Regulating the Black Box – Prevention of Discrimination in Automated Decision-Making

Anna Haipola
503763
Fundamental Rights and Human Rights in the Information Society
University of Turku, Faculty of Law
28 November 2019
Graduate Thesis, Master of Laws

UNIVERSITY OF TURKU
Faculty of Law

ANNA HAIPOLA: Regulating the Black Box – Prevention of Discrimination in Automated Decision-Making
Graduate Thesis, Master of Laws, XX + 76 pages
Constitutional Law
November 2019

Keywords: discrimination, data protection, automated decision-making, general principle of equal treatment, prohibition of discrimination, indirect discrimination, direct discrimination, human rights, fundamental rights, purpose limitation, data minimisation, data accuracy

Machine learning enables efficient processing of big data, which can be utilised in algorithmic decision-making. When decisions that affect individuals, such as recruitment and credit decisions, are outsourced to machines, questions of discrimination and data protection arise. Discrimination may occur in automated decision-making if the decisional rules in the algorithm are either directly or indirectly discriminating or if there are issues with data quality. In the European Union, individuals are protected from discrimination by private entities based on several directives with different scopes when it comes to the protected grounds (e.g. gender, ethnicity or age) and area of regulation (e.g. employment or sale of goods and services). In addition, the non-discrimination provisions of the European Convention of Human Rights and the Charter of Fundamental Rights of the European Union have a horizontal direct effect to some extent.

The general principle of equality and the prohibition of discrimination based on a protected ground both have an effect to automated decision-making. Direct discrimination can be found in the context of machine learning if the automated decision treats a person belonging to a protected group in a less favourable way in comparison to another person in a similar situation and this difference is based directly on a forbidden ground. Indirect discrimination is more challenging to spot in algorithms because the rules in the decision-making model may appear neutral while having the side effect of discriminating against one of the specific forbidden grounds. The differentiation may be based on an objective attribute, such as a ZIP code, but in fact, the objective attribute may be correlated with a protected ground, such as ethnicity.

The General Data Protection Regulation sets a general prohibition for the use of automated decision-making, as well as several exceptions to the prohibition. The principle of purpose limitation restricts the possibilities to reuse personal data originally collected for a certain purpose in an automated decision-making model. The principle of data minimisation obliges the controllers to keep the amount of data at a minimum for the purpose of processing. The principle of data accuracy supports non-biased automatic decision-making by requiring that inaccurate data is rectified.

There are several technical solutions aimed at the creation of accountable algorithms but a legal analysis on the legitimacy of such methods is missing. The main purpose of this thesis is to bring clarity to the issue of discrimination in automated decision-making from a legal point of view, suggesting that wider interdisciplinary research, especially with legal expertise on non-discrimination legislation, is needed.

TURUN YLIOPISTO
Oikeustieteellinen tiedekunta

ANNA HAIPOLA: Regulating the Black Box – Prevention of Discrimination in Automated Decision-Making
Pro gradu -tutkielma, XX + 76 s.
Valtiosääntöoikeus
Marraskuu 2019
Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck -järjestelmällä.

Asiasanat: syrjintä, tietosuoja, automaattinen päätöksenteko, yhdenvertaisuus, syrjinnän kielto, välillinen syrjintä, välitön syrjintä, ihmisoikeudet, perusoikeudet, käyttötarkoitussidonnaisuus, tietojen minimointi, täsmällisyys

Koneoppiminen mahdollistaa big datan tehokkaan käsittelyn, mitä voidaan hyödyntää algoritmisessa päätöksenteossa. Etenkin yksilöille merkittävien päätösten, kuten rekrytoinnin ja lainapäätösten, ulkoistaminen automaattisille järjestelmille herättää kysymyksiä liittyen tietosuojaan ja syrjimättömyyteen. Automaattinen päätöksenteko voi johtaa syrjintään, jos algoritmin päätössäännöt ovat välillisesti tai välittömästi syrjiviä, tai jos data on huonolaatuista. Euroopan unionissa yksilöitä suojataan yksityisten toimijoiden syrjinnältä erilaisin direktiivein, jotka eroavat toisistaan siinä, mitkä henkilöön liittyvät syyt (esim. sukupuoli, etninen tausta tai ikä) otetaan huomioon ja mitä osa-aluetta (esim. työ tai tavaroiden tai palveluiden tarjoaminen) sääntely koskee. Lisäksi Euroopan ihmisoikeussopimuksen ja Euroopan unionin perusoikeuskirjan syrjintää koskevia säännöksiä voidaan jossain määrin suoraan soveltaa myös yksityisten välisissä suhteissa.

Yhdenvertaisuusperiaate ja syrjinnän kielto tiettyjen henkilöön liittyvien syiden perusteella tulee ottaa huomioon automaattisessa päätöksenteossa. Koneoppiva malli voidaan todeta välittömästi syrjiväksi, jos automaattisesti tehty päätös asettaa huonompaan asemaan henkilön, joka kuuluu suojattuun vähemmistöön ja erona tämän ja verrattavan henkilön välillä on ainoastaan kuuluminen kyseiseen vähemmistöön. Välillistä syrjintää on vaikeampi havaita algoritmeista, koska tuolloin säännöt, joihin päätöksenteko perustuu, vaikuttavat neutraaleilta mutta asettavat tosiasiassa vähemmistöön kuuluvat heikompaan asemaan. Erottelu päätöksenteon kohteiden välillä saattaa perustua objektiiviseen tekijään, kuten postinumeroon, mutta tämän tekijän ja henkilöön liittyvän syyn, kuten etnisen taustan, välillä voi olla korrelaatio.

Yleinen tietosuoja-asetus pääsääntöisesti kieltää automaattisen päätöksenteon, mutta tähän kieltoon on useita poikkeuksia. Käyttötarkoitussidonnaisuuden periaate rajoittaa mahdollisuuksia käyttää automaattisen päätöksentekosovelluksen kehittämisessä henkilötietoja, jotka on alun perin kerätty toista tarkoitusta varten. Tietojen minimoinnin periaate velvoittaa rekisterinpitäjiä käsittelemään ainoastaan tarkoitusta varten tarpeellisen määrän tietoja. Täsmällisyyden periaate tukee syrjimätöntä automaattista päätöksentekoa siten, että se velvoittaa oikaisemaan epätarkat ja virheelliset henkilötiedot.

On olemassa useita teknisiä ratkaisuja algoritmien vastuulliseen ja läpinäkyvään käyttöön. Näitä keinoja ei ole kuitenkaan analysoitu oikeustieteen näkökulmasta. Tämän tutkielman ensisijainen tarkoitus on selkeyttää syrjintään liittyviä ongelmia automaattisessa päätöksenteossa oikeustieteen keinoin. Laajempi poikkitieteellinen tutkimus aiheesta on tarpeen, erityisesti tutkimus, jossa hyödynnetään syrjimättömyyttä koskevan lainsäädännön asiantuntemusta.

Table of Contents

Sources

Literature

Barocas – Selbst (2016)
Barocas, Solon – Selbst, Andrew D., Big Data's Disparate Impact. California Law Review 104 (3) 2016, pp. 671–732.

Bayamlioglu (2018)
Bayamlioglu, Emre, Contesting Automated Decisions. European Data Protection Law Review (EDPL), 4 (4), 2018, pp. 433–446.

Calders – Žliobaitė (2013)
Calders, Toon – Žliobaitė, Indrė, Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures, pp. 43–57 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Citron (2008)
Citron, Danielle Keats, Technological Due Process. Washington University Law Review 85 (6) 2008, pp. 1249–1313.

Colonna (2014)
Colonna, Liana, Data Mining and Its Paradoxical Relationship to the Purpose Limitation Principle, in Re-loading Data Protection, pp. 299–321 in Gutwirth, Serge – Leenes, Ronald – de Hert, Paul (eds.), Multidisciplinary Insights and Contemporary Challenges. Springer, Dordrecht 2014.

Comandé (2017)
Comandé, Giovanni, Regulating Algorithms' Regulation? First Ethico-Legal Principles, Problems and Opportunities of Algorithms, pp. 169–206 in Cerquitelli, Tania – Quercia, Daniele – Pasquale, Frank (eds.), Transparent Data Mining for Big and Small Data. Springer International Publishing AG 2017.

Crawford – Schultz (2014)
Crawford, Kate – Schultz, Jason, Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. Boston College Law Review 55 (1) 2014, pp. 93–128.

Custers (2013)
Custers, Bart, Data Dilemmas in the Information Society: Introduction and Overview, pp. 4–26 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Edwards – Veale (2017)
Edwards, Lilian – Veale, Michael, Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. Duke Law & Technology Review 16 (18) 2017, pp. 18–84.

Edwards – Veale (2018)
Edwards, Lilian – Veale, Michael, Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'? IEEE Security & Privacy 16 (3) 2018, pp. 46–54.

Finocchiaro – Ricci (2013)
Finocchiaro, Giusella – Ricci, Annarita, Quality of Information, the Right to Oblivion and Digital Reputation, pp. 289–300 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Forgó – Hänold – Schütze (2017)
Forgó, Nikolaus – Hänold, Stefanie – Schütze, Benjamin, The Principle of Purpose Limitation and Big Data, in New Technology, Big Data and the Law, edited by Corrales, Marcelo – Fenwick, Mark – Forgó, Nikolaus. Springer Nature Singapore Pte Ltd. 2017, pp. 17–42.

