are in turn based on statistical surveys.'[126] Note that these rules refer to recognised procedures of other disciplines like mathematics, statistics, and actuarial sciences which guarantee that certain aspects of profiling are reasonable from a methodological point of view, that is, that using personal data is 'essential' for probability calculation or that relying on a protected characteristic like sex is a 'determining factor' for risk assessment.

In other contexts, statutes do not refer to methodology in the narrower sense, but to other aspects related to the validity of profiling and establish review obligations. Thus, the EU PNR Directive stipulates that the profiling criteria have to be 'regularly reviewed'.[127] The risk management system used by the German revenue authorities must ensure that 'regular reviews are conducted to determine whether risk management systems are fulfilling their objectives'.[128]

But even if explicit standards do not exist, implicit methodological requirements flow from the justification test – in other words, the proportionality test – of anti-discrimination law. Discriminatory decisions based on automated profiling need to pass the proportionality test, and this includes the methodology of profiling.[129] It is a matter of suitability (step 1) that automated profiling produces valid probability statements. Only then does it further a legitimate goal if a discriminatory decision is based on the result of profiling. Furthermore, it needs to be discussed in the context of necessity (step 2) and appropriateness (step 3) whether a different methodology of profiling and decision-making would have a less discriminatory effect. If the profiling methodology can be improved, if its harms can be reduced, the costs and benefits of these improvements will be relevant for considerations of necessity and appropriateness.

For the sake of completeness, this chapter argues that methodological profiling standards can also be derived from data protection law. In accordance with Article 6(1) of the GDPR the processing of personal data, which is essential for profiling a particular person,[130] requires a legal basis. All legal bases for data processing except consent demand that data processing is 'necessary' for certain purposes, that is, for the performance of a contract,[131] for compliance with a legal obligation,[132] for the performance of a task carried out in the public interest,[133] or for the purposes of legitimate interests.[134] For automated profiling and decision-making, Article 22(2) and (3) GDPR also require suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, which includes non-discrimination. Thus, the necessity test of Article 6 (1) GDPR and the safeguarding clause of Article 22(2) and (3) GDPR also imply a minimum standard of profiling methodology. Data processing for profiling is only necessary for the

[126] Section 20(2) German General Act on Equal Treatment of 14 August 2006 *Bundesgesetzblatt* I 1897, as last amended by Article 8 of the SEPA Accompanying Act of 3 April 2013 *Bundesgesetzblatt* I 610. *Cf.* Section 33(5) General Act on Equal Treatment, on old insurance contracts and gender discrimination.

[127] Article 6(4) PNR Directive (EU) 2016/681.

[128] Section 88(5) German Fiscal Code of 1 October 2002 *Bundesgesetzblatt* I 3866; 2003 I 61, last amended by Article 17 of the Act of 17 July 2017 *Bundesgesetzblatt* I 2541.

[129] See Sub-section IV 1(b).

[130] This is the third step of the profiling process, see II1.

[131] GDPR, Article 6(1)(b).

[132] GDPR, Article 6(1)(c).

[133] GDPR, Article 6(1)(e).

[134] GDPR, Article 6(1)(f).

above-mentioned goals, if the profiling method produces valid predictions and if no alternative profiling method exists which makes equally good predictions while discriminating less. Similar standards can be derived from Article 22 GDPR for automated decision-making based on profiling.

These implicit methodological standards can be developed from the proportionality requirements of anti-discrimination and data protection law even if the legislator has also enacted specific methodological standards with a limited scope of application. Specific methodological standards have long existed in areas of law like insurance and credit law, which refer to established mathematical-statistical standards. Anti-discrimination lawyers, however, have only recently started to call for methodological standards of profiling,[135] long after today's anti-discrimination laws were formulated.[136] Admittedly, the 2016 GDPR addresses the dangers of profiling without also formulating an explicit legal methodological requirement. But Recital 71 requires that 'the controller should use appropriate mathematical or statistical procedures for the profiling [. . .] in a manner [. . .] that prevents [. . .] discriminatory effects'.[137] This non-binding recital expresses the lawmakers' intentions and can help to interpret the legal obligations of the GDPR. Several provisions of GDPR and other recitals also show that the Regulation intends to effectively address the dangers of profiling, including the danger of discrimination.[138] As a consequence, even if the GDPR does not establish an explicit profiling methodology, a minimum standard is implicitly included in the requirement of 'necessary' data protection. In this respect, profiling differs from activities governed by standards outside of data protection law. For example, evaluating exam papers and inferring from these pieces of personal data whether the candidate qualifies for a certain grade follows criteria that have been developed in the examination subject. These criteria cannot be found in data protection law.[139] Inferring information by means of profiling, however, is an activity inextricably linked to data processing and clearly covered by the GDPR.

This minimum standard of a proportionate profiling methodology does not amount to a free-standing 'right to reasonable inferences'[140]. It is a justification requirement triggered by discrimination, this means by different treatment and detrimental impact. However, many decisions based on profiling will involve different treatment or detrimental impact. As a consequence, this

[135] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10) (2019).

[136] Article 21 European Charter of Fundamental Rights (2000), Framework Directive 2000/78/EC, Race Directive 2000/43/EC, Goods and Services Sex Discrimination Directive 2004/113/EC, Equal Treatment Directive 2006/54/EC; German General Equal Treatment Act (2006); not to mention Article 3 German Basic Law (1949) or Article 14 European Convention of Human Rights (1950).

[137] The full sentence reads: "In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect."

[138] Automated decision-making based on profiling is not only addressed in Article 22 GDPR, but also in Articles 13(2)(f), 14(2)(g), 15(1)(h) GDPR (rights to information), Article 35(3)(a) GDPR (data protection impact assessment), Article 47 (2)(e) GDPR (binding corporate rules), Article 70(1)(f) GDPR (guidelines of the European Data Protection Board); profiling as such is addressed in Article 21(1) and (2) GDPR (right to object to certain forms of profiling); moreover Recitals 24, 60, 63, 70–73, 91 concern aspects of profiling. The aim to prevent discrimination is not only expressed in Recital 71, but also in Recital 75 (concerning risks to the rights and freedoms resulting from data processing) and in Recital 85 (concerning damage due to personal data breach).

[139] This is why the right to rectification does not extend to incorrect answers, CJEU, *Nowak* C-434/16, [2017] (n 52–57); *cf.* already Sub-section II 2.

[140] Wachter and Mittelstadt, 'A Right to Reasonable Inferences' (n 10).

minimum standard of proportionate profiling methodology has a wide scope of application. What's more, this standard does not only entail the need for 'reasonable' inferences. Proportionality comprises more than the validity of inferences, it also calls for the least discriminatory methodology that is possible or that can be reasonably expected of the decision-maker.

b. Technical and Legal Elements of Profiling Methodology

The practical challenge now lies in developing appropriate methodological standards.[141] From a technical point of view, disciplines such as data science, mathematics, and computer science shape these standards. At the same time, legal considerations play a decisive role as these methodological standards have a legal basis in the proportionality test. Both technical and legal elements are relevant for assessing the suitability (step 1), the necessity (step 2), and appropriateness (step 3) of profiling.

Returning to the elements of profiling[142] and to the factors identified as causing and affecting discriminatory decisions,[143] it is important to emphasise how technical *and* legal considerations are crucial in developing the right profiling methodology. In regards to error rates, first, it is a technical question to determine how reliable predictions are and how different error rates affect different groups of people depending on allocation decisions.[144] But it is a legal matter to define the minimum standard for the validity of profiling (relevant for suitability, step 1)[145] and to assess whether differences in error rates are significant when comparing the effects and costs of different profiling methods (relevant for necessity and appropriateness, steps 2 and 3). It is also a legal question whether different error rates among different groups are acceptable (i.e. necessary and appropriate).

Second, technical and legal assessments are also required for avoiding or evaluating bias, such as sampling, labelling, or feature selection biases, in the process of profiling. Sampling bias can be prevented by using representative training and testing data. How representative data sets can be obtained or created, and what amount of time, money, and effort this involves, are both technical questions. Moreover, data and computer scientists are also working on alternative methods to simulate representativeness by using synthetic data or processed data sets.[146] The legal evaluation includes the extent to which these additional efforts can be reasonably expected of the decision-maker. Similarly, there are attempts to counteract labelling bias by technical means, such as neutralising pejorative terms in target or predictor variables. But again, these options must also be assessed from a legal point of view, accounting for possible costs and legal harms, such as a loss of free speech in evaluation schemes. Feature selection bias can be reduced by replacing less relevant predictor variables with more relevant ones. Again, aspects of technical feasibility (for instance data availability) and technical performance (like error rate reduction)