Gellert et al. (2013)
Gellert, Raphaël – de Vries, Katja – de Hert, Paul – Gutwirth, Serge, A Comparative Analysis of Anti-Discrimination and Data Protection Legislations, pp. 61–90 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Goodman – Flaxman (2016)
Goodman, Bryce – Flaxman, Seth, European Union Regulations on Algorithmic Decision-making and a "Right to Explanation". Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, 2016, pp. 1–6.

Hajian – Domingo-Ferrer (2013)
Hajian, Sara – Domingo-Ferrer, Josep, Direct and Indirect Discrimination Prevention Methods, pp. 241–255 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Harvard Law Review (2017)
Recent Cases, State v. Loomis. Harvard Law Review 130 (5) 2017, pp. 1530–1537.

de Hert – Papakonstantinou (2016)
de Hert, Paul – Papakonstantinou, Vagelis, The new General Data Protection Regulation: Still a sound system for the protection of individuals? Computer Law & Security Review 32 (2) 2016, pp. 179–194.

Kaminski – Malgieri (2019)
Kaminski, Margot – Malgieri, Gianclaudio: Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations. University of Colorado Law Legal Studies Research Paper No. 19–28.

Kamiran – Calders – Pechenizkiy (2013)
Kamiran, Faisal – Calders, Toon – Pechenizkiy, Mykola, Techniques for Discrimination-Free Predictive Models, pp. 223–241 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Kamiran – Žliobaitė (2013)
Kamiran, Faisal – Žliobaitė, Indrė, Explainable and Non-explainable Discrimination in Classification, pp. 155–171 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Karanasiou – Pinotsis (2017)
Karanasiou, Argyro – Pinotsis, Dimitris, A Study into the Layers of Automated Decision-making: Emergent Normative and Legal Aspects of Deep Learning. International Review of Law, Computers & Technology 31 (2) 2017, pp. 170–187.

Kemppinen (2011)
Kemppinen, Jukka, Informaatio-oikeuden alkeet. Tietosanoma 2011.

Koskinen (2018)
Koskinen, Ida, Koneoppiminen ja EU:n yleisen tietosuoja-asetuksen vaatimus lainmukaisesta, kohtuullisesta ja läpinäkyvästä käsittelystä. Defensor Legis 2/2018, pp. 240–256.

Kroll et al. (2017)
Kroll, Joshua A. – Huey, Joanna – Barocas, Solon – Felten, Edward W. – Reidenberg, Joel R. – Robinson, David G., – Yu, Harlan, Accountable Algorithms. University of Pennsylvania Law Review 165 (3) 2017, pp. 633–706.

Kuner et al. (2012)
Kuner, Christopher – Cate, Fred H. – Millard, Christopher – Svantesson, Dan Jerker B, The challenge of 'big data' for data protection. International Data Privacy Law 2 (2) 2012, pp. 47–49.

Leonard (2014)
Leonard, Peter, Customer Data Analytics: Privacy Settings for 'Big Data' Business. International Data Privacy Law 4 (1) 2014, pp. 53–68.

Mittelstadt et al. (2016)
Mittelstadt, Brent Daniel – Allo, Patrick – Taddeo, Mariarosaria – Wachter, Sandra – Floridi, Luciano, The Ethics of Algorithms: Mapping the Debate. Big Data & Society, 3 (2) 2016, pp. 1–21.

Moerel – Prins (2016)
Moerel, Lokke – Prins, Corien, Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things. Tilburg Institute for Law, Technology, and Society (TILT), 25 May 2016, pp. 1–98. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2784123. Accessed 26 September 2019.

Ohm (2010)
Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review 57 (6) 2010, pp. 1703–1777.

Ojanen – Scheinin (2011)
Ojanen, Tuomas – Scheinin, Martin, Yhdenvertaisuus ja syrjinnän kielto (part III, chapter 2), in Perusoikeudet, edited by Hallberg, Pekka – Karapuu, Heikki – Ojanen, Tuomas – Scheinin, Martin – Tuori, Kaarlo – Viljanen, Veli-Pekka. E-book, Talentum. The authors have updated part III on 13 January 2013. Accessed 28 April 2019. Available at Alma Talent Fokus service.

O'Neil (2017)
O'Neil, Cathy, Weapons of Math Destruction – How Big Data Increases Inequality and Threatens Democracy. Crown 2017.

Papadopoulos (2011)
Papadopoulos, Thomas, Criticising the Horizontal Direct Effect of the EU General Principle of Equality. European Human Rights Law Review (4) 2011, pp. 437–447.

Pasquale (2015)
Pasquale, Frank, The Black Box Society – The Secret Algorithms That Control Money and Information. Harvard University Press 2015.

Pedreschi – Ruggieri – Turini (2008)
Pedreschi, Dino – Ruggieri, Salvatore – Turini, Franco, Discrimination-Aware Data Mining. Conference paper. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24–27, 2008.

Pedreschi – Ruggieri – Turini (2013)
Pedreschi, Dino – Ruggieri, Salvatore – Turini, Franco, The Discovery of Discrimination, pp. 91–109 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Pöysti (2018)
Pöysti, Tuomas, Kohti digitaalisen ajan hallinto-oikeutta. Lakimies 7–8/2018, pp. 868–903.

Reed (2007)
Reed, Chris, Taking Sides on Technology Neutrality. SCRIPTed 4 (3) 2007, pp. 263–284.

Reinisch (2012)
Reinisch, August, Essentials of EU Law. Cambridge University Press 2012.

Romei – Ruggieri (2013)
Romei, Andrea – Ruggieri, Salvatore, Discrimination Data Analysis: A Multi-disciplinary Bibliography, pp. 109–137 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Rosas (2015)
Rosas, Allan, Five Years of Charter Case Law: Some Observations, pp. 11–20 in de Vries, Sybe – Bernitz, Ulf – Weatherill, Stephen (eds.), The EU Charter of Fundamental Rights as a Binding Instrument. Hart Publishing 2015.

Sandvig et al. (2014)
Sandvig, Christian – Hamilton, Kevin – Karahalios, Karrie – Langbort, Cedric, Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. Paper presented to 'Data and Discrimination: Converting Critical Concerns into Productive Inquiry,' a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.

van der Sloot (2013)
Van der Sloot, Bart, From Data Minimization to Data Minimummization, pp. 273–288 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Staab – Stalla-Bourdillon – Carmichael (2016)
Staab, Steffen – Stalla-Bourdillon, Sophie – Carmichael, Laura, Observing and Recommending from a Social Web with Biases. Web Science Institute (WSI) Pump-priming Project, 26 January – 11 March 2016, University of Southampton.

Tutt (2016)
Tutt, Andrew, An FDA for Algorithms. Administrative Law Review 69 (83) 2017, pp. 83–123.

Veale – Binns (2017)
Veale, Michael – Binns, Reuben, Fairer Machine Learning in the Real World: Mitigating Discrimination Without Collecting Sensitive Data. Big Data & Society 4 (2) 2017, pp. 1–17.

Verwer – Calders (2013)
Verwer, Sicco – Calders, Toon, Introducing Positive Discrimination in Predictive Models, pp. 255–273 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Viljanen (2018)
Viljanen, Mika, Algoritmien haaste – uuteen aineelliseen oikeuteen? Lakimies 7–8/2018, pp. 1070–1087.

Voigt – von dem Bussche (2017)
Voigt, Paul – von dem Bussche, Axel, The EU General Data Protection Regulation (GDPR): A Practical Guide. Springer International Publishing.

Wachter – Mittelstadt – Floridi (2017)
Wachter, Sandra – Mittelstadt, Brent – Floridi, Luciano, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7 (2) 2017, pp. 76–99.

White – Ovey (2010)
White, Robin C.A. – Ovey, Clare, Jacobs, White, and Ovey: The European Convention on Human Rights. Sixth edition. Oxford University Press 2010.

Zarsky (2013)
Zarsky, Tal, Transparency in Data Mining: From Theory to Practice, pp. 301–324 in Custers, Bart – Calders, Toon – Schermer, Bart – Zarsky, Tal (eds.), Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases. Springer-Verlag Berlin Heidelberg 2013.

Zarsky (2016)
Zarsky, Tal, Incompatible: The GDPR in the Age of Big Data. Seton Hall Law Review 47 (995) 2016, pp. 995–1020.

Žliobaitė (2017)
Žliobaitė, Indrė, Measuring Discrimination in Algorithmic Decision Making. Data Mining and Knowledge Discovery, 31 (4) 2017, pp. 1060–1089.

Žliobaitė – Custers (2016)
Žliobaitė, Indrė – Custers, Bart, Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models. Artificial Intelligence and Law 24 (2) 2016, pp. 183–201.

Official Sources

Ailisto et al. (2018)
Ailisto, Heikki (ed.) – Heikkilä, Eetu – Helaakoski, Heli – Neuvonen, Anssi – Seppälä, Timo, Artificial intelligence and its capability assessment. Publications of the Government's analysis, assessment and research activities 46/2018. Prime Minister's Office, Helsinki 2018.

Council of Europe (2000)
Council of Europe, Explanatory Report to the Protocol No. 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 177). Available at https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/177.

Council of the European Union (2019)
Council of the European Union, Report from Presidency to Permanent Representations Committee (Part 1) / Council on the Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation – Progress Report. ST 9567 2019 REV 1. Brussels, 27 May 2019.

Council of the European Union (2016)
Council of the European Union, Position (EU) No 6/2016 of the Council at first reading with a view to the adoption of a Regulation of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Adopted by the Council on 8 April 2016 (2016/C 159/01).

European Commission (2010)
Comparative Study on Different Approaches to New Privacy Challenges, in Particular in the Light of Technological Developments. Final Report. 20 January 2010.