[141] See also Orwat, *Diskriminierungsrisiken* (n 8) 114.
[142] Sub-section II 1.
[143] Sub-section III 2.
[144] See Sub-section III 2(d).
[145] Similar legal assessments can be found, for example, in Criminal Procedural Law regarding the reliability of DNA testing methods.
[146] Cofone, 'Algorithmic Discrimination' (n 10) 1431; German Data Ethics Commission, *Opinion* (n 121) 132. On further technical solutions see for example F Kamiran, T Calders, and M Pechenizkiy 'Techniques for Discrimination-Free Predictive Models' in T Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 223; S Hajian and J Domingo-Ferrer, 'Direct and Indirect Discrimination Prevention Methods' in T Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 241; S Verwer and T Calders, 'Introducing Positive Discrimination in Predictive Models' in T Custers and others (eds), *Discrimination and Privacy in the Information Society* (2013) 255.

have to be combined with a legal assessment of technical and legal costs (e.g. a loss of data protection). These considerations concerning possible alternatives to avoid biases are part of the necessity and appropriateness test (steps 2 and 3). Apart from looking at error rates and bias, the proportionality assessment can finally also extend to the profiling model as such. One may argue, for example, that some decisions require a profiling model based on (presumed) causalities, not on mere correlations.

As a consequence, developing appropriate methodological profiling standards will require exchange and cooperation between lawyers and data and computer scientists. In this process, scientists have to explain the validity and the limits of existing methods as well as to explore less discriminatory alternatives, and lawyers have to specify and to weigh benefits and harms of these methods from a legal perspective.

4. *Direct and Indirect Discrimination*

One final aspect of justification concerns direct and indirect discrimination, or differential treatment and detrimental impact. Distinguishing direct and indirect discrimination has been a central tenet of discrimination law up to now. In the age of intelligent profiling, this distinction will become blurred, and indirect discrimination will become increasingly important.

a. Justifying Differential Treatment

In some contexts, even differential treatment based on protected characteristics such as gender, race, nationality, or religion is claimed to be justified based on statistical correlations. This is the case, for example, if unemployed women are less likely to get hired than men and job agencies allocate their services accordingly, if the Swedish minority in Finland has higher credit scores than the Finish majority and, hence, the Swedish can access credit more easily and at lower cost than the Finish, or if Muslims are presumed to have a stronger link to terrorism than the rest of the population and law enforcement agencies more closely scrutinise Muslims.[147] A justification of these forms of different treatment is not entirely ruled out. But the justification should be limited to extremely narrow conditions, especially in the case of particularly problematic characteristics. Even if race, gender, nationality, or religion happened to statistically correlate with certain risks, the harm inflicted by classifying people by these sensitive characteristics is too severe to be generally acceptable. It would not be appropriate (step 3), provided the measure passes the first two steps.[148]

b. Justifying Detrimental Impact

With regard to indirect discrimination, anti-discrimination law has to-date tended to concentrate on evident phenomena. In these cases, clear proxies exist, notably when employers disadvantage (predominantly female) part-time workers[149] or (predominantly Black) applicants who lack certain educational qualifications,[150] or when EU member states make rights or benefits conditional on domestic residence or language skills, which are requirements that are easily met by

[147] On these examples J Holl, G Kernbeiß, and M Wagner-Pinter, *Das AMS-Arbeitsmarktchancen-Modell* (2018) www.ams-forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf; AlgorithmWatch, *Automating Society* (n 64) 59–60; Ungern-Sternberg, 'Religious Profiling'(n 65) 191–193.

[148] Ungern-Sternberg, 'Religious Profiling' (n 65) 205–211.

[149] CJEIU, C- 96/80 *Jenkins* [1981]; CJEU, C-170/84 *Bilka* [1986].

[150] *Griggs v. Duke Power Co*, 401 US 424 (1971).

most nationals, but not by EU foreigners.[151] Thus, indirectly disadvantaging women, Blacks, or aliens has to be justified by establishing that a measure is proportionate to reach a legitimate aim. However, do justification standards need to be equally high in the context of profiling, for example, if group profiles are much more refined and if overlaps with protected groups less clear? Or is it sufficient if profiling is based on a sound methodology? Lawyers will have to clarify why indirect discrimination is problematic and what amounts to such an instance of indirect discrimination.