European Commission (2012)
Comparative study of anti-discrimination and equality laws of the US, Canada, South Africa and India. Written by Sandra Fredman. European network of legal experts in the non-discrimination field. Directorate-General for Justice. February 2012.

European Commission (2015)
Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee: Report on the application of Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services. 5 May 2015.

European Commission (2018)
A Comparative Analysis of Non-Discrimination Law in Europe 2018. Prepared by Isabelle Chopin, Carmine Conte and Edith Chambrier for the European network of legal experts in gender equality and non-discrimination. Directorate-General for Justice and Consumers. November 2018.

European Data Protection Board (2018)
European Data Protection Board, Endorsement 1/2018. Brussels, 25 May 2018.

European Data Protection Supervisor (2012)
Opinion of the European Data Protection Supervisor on the Data Protection Reform Package. 7 March 2012.

European Data Protection Supervisor (2016)
Opinion of the European Data Protection Supervisor on Coherent Enforcement of Fundamental Rights in the Age of Big Data. Opinion 8/2016. 23 September 2016.

European Data Protection Supervisor (2017)
Assessing the Necessity of Measures that Limit the Fundamental Right to the Protection of Personal Data: A Toolkit. 11 April 2017.

European Parliament (2013)
Report on transposition and application of Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services (2010/2043(INI)). Committee on Women's Rights and Gender Equality.

European Union Agency for Fundamental Rights and Council of Europe (2018)
European Union Agency for Fundamental Rights and Council of Europe, Handbook on European Non-Discrimination Law. 2018 edition. Publications Office of the European Union 2018.

Government proposal 309/1993
Government proposal 309/1993 (Finland), Hallituksen esitys Eduskunnalle perustuslakien perusoikeussäännösten muuttamisesta, HE 309/1993 vp.

Government proposal 44/2003
Government proposal 44/2003 (Finland), Hallituksen esitys Eduskunnalle laiksi yhdenvertaisuuden turvaamisesta sekä eräiden siihen liittyvien lakien muuttamisesta, HE 44/2003 vp.

Information Commissioner's Office (2017)
The United Kingdom's Information Commissioner's Office, Big data, artificial intelligence, machine learning and data protection. Version 2.2, 4 September 2017. Available at https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf. Accessed 5 June 2019.

Koulu et al. (2019)
Koulu, Riikka – Mäihäniemi, Beata – Kyyrönen, Vesa– Hakkarainen, Jenni– Markkanen, Kalle, Algorithm as a decision-maker? The possibilities and challenges of artificial intelligence in the national regulatory environment. Publication series of the Government's analysis, assessment and research 2019:44. Prime Minister's Office, Helsinki 2019.

Parliamentary reply 95/2003
Parliamentary reply 95/2003 (Finland), Eduskunnan vastaus 95/2003 vp.

WP29 Guidelines (2017a)
Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. 17/EN, WP 251. Adopted on 3 October 2017. As last Revised and Adopted on 6 February 2018.

WP29 Guidelines (2017b)
Article 29 Data Protection Working Party, Guidelines on Consent under Regulation 2016/679. 17/EN, WP 251. Adopted on 28 November 2017. As last Revised and Adopted on 10 April 2018.

WP29 Opinion 03/2013
Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation, 00569/13/EN, WP 203. 2 April 2013.

WP29 Opinion 05/2014
Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 0829/14/EN, WP216. 10 April 2014.

Case Law

European Court of Justice

C-144/04
Court of Justice of the European Communities, Case C-144/04 *Werner Mangold v Rüdiger Helm*. 22 November 2005. ECLI:EU:C:2005:709.

C-303/06
Court of Justice of the European Communities, Case C-303/06 *S. Coleman v Attridge Law and Steve Law*. 17 July 2008. ECLI:EU:C:2008:415.

C-54/07
Court of Justice of the European Communities, Case C-54/07 *Centrum voor gelijkheid van kansen en voor racismebestrijding v Firma Feryn NV*. 10 July 2008. ECLI:EU:C:2008:397.

C-555/07
Court of Justice of the European Communities, Case C-555/07 *Seda Kücükdeveci v Swedex GmbH & Co. KG*. 19 January 2010. ECLI:EU:C:2010:21.

C-236/09
European Court of Justice, Case C-236/09 *Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres*. 1 March 2011. ECLI:EU:C:2011:100.

C-617/10
European Court of Justice, Case C-617/10 *Åklagaren v Hans Åkerberg Fransson*. 26 February 2013. ECLI:EU:C:2013:105.

Joined Cases C-293/12 and C-594/12
European Court of Justice, Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others*. 8 April 2014. ECLI:EU:C:2014:238.

C-414/16
European Court of Justice, Case C-414/16 *Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung eV.* 17 April 2018. ECLI:EU:C:2018:257.

C-68/17
European Court of Justice, Case C-68/17 *IR v JQ*. 11 September 2018. ECLI:EU:C:2018:696.

Joined Cases C-569/16 and C-570/16
European Court of Justice, Joined Cases C-569/16 and C-570/16 *Stadt Wuppertal v Maria Elisabeth Bauer* (C-569/16) and *Volker Willmeroth, in his capacity as owner of TWI Technische Wartung und Instandsetzung Volker Willmeroth e.K. v Martina Broßonn* (C-570/16). 6 November 2018. ECLI:EU:C:2018:871.

European Court of Human Rights

Belgian Linguistic Case
Case 'relating to certain aspects of the laws on the use of languages in education in Belgium' v. Belgium, 1474/62; 1677/62; 1691/62; 1769/63; 1994/63; 2126/64, 23 July 1968. ECLI:CE:ECHR:1968:0723JUD000147462.

Biao v. Denmark
Biao v. Denmark, 38590/10, Grand Chamber, 24 May 2016. ECLI:CE:ECHR:2016:0524JUD003859010.

D.H. and Others v. the Czech Republic
D.H. and Others v. the Czech Republic, 57325/00, Grand Chamber, 13 November 2007. ECLI:CE:ECHR:2007:1113JUD005732500.

Posti and Rahko v. Finland
Posti and Rahko v. Finland, 27824/95, 24 September 2002. ECLI:CE:ECHR:2002:0924JUD002782495.

Thlimmenos v. Greece
Thlimmenos v. Greece, 34369/97, Grand Chamber, 6 April 2000. ECLI:CE:ECHR:2000:0406JUD003436997.

Weller v. Hungary
Weller v. Hungary, 44399/05, 31 March 2009. ECLI:CE:ECHR:2009:0331JUD004439905.

X and Y v. the Netherlands
X and Y v. the Netherlands, 8978/80, 26 March 1985. ECLI:CE:ECHR:1985:0326JUD000897880.

Z and others v. the United Kingdom
Z and others v. the United Kingdom, 29392/95, Grand Chamber, 10 May 2001. ECLI:CE:ECHR:2001:0510JUD002939295.

Others

Finland's National Non-Discrimination and Equality Tribunal 216/2017
Decision 216/2017 by Finland's National Non-Discrimination and Equality Tribunal, dated 21 March 2018.

Finland's National Non-Discrimination and Equality Tribunal 337/2018
Decision 337/2018 by Finland's National Non-Discrimination and Equality Tribunal, dated 19 December 2018.

Office of the Data Protection Ombudsman (2019a)
Decision 2278/452/17 by the Office of the Data Protection Ombudsman in Finland, dated 15 February 2019.

Office of the Data Protection Ombudsman (2019b)
Decision 1387/44/19 by the Office of the Data Protection Ombudsman in Finland, dated 15 February 2019.

Parliamentary Ombudsman of Finland 3116/2017
Decision 3116/2017 by the Parliamentary Ombudsman of Finland, dated 29 June 2018.

Parliamentary Ombudsman of Finland 3393/2017
Decision 3393/2017 by the Parliamentary Ombudsman of Finland, dated 29 June 2018.

Parliamentary Ombudsman of Finland 3379/2018
Decision 3379/2018 by the Parliamentary Ombudsman of Finland, dated 10 September 2018.

State v. Loomis
State v. Loomis, Supreme Court of Wisconsin, No. 2015AP157-CR. 881 N.W.2d 749 (Wis. 2016). 13 July 2016.

Online Sources

Elements of AI (2018)
Elements of AI, online course by Reaktor and University of Helsinki, 2018. Available at https://www.elementsofai.com. Accessed 1 May 2019.

The English Oxford Living Dictionary
The English Oxford Living Dictionary, https://www.lexico.com/en. Accessed 19 October 2019.

Espinhaço Gomes (2019)
Espinhaço Gomes, Inês, Queering European Union Law: Sex and Gender Beyond the Binary and Cisnormativity. Europa-Kolleg Hamburg, Institute for European Integration, Study Paper No 04/19. Available at http://www.europa-kolleg-hamburg.de. Accessed 14 October 2019.

European network of legal experts in gender equality and non-discrimination (2019)
Flash report 'Police stop and search found to be discriminatory' by the European network of legal experts in gender equality and non-discrimination, 20 February 2019. Available at https://www.equalitylaw.eu/downloads/4840-finland-police-stop-and-search-found-to-be-discriminatory-pdf-111-kb. Accessed 15 October 2019.

FATML Website
Website of Fairness, Accountability, and Transparency in Machine Learning https://www.fatml.org/.

Finlex Tool for the Assessment of Equality
Tool for the Assessment of Equality, published in the internet service on legal information Finlex, owned by the Finnish Ministry of Justice. Available at http://yhdenvertaisuus.finlex.fi/en/. Accessed 14 October 19.

Forbes (2016)
Marr, Bernard, What Is The Difference Between Deep Learning, Machine Learning and AI? Forbes, 8 December 2016. Available at https://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/#87f9d426cfa4. Accessed 14 September 2019.

Frantziou (2018)
Frantziou, Eleni, Joined cases C-569/16 and C-570/16 Bauer et al: (Most of) the Charter of Fundamental Rights is Horizontally Applicable. European Law Blog, 19 November 2018. Available at https://europeanlawblog.eu/2018/11/19/joined-cases-c-569-16-and-c-570-16-bauer-et-al-most-of-the-charter-of-fundamental-rights-is-horizontally-applicable/. Accessed 12 November 2019.

Gartner IT Glossary
Gartner IT glossary, http://www.gartner.com/it-glossary/. Accessed 14 September 2019.

The Independent (2018)
Beduschi, Ana, How artificial intelligence could be violating our human rights. The Independent, 8 October 2018. This article first appeared on The Conversation (theconversation.com). Available at https://www.independent.co.uk/life-style/gadgets-and-tech/artificial-intelligence-ai-human-rights-data-protection-privacy-algorithms-gdpr-discrimination-a8563341.html. Accessed 14 September 2019.

Indrė Žliobaitė's home page
Indrė Žliobaitė's home page https://www.zliobaite.com/. Link to a slide presentation 'Fairness-aware Machine Intelligence': https://4c7b2f8e-a-62cb3a1a-s-sites.googlegroups.com/site/zliobaitefiles2/discrimination_MLcoffee.pdf?attachauth=ANoY7crpLVG4r4vBs3kgZz71QZTdqO9gaYEvAQXTC2QZR0UmTu8w8bCmcejBbv1jX7rjHMk70f5KxojXahhcU03CrMHv8bwK5q800rsXnLueWWDDiE7p8GwGiRT-GYDAjq0FjQh_m5T6HGTwDqayeVFbt3Q9XkfRqbOGH9JZ4x2C2_LfZA0byIV1fiqaN-qiyvr9vzEXkHjpMyrQZyIUNzrl1LLUJrpTTZ_EZ-AWzi497CqRibaHhp0-Y%3D&attredirects=0

International Open Data Charter
International Open Data Charter, available at https://opendatacharter.net/wp-content/uploads/2015/10/opendatacharter-charter_F.pdf. Accessed 1 May 2019.

The New York Times (2016)
Smith, Mitch, In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures. The New York Times, 22 June 2016. Available at https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html?module=inline. Accessed 3 October 2019.

Politico (2019)
Kayali, Laura, Next European Commission takes aim at AI – Artificial intelligence will be the next GDPR. Politico, 18 July 2019. Available at https://www.politico.eu/article/ai-data-regulator-rules-next-european-commission-takes-aim/. Accessed 19 September 2019.

Reuters (2018)
Dastin, Jeffrey, Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 10 October 2018. Available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G. Accessed 14 September 2019.

University of Helsinki Website
Course description of the course 'Fairness Aware AI' by the University of Helsinki at https://courses.helsinki.fi/en/data20006.

Washington Post (2012)
Shaver, Katherine, Female dummy makes her mark on male-dominated crash tests. Washington Post, 25 March 2012. Available at https://www.washingtonpost.com/local/trafficandcommuting/female-dummy-makes-her-mark-on-male-dominated-crash-tests/2012/03/07/gIQANBLjaS_story.html?noredirect=on&utm_term=.7f4d99583bf1. Accessed 15 June 2019.

Zhang – Lemoine – Mitchell (2018)
Zhang, Brian Hu – Lemoine, Blake – Mitchell, Margaret, Mitigating Unwanted Biases with Adversarial Learning. Association for the Advancement of Artificial Intelligence 2018. Available at https://arxiv.org/pdf/1801.07593.pdf. Accessed 1 May 2019.

Abbreviations

ECHR	European Convention of Human Rights
ECJ	European Court of Justice
ECtHR	European Court of Human Rights
EUCFR	Charter of Fundamental Rights of the European Union
GDPR	General Data Protection Regulation
TEU	Treaty on European Union
TFEU	Treaty on Functioning of the European Union
WP29	Article 29 Data Protection Working Party

1. Introduction

1.1. Research Topic and Aims of the Research

A current trend in the society is to utilise automation in increasing amounts, including in the context of decisions affecting individuals. Examples of such situations are admission to the university, recruitment and credit decisions. In order to increase efficiency, both the private and public sector are creating tools to eliminate humans from the decision-making process, aiming at replacing them with machines. Often these automated decision-making models use artificial intelligence. While the development may reduce the amount of manual work and enable humans to focus on more complex tasks, there is a risk that automated decision-making may not be objective. The concerns relate especially to cases where an allegedly discriminatory decision was made in an automated process and the reasoning behind the decision by a machine cannot be presented in an understandable manner. One aspect of the difficulty to both build and explain such automated models is that of data protection, which may restrict the creation and deployment of the technology.

This research focuses on the conflict between non-discrimination and privacy in the context of automated decision-making based on machine learning.[1] As a principle, it seems that building non-discriminating machine learning algorithms for automated decision-making requires as much data as possible, even the sensitive personal data.[2] This is due to the fact that algorithmic bias usually occurs due to some groups not being represented well in the training data[3]. Consequently, the machine learning model faces difficulties making decisions on the minorities due to lack of historical data, and it may even generalise the few cases that have been used as an

[1] About the issues of discrimination with regard to big data and algorithms in general, see O'Neil (2017), and Pasquale (2015). O'Neil writes, inter alia, about the inequality in the courts, recruitment, workplace, as well as credit and insurance applications, all caused by the use of algorithms in the decision-making. Pasquale presents technologies that profile people, so-called 'filter bubble' technologies, and finance technologies that run algorithmic trade. Edited books on the subject include, for example, Transparent Data Mining for Big and Small Data edited by Cerquitelli, Quercia and Pasquale (2017) and Discrimination and Privacy in the Information Society – Data Mining and Profiling in Large Databases edited by Custers, Calders and Zarsky (2013).

[2] Žliobaitė – Custers (2016), especially pp. 8–17, and Žliobaitė (2017), p. 1068. See also Koskinen (2018), p. 243, about the removal of data from an algorithm potentially making the algorithm less accurate. Koskinen also recognises the issue of discrimination in this context.

[3] Training data refers to the data used in order to train an algorithm that operates in a machine learning application. See Koulu (2019), p. 22.

input so that it will offer a similar decision for everyone in the group. The principles used in data protection, on the other hand, aim at minimising the amount of data processed. This is due to the fact that usually, the training data is personal data[4], and the principles of purpose limitation and data minimisation cover the processing of personal data, as regulated in Articles 5(b) and 5(c) of the General Data Protection Regulation (GDPR).[5] However, would it be possible to define the purpose as creation of a non-biased automated decision-making system, for which reason the amount of data used could be larger and still in compliance with the principle of data minimisation, because the amount of data is necessary for the limited purpose?

The response to the question of balance may also lie in the principle of accuracy as set forth in Article 5(1)(d) of the GDPR: '[Personal data shall be] accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy')'. As noted above, the issue with discrimination occurs when decisions are generally applied to a certain group regardless of the individual's qualities. If the data concerning each individual is accurate, and the individual data is used as a basis for an automated decision in addition or instead of the data on the group, such discrimination should not occur. If accurate data is available on every individual subject to an automated decision, and the decisions are based on their personal data and not only historical data on other individuals, discrimination should not occur. The prerequisite for this is that personal data is available and that it is processed in compliance with the data protection laws when making automated decisions.

The main research question of this thesis is whether the fundamental rights of non-discrimination and data protection can be protected in a balanced way when using machine learning in automated decision-making. In order to come to conclusions with regard to the research question, some sub-questions need to be studied. Firstly, the scope of protection needs to be understood, i.e. what kind of situations are protected by prohibition of discrimination, and whether

[4] The General Data Protection Regulation defines personal data in the following way: 'any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'. There are studies that state that all data is personal data because any data can be connected to an individual, see generally Ohm (2010), especially pp. 1716–1730, and Leonard (2014), p. 60. Given the wide scope of the term 'personal data', it is likely that the data used to train a machine learning model qualifies as personal data.

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, hereinafter the General Data Protection Regulation or GDPR.

the use of machine learning and automated decision-making is relevant for those situations. Secondly, from the point of view of data protection, the requirements of data protection when building and deploying a model for the automated decision-making situation need to be taken into consideration. Thirdly, the requirements of non-discrimination when deploying the model shall be examined.

The most important exclusion from the scope of this research is that automated decisions made by public authorities are left out of the scope. The focus is therefore on the private companies using automation in decisions affecting individuals. In addition, in the sphere of data protection, there are several closely related topics that have been left out of the scope of this study due to the limited length of the work. These include storing of personal data and anonymisation of personal data. Even though the limitations that legislation such as the GDPR set on the storing of personal data make it more difficult to build machine learning models because the period for which the personal data are stored should be limited to a strict minimum[6], the focus of this study is on the training and use of automated decision-making models instead of the collecting and storing of personal data that is used for the training of the model. The use of anonymised data is not dependent on the restrictions laid down in data protection legislation, and therefore allowed also in the training of machine learning models.[7] If anonymising the training data eliminates the compliance issues, it is not necessary to write about the use of anonymised data. The questions whether personal data can be anonymised at all, at least permanently, and whether using anonymised data would be useful at all in the context of machine learning, are so diverse that in order to keep the writing within the page limits allocated, it is only possible to touch upon these subjects.

When it comes to discrimination, the study focuses on prevention of discrimination in the phases of building and deploying an automated decision-making model. Remedies for a situation in which such alleged discrimination occurs have therefore been left out of the scope. The main rule is that the burden of proof shall be on the respondent, i.e. the respondent shall prove

[6] See more about the restrictions on storing personal data in recital 39 and Article 5(1)(e) of the GDPR.