There are good arguments in favour of extending stricter standards to situations in which proxies are less established and group profiles and protected groups overlap less significantly. Traditionally, one can distinguish 'weak' and 'strong' models of indirect discrimination.[152] According to the 'weak' model, indirect discrimination is meant to back the prohibition of direct discrimination by interdicting ways to circumvent direct discrimination.[153] 'Stronger' models pursue more far-reaching aims such as equality of chances[154] or equality of results correcting existing inequalities[155]. Furthermore, indirect discrimination might also be seen as a functional instrument to secure effective protection of non-discrimination where it overlaps with liberties like freedom of movement or freedom of religion.[156] Stronger models of indirect discrimination require that responsibilities and burdens of state and private actors are specified. In many cases it will be fair, for example, that employers do not have to bear the burden of existing societal inequalities, but that they refrain from perpetuating or deepening these inequalities.[157] Moreover, it seems helpful to specify particular harms caused in different situations that merit different forms of responses by non-discrimination law, for example redressing disadvantaging, addressing stereotypes, enhancing participation, or achieving structural change as proposed by *Sandra Fredman*.[158]

This chapter submits that the use of indirectly discriminatory algorithms also merits considerable scrutiny, for at least two reasons. First, big data analysis facilitates the linkage of innocuous data to sensitive characteristics. If internet platforms can infer characteristics like gender, sexual orientation, health conditions, or purchasing power from your online behaviour, they do not need to ask for this sensitive data in order to use it. This situation can be compared to the circumvention scenario that even 'weak' models of indirect discrimination intend to prevent. Second, it is increasingly difficult to distinguish between direct and indirect discrimination. The more complex profiling algorithms become and the more autonomously they operate, the more difficult it is to identify the relevant predictor variables (i.e. to tell whether profiling directly

[151] *Cf.* CJEU, C-152/73 *Sotgiu* [1974]; P Craig and J de Búrca, *EU Law* (7th ed., 2020) 796–797.

[152] Different weak and strong models are developed by Schiek, 'Indirect Discrimination'(n 41) 323–333 (circumvention vs. social engineering); Connolly, *Discrimination Law* (n 38) 153–156 (pretext, functional equivalency, quota model); Tobler, *Limits* (n 45) 24 (effectiveness of discrimination law and challenges the underlying causes of discrimination); see also Morris, 'On the Normative Foundations' (n 45) (corrective and distributive justice); M Grünberger, *Personale Gleichheit* (2013) 657–661 (hereafter Grünberger, *Personale Gleichheit*) (individual and group justice); S Fredman, 'Substantive Equality Revisited' (2016) 14 *I·CON* 713(hereafter Fredman '*Substantive Equality Revisited*') (formal and substantive equality); Wachter, Mittelstadt, and Russell, 'Bias Preservation'(n 12) para 2 (formal and substantive equality).

[153] This is a common position in Germany, *cf.* M Fehling, 'Mittelbare Diskriminierung und Artikel 3 (Abs. 3) GG' in D Heckmann, R Schenke, and G Sydow (eds) *Festschrift für Thomas Würtenberger* (2013) 668 (675).

[154] Wachter, Mittelstadt, and Russell, 'Bias Preservation'(n 12) para 2.1.1.

[155] Schiek, 'Indirect Discrimination' (n 41) 327.

[156] *Cf.* n 151 on freedom of movement; CJEU, Case C-157/15 *Achbita* [2017], and CJEU, Case C-188/15 *Bougnaoui* [2017] on freedom of religion, *cf.* also L Vickers, 'Indirect Discrimination and Individual Belief: Eweida v British Airways plc' (2009) 11 *Ecclesiastical Law Journal* 197.

[157] Grünberger, *Personale Gleichheit* (n 152) 660–661.

[158] Fredman, '*Substantive Equality Revisited*' (n 152).

https://doi.org/10.1017/9781009207898.020 Published online by Cambridge University Press

includes a forbidden characteristic or not). In addition to this epistemic challenge, normative questions concerning the difference between direct and indirect discrimination arise. If a complex profile comprises 250 data points, among them one sensitive one (for instance gender) and 50 data points related to this sensitive characteristic (for example attributes typical of a certain gender), does using this profile involve different treatment or lead to detrimental impact? What if it cannot be established if the one sensitive data point was decisive for a particular outcome? The detrimental effect of profiling might be easier to prove than differential treatment because the output of profiling algorithms can be more easily tested than its internal decision-making criteria, especially with increasingly autonomous, self-learning, and opaque algorithms.[159] Because of this, it might be more helpful for the people affected and also more predictable for the users of profiling algorithms to assume indirect discrimination, but at the same time also to apply stricter scrutiny.