[7] Recital 21 of the GDPR states: 'The principles of data protection should apply to any information concerning an identified or identifiable natural person. -- The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.' Note that the GDPR does apply to pseudonymised data. Recital 26 states: 'Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person.' See also Koskinen (2018), p. 241.

that no discrimination has occurred. In the European Union, the Council Directive 97/80/EC of 15 December 1997 on the burden of proof in cases of discrimination based on sex, laid down this principle, and it was renewed in the Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation, Section 2.

1.2. Methods and Structure

The methodology used in this study is doctrinal legal research when it comes to discrimination and data protection legislation. I will identify and analyse the content of laws governing discrimination and data protection, as well as some court cases related to those subjects. The main objective of this study is to analyse the existing legislation in the context of automated decisions based on the use of machine learning.

In addition, the secondary objective of this work is to conclude some recommendations *de lege ferenda* when it comes to the algorithmic solutions used in order to create unbiased decision-making tools. With this, I present some fields to study even further especially from the point of view of legal research. I will also touch upon the question on whether new legislation is necessary in the first place. The president of the European Commission, Ursula von der Leyen, has stated that the Commission aims at proposing legislation for a coordinated European approach on the human and ethical implications of artificial intelligence within the first 100 days in office.[8] It is reasonable to ask whether such legislation is needed, given that we already have non-discrimination laws and data protection laws that all applications of artificial intelligence need to comply with at present. Ideally, compliance with these laws alone should guarantee that human rights as well as ethical aspects have been taken into consideration.

Further research especially from the legal point of view is undoubtedly needed. It is visible in the list of sources of this study, more precisely in the fact that there are few sources concerning discrimination in automated decision-making purely from the point of view of anti-discrimination laws. Most legal sources regarding automated decision-making take the point of view of data protection instead, especially the relevant articles in the General Data Protection Regulation. Some of the sources used are research papers by technical experts who recognize the issue of discrimination and present technical solutions to the problem, without analysing whether

[8] Politico (2019).

those solutions would be compliant with the existing legislation. There are also papers written in close cooperation between legal professionals and technical professionals, mostly contemplating the issue from a general perspective rather than touching upon concrete technical ways to eliminate the bias in machine learning and automated decision-making models. All in all, legal analysis on the suggested technical solutions was difficult to find.

Regarding the general scope of legislation examined in this study, both discrimination and data protection are mainly approached from the European perspective. In addition to the relevant articles in the European Convention on Human Rights (ECHR) and the Charter of Fundamental Rights of the European Union (EUCFR), provisions related to equality and non-discrimination can be found in the Treaty on European Union (TEU) and the Treaty on Functioning of the European Union (TFEU). Certain EU Directives are also relevant for the study. Regarding data protection, the legal framework used in the research consists mainly of the General Data Protection Regulation. Despite the focus on the European legislation, a few sources from the United States have been used. The issue of bias in data that leads to discrimination when automating decisions seems to be even more visible overseas, thanks to Cathy O'Neil's bestseller Weapons of Math Destruction and the so-called 'COMPAS' case that gained publicity in 2016, as two examples.[9] Apart from the aforementioned case from the Supreme Court of Wisconsin, the case law utilised in this study is from the European courts, namely the European Court of Human Rights (ECtHR) and the European Court of Justice (ECJ). One case from a national tribunal in Finland is also examined as the forerunner on decisions regarding discrimination in automated decision-making in Europe. Several official sources by both European and Finland's national authorities have been studied as well.

This work is divided into six chapters. Throughout the study, imaginary examples of situations of automated decision-making will be presented, in order to bring concretion to the theoretical concepts of each chapter. Chapter two is a walkthrough of the technologies used in the automated decision-making applications. Chapter three takes a deep dive into the legislation on anti-discrimination in place in Europe. This chapter also includes case analyses, including the first known case regarding discrimination in automated decision-making from Finland's National Non-Discrimination and Equality Tribunal. Some imaginary examples are also handled from the point of view of non-discrimination legislation. Chapter four sheds light on the data protection laws, especially four selected areas, namely automatic decision-making with regard to data

[9] *State v. Loomis*. The case was widely presented in the media, see for example The New York Times (2016).

protection, purpose limitation principle, data minimisation principle, and data accuracy principle. In the fifth chapter, some technical solutions as well as legislative initiatives to the bias in automated decision-making are presented and shortly analysed from the legal point of view. Lastly, the sixth chapter concludes the study.

2. Technology Behind Automated Decision-Making

2.1. Big Data and Data Mining

The value of data has been rising remarkably in the recent years. There are numerous examples on how companies have managed to customise their products and services thanks to collecting their customers' data, for instance by following the users' actions on the company's website or by asking the customers to register as a customer before granting access to the products and collecting their data in the registration form. Perhaps one of the less harmful examples of such customised services is a streaming service that recommends movies and music to their customers based on their previous consumption.[10] A more controversial example would be targeted advertisement on medicine or health services in the social media of an individual who has used a search engine to find information about a certain disease.

Big data is the buzz word of data science and commercialisation of data. It usually refers to 'the three Vs', i.e. volume, velocity and variety of data.[11] These are the important elements of one of the popular definitions for big data: 'high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation'.[12] Thanks to modern technology, it is possible to collect and store gigantic datasets in real time, and from several different sources.[13] It would therefore be challenging to use the big data in the more traditional way[14], e.g. looking for a confirmation in the data for an assumption made prior to collecting the data. This can also be called verification-driven data mining, or data analytics.[15] Instead, big data makes it possible to find correlations and conclusions in the data without any prior

[10] Another such example, targeted advertising, is mentioned at Information Commissioner's Office (2017), p. 20. The report also specifies that targeted advertising may be harmful if it is discriminating, such as arrest records in the search results for 'black-identifying' names. About the Information Commissioner's Office's publication as a source in general, it is to be noted that is not intended as a guidance document or a code of practice, and it is not a complete guide to the relevant law.

[11] Information Commissioner's Office (2017), p. 6. Zarsky (2016) adds a fourth V for veracity, however mentioning that the veracity of big data is arguable, pp. 998–999.

[12] Gartner IT glossary, 'big data', http://www.gartner.com/it-glossary/big-data.

[13] Information Commissioner's Office (2017), p. 6 and 41. See also WP29 Guidelines (2017a), p. 11, Colonna (2014), p. 306, and Forgó – Hänold – Schütze (2017), p. 18–19.

[14] Information Commissioner's Office (2017), p. 6.

[15] Colonna (2014), pp. 307–308.

expectations.[16] This practice is known as discovery-driven data mining.[17] For example, a law firm could use data analytics in order to find anomalies in contracts related to a merger or an acquisition in the due diligence process. At least from data protection legislation's point of view, as well as perhaps with regard to non-discrimination, the problems related to big data are due to the use of algorithms, the opacity of the processing, the massive amounts of data collected, the repurposing of data, and the use of new types of data.[18]

What is interesting about big data is how the society can become more efficient by analysing it.[19] For example, it can be understood how people and vehicles move in public spaces in order to make them more secure and avoid traffic jams. However, as mentioned, traditional data analysis techniques are usually not efficient enough to make sense of big data. That is why data mining, artificial intelligence and machine learning are needed.[20] A congressional research service (CRS) report defines data mining as the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.[21] The fact that the data is analysed is central from legal point of view, since it means that if the data is personal data, analysing the data is processing of personal data in accordance with data protection laws.[22] From anti-discrimination laws' viewpoint, the patterns and relationships in data sets may be relevant if it is found that there are discriminatory factors that have led to a certain pattern. This will be explained in more detail in chapter 3.

[16] Information Commissioner's Office (2017), p. 10. See also Forgó – Hänold – Schütze (2017), p. 17.

[17] More on discovery-driven data mining, e.g. its subdivision to descriptive data mining and predictive data mining, Colonna (2014), p. 308.

[18] Information Commissioner's Office (2017), p. 9.

[19] Zarsky (2016), p. 1000.

[20] Information Commissioner's Office (2017), pp. 7–8.

[21] Information Commissioner's Office (2017), p. 6, and Zarsky (2013), p. 304. Most, if not all, definitions of data mining raise the importance of discovery of previously unknown, valid patterns and relationships, see also Comandé (2017), p. 183. However, there is no clear consensus on the definition of big data, see Colonna (2014), p. 307.

[22] Article 4(2) of the GDPR defines processing of data in the following way: any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. Already the collection of data for purposes of data mining is therefore considered processing of data, and the GDPR shall be applied to such processing as long as the data is personal data.

2.2. Machine Learning

Due to the increased interest in data, the popularity of data science as a branch of science has expanded. Within data science, there are elements of computer science and statistics. Artificial intelligence[23] is considered to be part of computer science, and machine learning[24] a field of artificial intelligence.[25] Artificial intelligence usually refers to computer systems that are able to perform tasks normally requiring human intelligence, such as logical reasoning and decision-making.[26] Applications of artificial intelligence are able to solve problems independently by recognising patterns in the data or by learning from experience. Artificial intelligence refers to a combination of different technologies and methods, such as machine learning and neural networks, and the aim is to enable machines to perform cognitive tasks.[27] Machine learning refers to a method in which artificial intelligence operates without precise instructions from humans, i.e. comes to conclusions independently.[28] The combination of big data, artificial intelligence and machine learning can be addressed as big data analytics, although other forms of big data analytics exist as well.[29]

Before jumping into the details of machine learning, a brief introduction to algorithms will follow. An algorithm is a defined sequence of instructions that can be used to solve a certain problem.[30] In the context of computer science, an algorithm is a sequence of instructions telling a computer what to do.[31] In its simplest form, an algorithm can be an 'if-then' phrase, for example 'if age<18, then voting right=positive'. In fully automated decision-making, an algorithm uses data about an individual, often in addition to data about a group of individuals, in order to

[23] There is no standard definition for artificial intelligence. It usually refers to computing models that autonomously analyse data in order to adapt to a certain task that has been given to the model. See, for example Elements of AI (2018), chapter 1, section I.