The broader the reach of indirect discrimination becomes, the more relevant the standards of justification will be.[160] Developing these standards will, therefore, be a crucial task in coping with discriminatory AI and in attributing responsibilities in the fight against factual discrimination. In part, these standards might be developed in view of existing ones. EU anti-discrimination law establishes, for example, that companies cannot justify discrimination against their employees by relying on customers' preferences, for these are not considered 'genuine and determining occupational requirements'.[161] The reasoning is also applicable to indirect forms of discrimination based on (predicted) customers' preferences and could therefore exclude a justification of policies or measures based on profiling. Moreover, as explained earlier, justification standards for both direct and indirect discrimination also depend on technical factors such as the possibilities and costs of avoiding discrimination. In the context of indirect discrimination, this might be relevant for errors in personalised (as opposed to group) profiling. Take the example of face recognition which yields particularly high error rates for Black women and low error rates for White men.[162] This could mean that Black women cannot use technical devices based on image recognition or that unnecessary law enforcement activities are directed against them. Provided that applying an algorithm with unequal error rates is covered by anti-discrimination law, that is, if it amounts to an apparently neutral practice that puts members of a protected group at a particular disadvantage,[163] one should ask how costly it would be to reduce error rates and how useful it would be to rely on other techniques until error rates are reduced.

V. CONCLUSION

Law is not silent on discriminatory AI. Existing rules of anti-discrimination law and data protection law do cover decision-making based on profiling. This chapter aims to show that the legal requirement to justify direct and indirect forms of discrimination implies that profiling

[159] On this F Pasquale, *The Black Box Society* (2015).

[160] Generally, on this point C McCrudden, 'The New Architecture of EU Equality Law after CHEZ' (2016) *European Equality Law Review* 1 (9).

[161] CJEU, C-188/15 *Bougnaoui* [2017] para 37–41.

[162] J Buolamwini and T Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' (2018) 81 *Proceedings of Machine Learning Research* 1.

[163] The question which factual disadvantages are covered by anti-discrimination law cannot be treated here in detail. Traditionally, anti-discrimination law applies to differential treatment or detrimental impact as a result of legal acts (e.g. contractual terms, the refusal to conclude a contract, employers' instructions, statutes, law enforcement acts). But the wording of anti-discrimination law does not exclude factual disadvantages like a malfunctioning device, which might thus also trigger anti-discrimination provisions.

must follow methodological minimum standards. It remains a very important task for lawyers to specify these standards in case law or – preferably – legislation. For this, lawyers need to cooperate with data or computer scientists in order to assess the validity of profiling and to evaluate alternative methods by considering the discriminatory effects of sampling bias, labelling bias, and feature selection bias or the distribution of error rates.

The EU commission has recently published a proposal for the regulation of AI, the 'EU Artificial Intelligence Act'.[164] This piece of legislation would indeed specify relevant standards significantly. According to the proposal, AI systems classified as 'high risk' have to comply with requirements which reflect the idea that AI systems should produce valid results and must not cause any harm that cannot be justified. The Act stipulates, for example, that high risk systems have to be tested 'against preliminary defined metrics and probabilistic thresholds that are appropriate to the intended purpose',[165] that training, validation, and testing data must be 'relevant, representative, free of errors and complete' and shall have the 'appropriate statistical properties',[166] that data governance must include bias monitoring,[167] that the systems achieve 'in the light of their intended purpose, an appropriate level of accuracy'[168] and that 'levels of accuracy and the relevant accuracy metrics' have to be declared in the instructions of use.[169] As many of the AI systems known for their discrimination risks are classified as 'high risk'[170] or may be classified accordingly by the Commission in the future,[171] this is already a good start.

[164] EU Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 21st April 2021, COM/2021/206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206; *Cf.* T Burri, Chapter 7, in this volume.

[165] EU Artificial Intelligence Act, Article 9(7).

[166] EU Artificial Intelligence Act, Article 10(3).

[167] EU Artificial Intelligence Act, Article 10(2)(f) and (5).

[168] EU Artificial Intelligence Act, Article 15(1).

[169] EU Artificial Intelligence Act, Article 15(2).

[170] For example those used for predicting job performance, creditworthiness, or crime. See EU Artificial Intelligence Act, Annex III.

[171] EU Artificial Intelligence Act, Article 7.

https://doi.org/10.1017/9781009207898.020 Published online by Cambridge University Press