[24] Machine learning can be defined as algorithms that are composed of many technologies used in unsupervised and supervised learning, and that operate guided by lessons from existing information. Gartner IT glossary, 'machine learning', https://www.gartner.com/it-glossary/machine-learning/.

[25] Elements of AI (2018), chapter 1, section II.

[26] The English Oxford Living Dictionary, 'artificial intelligence', https://en.oxforddictionaries.com/definition/artificial_intelligence.

[27] Ailisto et al. (2018), p. 6.

[28] Koulu et al. (2019), p. 22.

[29] Information Commissioner's Office (2017), p. 8.

[30] Kemppinen (2011), p. 76. See also Koulu et al. (2019), p. 21, where algorithm is defined as follows: 'a description or an instruction, either mathematical or written in a programming language and divided into stages, on how to execute a task or how a program should react'.

[31] Viljanen (2018), p. 1071.

make a decision about the individual. For example, the algorithm can use inputs 'gender=female', 'city=Helsinki', and 'occupation=unemployed', and make a decision 'loan=denied'. If the inputs were something different, the recommendation would be 'loan=granted'. This type of an algorithm is still a very simple one, and it would probably be easy to explain the reasoning behind the decision. Problems occur when new layers are added to the algorithm, i.e. when using deep learning[32], and it is no longer clear which inputs exactly made the algorithm come to its decision. In addition, it can happen that the data about the individual is incomplete, in which case the algorithm would use data about the group and generalise it to the individual in question. For example, the data about the individual could include their gender and place of residence but not their occupation. The algorithm could then observe that in general, females in the city of residence are in low-income occupations, in which case the decision would be to deny the loan without knowing the actual income situation of the loan applicant. A good definition for algorithms in the context of machine learning is 'predictive models (decision rules) captured from historical data using data mining'.[33]

There is a difference between a simple programmed model and a model that uses machine learning. A traditional computer program will follow certain rules to solve well-defined problems.[34] This is a linear analysis executed based on how the program was originally programmed.[35] In machine learning, the algorithms are programmed to learn to solve problems independently.[36] The algorithms are able to adapt their outputs according to new data fed to the program as an input.[37] Because of this autonomous nature of the machine learning tools, it may be challenging to show the causal relations between the input and the output, for example why certain loan application was denied by an automated decision-making model.[38] Within machine learning models, there is a difference between models that use supervised learning and models that use unsupervised learning.

Supervised learning is a form of machine learning where the algorithm is given an input and the output in the training set, for example images of traffic signs that are all labelled ('stop

[32] Deep learning involves feeding vast quantities of data through non-linear neural networks that classify the data based on the outputs from each successive layer. Forbes (2016).

[33] Žliobaitė – Custers (2016), p. 183.

[34] Tutt (2016), p. 85. See also Koulu et al. (2019), p. 13.

[35] Information Commissioner's Office (2017), p. 7.

[36] Tutt (2016), p. 85.

[37] Information Commissioner's Office (2017), p. 7.

[38] Koulu et al. (2019), p. 13.

sign', 'speed limit', etc.). The task for the algorithm after the training phase would be to predict the correct output, or label, for example which traffic sign is in the picture.[39] In the simplest cases, the answers are in the form of yes/no (a binary classification problem).[40]

When using unsupervised learning, there are no labels or correct outputs. The task is to discover the structure of the data: for example, grouping similar items to form clusters, or reducing the data to a small number of important dimensions.[41] Using the example of traffic signs, in the context of unsupervised learning, the images would not have been labelled before being submitted to the algorithm. Instead, the task of the algorithm would be to find similarities between the training data and categorise the data, for example by grouping the images into two different clusters, the red and white signs (stop signs) and the yellow and red signs (speed limit signs). The algorithm would therefore not know the labels of the two groups, it would only know that the images in each group belong together.

Unsupervised learning can be used for profiling, which may also be considered automated decision-making. Article 4(4) of the GDPR defines profiling as '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'. Unsupervised machine learning, clustering, can be used to analyse the data collected about individuals in order to evaluate their characteristics or behaviour patterns, and as a result to place them into a certain category. The aim of profiling is to make predictions about the individuals, for example about their interests or likely behaviour, and these predictions can be further used in targeted advertising, for example.[42]

Typically, the algorithm does not store the data that it has used as an input. Instead, the algorithm creates a decisional rule based on the training data and stores these rules instead of the original training data.[43] It is challenging to understand why a machine learning model has come

[39] Information Commissioner's Office (2017), p. 7.

[40] Elements of AI (2018), chapter 4, section I. See also Karanasiou – Pinotsis (2017), p. 174.

[41] Elements of AI (2018), chapter 4, section I, Karanasiou – Pinotsis (2017), p. 174, and Information Commissioner's Office (2017), p. 8.

[42] WP29 Guidelines (2017a), pp. 6–8.

[43] Koskinen (2018), p. 243. However, Article 29 Data Protection Working Party states that the right to rectification and erasure of personal data apply to both the 'input personal data' (the personal data used to create the profile) and the 'output data' (the profile itself or 'score' assigned to the person), which suggests that WP29 would consider

to a certain decision because even if the source code of the machine learning model is available, only the machine learning method used can be derived from the source code, and not the data-driven decision rule.[44] In addition, if more complex machine learning methods such as deep learning are used, the model used in automated decision-making may become a 'black box', meaning that the decisive factors in the decision-making process of the model are unclear even to those who built the model.[45] This is an issue because usually, it must be possible to provide the reasoning behind a decision to the individuals that are subject to the decision, for example why a loan application was declined. It is of utmost importance to be able to explain the logic behind an algorithm in a case of alleged discrimination, and it may be difficult to understand whether such discrimination detected in a machine learning model is systematic or not.[46]

2.3. Automated Decision-Making

Automated decision-making, also known as algorithmic decision-making, refers to the ability of algorithms to provide solutions in tasks that have been defined to the algorithm, determining the optimal among a set of possible answers without human interaction.[47] A concrete example is admission to university.[48] The test data, applications by different people, needs to be analysed by the automated decision-making model in order for it to be able to decide which candidates shall be admitted to the university. In practice, this happens by submitting the historical applications to the university, as well as the admittance decision related to each application (labelled training data) to the model as an input. The model then processes the applications and the historical decisions, 'learning' what the prerequisites for an application to be successful are. The aim is to make the model able to deduce whether an applicant should be admitted or not even when the input is a new application that was not part of the training data set. Therefore, a large amount of training data is needed to build a model that is capable of independently solving a certain issue.[49] This is considered supervised learning, as described above.

---

training data personal data that would be stored in the automated decision-making model. WP29 Guidelines (2017a), p. 18.

[44] Kroll et al. (2017), p. 638.

[45] Information Commissioner's Office (2017), pp. 10–11. Koskinen (2018), p. 240. Especially if it is unintentional that a decision-making model ends up being discriminatory, it is challenging to notice the issue and to understand based on which decisional rules the discrimination occurs. See Barocas – Selbst (2016), p. 674.

[46] Mittelstadt et al. (2016), p. 2.

[47] Karanasiou – Pinotsis (2017), p. 171 and 173. Koulu et al. (2019), p. 22. About the relationship between profiling and automated decision-making, see WP29 Guidelines (2017a), p. 8.

[48] See a real-life example of such a situation in Staab – Stalla-Bourdillon – Carmichael (2016), pp. 4–5.

[49] Žliobaitė – Custers (2016), especially pp. 16–17. Koulu et al. (2019), p. 13.

As mentioned above, also profiling can be automated decision-making. Article 29 Data Protection Working Party (WP29)[50] divides profiling into three categories in this context: (i) general profiling (without decision-making); (ii) decision-making based on profiling; and (iii) solely automated decision-making, including profiling, which produces legal effects or similarly significantly affects the data subject.[51] The importance between decision-making with human intervention and solely automated decision-making is material in terms of application of Article 22(1) of the GDPR, addressed in more detail below in chapter 4.2.

[50] Article 29 Data Protection Working Party 29 was set up under Article 29 of the Data Protection Directive (Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data) that preceded the GDPR. The Working Party was an independent European advisory body on data protection and privacy, and its tasks were described in Article 30 of Directive 95/46/EC and Article 15 of Directive 2002/58/EC. See WP29 Guidelines (2017a), p. 1. Note that WP29 Guidelines are based on the Data Protection Directive and therefore cannot necessarily be applied to the General Data Protection Regulation as such. Article 68 of the GDPR established the European Data Protection Board, and its task is to ensure the consistent application of the GDPR. The European Data Protection Board has already published certain guidelines, in addition to which it has endorsed several guidelines by WP29, including the Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. See European Data Protection Board (2018).

[51] WP29 Guidelines (2017a), pp. 8–9. General profiling could be, for example, a bank analysing their clientele by categorising them into different age groups. Decision-making based on profiling could be the bank's employee deciding whether to agree the loan based on a profile of the bank's customer produced by purely automated means. Profiling as solely automated decision-making, on the contrary, would be an algorithm deciding whether the loan is agreed and the automatically delivering the decision to the customer, without any prior and meaningful assessment by a human.

3. Discrimination

3.1. Discrimination and Automated Decision-Making

One of the pioneering publications about discrimination with relation to data mining was issued in 2008 by *Pedreschi, Ruggieri,* and *Turini*. Their conference paper, 'Discrimination-aware data mining', was presented at the International Conference on Knowledge Discovery and Data Mining by the Association for Computing Machinery. The paper raises the issue of use of historical data in data mining and machine learning in a rule-based setting, given that the task of machine learning is to classify people into different groups or to predict the group for an individual whose data was not included in the training data set. In order to classify people, the model needs to understand the rules based on which the individuals in the training data were classified. If such classification rules are used to make decisions that affect the individuals, for example their access to benefits, public services, or credit, it may lead to discrimination since the machine learning model will discover traditional prejudices in the training data and use those prejudices as a basis for classifying individuals. In order to tackle this issue, the authors introduce the notion of discriminatory classification rules as a criterion to identify the potential risks of discrimination.[52]

The concerns related to discrimination in data mining and machine learning have only been actively studied for a decade. An important milestone for the field of research was the first annual Fairness, Accountability, and Transparency in Machine Learning (FAT / ML) conference in 2014.[53] By 2019, dozens of articles and conference papers are published on an annual basis, and there are multidisciplinary courses offered in the universities focused on the prevention of discrimination in machine learning.[54] Despite these positive developments, a majority of the research in this field is from technical point of view, and even if the legal implications would be touched upon, in-depth legal analysis is missing. There is an urgent need especially for the assessment of existing non-discrimination legislation with relation to the suggested technical solutions. This is an area with very little prior work, whereas the application of data protection legislation to automated decision-making models has been covered to some extent.

[52] Pedreschi – Ruggieri – Turini (2008), p. 1.

[53] FATML Website, see the schedule of the conference at https://www.fatml.org/schedule/2014.

[54] University of Helsinki Website. See, for example, 'Fairness Aware AI' by the University of Helsinki at https://courses.helsinki.fi/en/data20006.

On a high level, discrimination may occur in automated decision-making for two reasons: because of discriminatory decisional rules in the model or because of data-related issues.[55] The problems with data can be further divided into subcategories. When it comes to the decisional rules in the machine learning model, depending on which factors the automated decision-making model takes into account, the logic behind decisions may be against anti-discrimination legislation. The rules may be directly discriminating, such as if a loan decision depends on a person's gender, ethnic background or age, i.e. the rule could dictate that women are to be granted a loan with more favourable terms than men. On the other hand, the algorithmic model could examine factors that seem objective and end up discriminating indirectly, as is the case when a postal code reveals the ethnicity of a person because of a high percentage of people of certain ethnic origin in a certain neighbourhood.[56] Lastly, when the automated decision-making model aims at a qualitative analysis of individuals, the qualities to be looked for are defined by humans and may not be objective.[57] As an example, ranking candidates based on their years of experience in an automated process may end up discriminating minorities, as will be discovered in chapter 3.4.2.

The data-related issues that can lead to discrimination in automated decision-making when machine learning is used can be divided into three subcategories.[58] Firstly, it is possible that there are prejudices in historical data, as presented above. Secondly, the data may be inaccurate. Thirdly, the data used as a basis for the decision-making may not be relevant to the person subject to the decision, i.e. the decision is made by generalising, not by assessing an individual.[59] In the first scenario, the data can reflect actual historical decisions and be correct in that

[55] Koskinen divides issues related to discrimination in automated decision-making into two categories, discrimination resulting from data and discrimination resulting from criteria given to the algorithm, such as decisional rules. Koskinen (2018), pp. 245–246.

[56] Kroll et al. (2017), p. 681.

[57] Koskinen (2018), p. 246.

[58] This categorisation is partly based on Žliobaitė's presentation Fairness-aware Machine Intelligence in May 2019. Slides shown at that presentation can be found on Žliobaitė's home page https://www.zliobaite.com/. See also WP29 Guidelines (2017a), p. 17, where the Working Party 29 raises the following possible deficiencies in data: 'The input data may be inaccurate or irrelevant, or taken out of context. There may be something wrong with the algorithm used to identify correlations.' WP29 also highlights the risks of inaccurate data related to profiling. When profiling individuals, the personal data on them is often used to make predictions, and these predictions may fail, especially if the personal data based on which the predictions were made was inaccurate.

[59] See Comandé (2017), pp. 181–182 on 'regulating anonymities, not only identities', by which he refers to the fact that in today's world, it does not matter whether information is personally identifiable because the information is often collected in order to generalise behavioural patterns and e.g. target advertising based on how individuals are classified in comparison to other individuals. The same applies to data collected in order to use it as a basis for automated decision-making: individuals do not matter, but patterns that can be generalised to individuals as soon as they are categorised into certain groups.

sense, but what makes the data 'incorrect' in the legal sense is that there are bias in the data, i.e. the historical decisions were discriminatory. In addition, the world and the society are changing, so historical data may not be reliable, or it may not be desirable to use it. For example, the number of females in a certain profession could have been much smaller in the past than it is today, so if recruitment data from five years ago is used, it would not give accurate decisions on whom to employ in that profession.[60] Therefore, it needs to be examined whether this kind of discrimination is prohibited by non-discrimination laws in order to come to the conclusion that the historical data cannot be used in automated decision-making as such. Whether the current non-discrimination legislation offers means to rectify this data is the question that has not been properly studied so far.

The second scenario, inaccurate data, may be caused by many different reasons, examples of which are listed as follows. Firstly, the data can be incomplete. This means that some groups are underrepresented in the data.[61] A concrete example would be airbag tests by car manufacturers. If airbags are only tested with male dummies, no data will be recorded on what would happen if a female was the victim of a car crash.[62] The machine learning model would learn from the test data how to create the safest possible airbag for males, but it could be that females would not be protected as well by the resulting product. Secondly, the sampling process may make the data distorted. This happens if the sample data is not collected at random, for example if the police are more likely to stop young male drivers rather than old female drivers for breathalysing.[63] Thirdly, the data could be missing something important or a wrong conclusion may have been drawn in the first phase of processing.[64] For example, a person had a false credit record default that was later corrected, and this affected a decision made about them. An

[60] This phenomenon got into the public eye when Amazon noticed that the machine learning model that they had been using in recruitment was discriminating against women due to the fact that the industry had been dominated by males for years. See Reuters (2018).

[61] As Koskinen points out, often the minorities in the society are also the underrepresented groups in the data. For example, people with low income may not be able to afford devices or have access to a network that collects data about individuals, hence the data available represents only people above a certain income rate. Since these groups are often discriminated against in the society to begin with, it is especially sensitive if the discrimination reproduces and becomes systematic in algorithmic decision-making. Koskinen (2018), p. 245.

[62] Washington Post (2012).

[63] A real-life example of such a sampling process is case 337/2018 decided in Finland's National Non-Discrimination and Equality Tribunal. Stop and search by the police was found to be ethnic profiling and direct discrimination on the grounds of racial or ethnic origin, prohibited in the Finnish Non-Discrimination Act. For an English summary of the case, see European network of legal experts in gender equality and non-discrimination (2019). See also Barocas – Selbst (2016), p. 687, on how disproportionate surveillance of protected groups in the workplace may lead to inaccuracy in the data.

[64] Comandé, p. 192, refers to this issue as 'arbitrariness-by-algorithm', which relates to the fact that algorithms may misinterpret even some accurate data regarding individuals.

individual could also be falsely profiled, for example to be a part of a sexual minority, and this could affect a decision made about them. The issues related to the second scenario of inaccurate data will be studied further in chapter 4 since they are more closely related to data protection than discrimination. In fact, accuracy in data processing is one of the principles set forth in the General Data Protection Regulation.

In the third scenario, it may be that no data relevant to the decision is collected from the individual, but instead, the person is profiled into a certain category and the basis for the decision are observations made on other individuals belonging to that category, often by looking for generalisations and average behaviour of that group. A concrete example would be a loan decision based on a person's gender, ethnic background or age instead of their property or income level. This could lead to direct discrimination based on a protected ground, as explained in more detail in chapter 3.4.3.

3.2. Scope of Non-Discrimination

This study focuses on discrimination in the relationship between individuals and private companies. Discrimination by public authorities towards individuals has been left out of the scope of the research, even though the public sector has also started deploying automated decision-making tools.[65] The reason for this exclusion is that in the vertical relationship between the state and the citizens, the state authorities have a special liability while in office that goes beyond that of the private companies making decisions affecting individuals.[66] This potentially makes it more difficult for state officials to adopt automated decision-making tools, given that they are liable for the decisions made by the tools in their official capacity. The examples used in this research are solely based on automated decisions made by private entities, which makes it worthwhile to examine the scope of non-discrimination laws, especially whether they can be applied in horizontal relationships between private actors.

[65] In Finland, the tax authority's use of automated decision-making has already led to two complaints to the Parliamentary Ombudsman of Finland. The complaints have been decided by the Ombudsman, see Parliamentary Ombudsman of Finland 3116/2017 and Parliamentary Ombudsman of Finland 3393/2017. The Ombudsman found issues with automation in these cases and requested further clarifications from the tax authority on how the rule of law, due process, liability while in office, as well as the duty of the authorities to advise and serve the citizens are secured in the automated process, see Parliamentary Ombudsman of Finland 3379/2018.

[66] See more about the problematic with automated decision-making in the public sector and application of administrative law in Pöysti (2018).

Primarily, the prohibition of discrimination based on certain grounds as stipulated in the European Convention on Human Rights and the Charter of Fundamental Rights of the European Union protects individuals from discrimination by the public authorities in a vertical relationship.[67] However, nowadays the scope of the protection is wider and covers also horizontal relationships in the private sector. The state therefore has both a negative obligation not to treat anyone in a discriminative way, and a positive obligation to further equality in the society, even in horizontal relationships.[68] In addition, legislation on the lower levels of normative hierarchy often applies equally to both public and private entities. Regulations have a direct horizontal effect in the European Union. Furthermore, many national laws implementing EU directives are applicable to private entities as well, even though the directives as such do not have a direct (horizontal) effect in the European Union law.[69] For example, national laws implementing the Gender Goods and Services Directive (2004/113/EC) guarantee that goods and services are offered to women and men equally, even by private companies.[70] European Union law therefore sets a positive obligation to Member States to prevent discrimination.[71] It should be noted here that the legislation on the lower level does not necessarily protect the rights of all protected groups, at least not in the same scope, as examined in more detail in chapter 3.3.2. That is why it is important that individuals have the right of appeal directly based on the Charter and the Convention even in private relationships, as these legal instruments provide the widest level of protection regardless of the protected group and the situation in which discrimination occurs.

---

[67] With regard to the European Convention on Human Rights, see White – Ovey (2010), p. 100: 'Though positive obligations were once thought to be the exception rather than the rule, there are now hardly any provisions of the Convention under which positive obligations have not been recognized.' For the European Union Charter of Fundamental Rights, see European Data Protection Supervisor (2016), p. 5.

[68] For ECHR, see White – Ovey (2010), p. 86. For EUCFR, see Frantziou (2018): 'The constitutional norm, now affirmed in Dansk Industri, Egenberger, IR and, most recently, Bauer, appears to be that the Charter of Fundamental Rights is horizontally applicable, at least indirectly and, in many cases, directly as well.' Frantziou has also published a book on the horizontal effect of fundamental rights in the European Union in 2019. An example directly from the EUCFR is its Article 23 that sets an obligation to guarantee equality between men and women in all areas, including employment by private employers. For horizontal effect in the context of equality in general, see Ojanen – Scheinin (2011), part III, chapter 2, section 'Yhdenvertaisuusnormien horisontaalivaikutus' (Horizontal Effect of Equality Norms), and Gellert et al. (2013), pp. 63–64.

[69] See, for example, Reinisch (2012), p. 64. The non-horizontality of directives has also been recognised by the ECJ, for example in Joined Cases C-569/16 and C-570/16 *Stadt Wuppertal v Maria Elisabeth Bauer* (C-569/16) and *Volker Willmeroth, in his capacity as owner of TWI Technische Wartung und Instandsetzung Volker Willmeroth e.K. v Martina Broßonn* (C-570/16) [2018], para 76.

[70] For example, in Finland the Gender Goods and Services Directive was implemented by amendments to the Act on Equality between Women and Men (609/1986). Section 8e was added to the Act, and the section specifically states: 'The action of a provider of goods or services shall be deemed to constitute discrimination prohibited under this Act if a person is treated less favourably than others on the basis of gender in the provision of goods and services available to the public *in the public or private sector*, or if the person is otherwise treated in the manner referred to in section 7 (emphasis added).'

[71] Ojanen – Scheinin (2011), part III, chapter 2, section 'Yhdenvertaisuusnormien horisontaalivaikutus' (Horizontal Effect of Equality Norms).

With regard to the European Convention on Human Rights, the positive obligation derives from Article 1: 'The High Contracting Parties shall secure to everyone within their jurisdiction the rights and freedoms defined in Section I of this Convention.' There is also case law from the European Court of Human Rights stating the positive obligation. The ECtHR has stated about the states' positive obligation that '[t]he obligation on High Contracting Parties under Article 1 of the Convention to secure to everyone within their jurisdiction the rights and freedoms defined in the Convention, taken in conjunction with Article 3, requires States to take measures designed to ensure that individuals within their jurisdiction are not subjected to torture or inhuman or degrading treatment, including such ill-treatment administered by private individuals'.[72] Article 3, prohibition of torture, is not the only right that the state has a positive obligation to protect.[73] In fact, the positive obligation has been recognised under nearly all provisions of the ECHR. It should be noted, however, that individuals may only bring applications to the ECtHR claiming to be the victim of a violation by one of the High Contracting Parties of the rights set forth in the Convention or the Protocols thereto, as stipulated in Article 34 of the ECHR. Consequently, there are limits to the responsibility of the states for violations committed by private persons, and the question often is whether the state has provided for an effective remedy against such violations.[74]

With regard to non-discrimination, Article 14 of the ECHR and Article 1 of Protocol No. 12, the applicability to horizontal relationships is not so clear. One of the reasons is that, as examined below, protection from discrimination applies only in relation to the exercise of another right guaranteed by the Convention.[75] The explanatory report to the Protocol No. 12 specifies the scope of protection against discrimination to cover cases where a person is discriminated against:

[72] *Z and others v. the United Kingdom* (2001), para. 73.

[73] See also, for example, *X and Y v. the Netherlands* (1985), para. 23: 'The Court recalls that although the object of Article 8 (art. 8) is essentially that of protecting the individual against arbitrary interference by the public authorities, it does not merely compel the State to abstain from such interference: in addition to this primarily negative undertaking, there may be positive obligations inherent in an effective respect for private or family life (see the Airey judgment of 9 October 1979, Series A no. 32, p. 17, para. 32). These obligations may involve the adoption of measures designed to secure respect for private life even in the sphere of the relations of individuals between themselves.'

[74] White – Ovey (2010), pp. 99–100.

[75] White – Ovey (2010), p. 547.

i. in the enjoyment of any right specifically granted to an individual under national law;
ii. in the enjoyment of a right which may be inferred from a clear obligation of a public authority under national law, that is, where a public authority is under an obligation under national law to behave in a particular manner;
iii. by a public authority in the exercise of discretionary power (for example, granting certain subsidies);
iv. by any other act or omission by a public authority (for example, the behaviour of law enforcement officers when controlling a riot).[76]

It is noteworthy that a majority of the cases above are related to discrimination by the public authorities towards individuals, hence outside of the scope of this study. The explanatory report to the Protocol No. 12 further specifies that also relationships between individuals may be in the scope of the Protocol, if they are in the public sphere normally regulated by law, for which the state has a certain responsibility.[77] For example, an electricity supplier could not refuse to provide electricity to a person's house based on the person's gender, race or other quality. Whether insurance companies or banks and other financial institutions are covered in this public sphere is debatable.

With regard to the Charter of Fundamental Rights of the European Union, the positive obligation of the state to protect individuals from discrimination by private entities has been enforced in the case law of the European Union. With regard to non-discrimination, the European Court of Justice recognised the horizontal direct effect of the general principle of non-discrimination on grounds of age in the *Mangold* case.[78] The question was primarily on the application of the Employment Equality Directive (2000/78/EC), but the transposition time of the Directive had not yet been expired at the time of the events that led to the court proceedings. The ECJ therefore referred to international instruments and constitutional traditions common to the Member States and regarded the principle of non-discrimination on grounds of age as a general principle of Community law.[79] The ruling was criticised because it gave effect to the Directive before the end of its transitional period and in a horizontal relationship despite the principle of prohibition of horizontal effect of the directives. The judgment was also claimed to have created legal uncertainty in the European Union.[80] Despite the criticism, the ECJ followed a similar approach on horizontal direct effect of the general principle of non-discrimination on grounds of age in

[76] Council of Europe (2000). Explanatory Report to the Protocol No. 12 to the Convention for the Protection of Human Rights and Fundamental Freedoms (ETS No. 177), para. 22.

[77] *Id.*, para. 28.

[78] Case C-144/04 *Werner Mangold v Rüdiger Helm* [2005] ECR I-09981, para 74.

[79] *Id.*, para 74–75.

[80] Papadopoulos (2011), p. 442.

a subsequent case of *Kücükdeveci* by referring directly to Article 21(1) of the EUCFR.[81] In 2018, the direct horizontal effect of non-discrimination was further confirmed in two different cases brought before the European Court of Justice, *Egenberger*[82] and *IR v JQ*[83], in both cases on grounds of religion or belief. What is considered to be the most remarkable case in terms of horizontal direct effect and non-discrimination is the *Bauer* case from November 2018. The question that was remarkable from the point of view of horizontal direct effect was whether the right to paid annual leave applied where the employment relationship is between two private persons. In this connection, the ECJ refers to the prohibition of horizontal effect of EU directives[84], which made it necessary to examine the scope of Article 31(2) of the EUCFR. The ECJ thereinafter affirms the horizontality of the Charter:

> '[A]lthough Article 51(1) of the Charter states that the provisions thereof are addressed to the institutions, bodies, offices and agencies of the European Union with due regard for the principle of subsidiarity and to the Member States only when they are implementing EU law, Article 51(1) does not, however, address the question whether those individuals may, where appropriate, be directly required to comply with certain provisions of the Charter and cannot, accordingly, be interpreted as meaning that it would systematically preclude such a possibility.'[85]

The judgment is viewed to have confirmed the horizontal effect of the EU Charter of Fundamental Rights by comparing the same scenario in employment and showing that not extending the effect of the Charter to the private employer in the same way as it affects a public employer would lead to an unequal situation for the employees. Regarding the direct horizontal effect, the Court does not provide clear guidance. It appears that it will depend on the national implementation of fundamental rights whether private individuals should be able to get protection directly based on the EUCFR.[86]

---

[81] Case C-555/07 *Seda Kücükdeveci v Swedex GmbH & Co. KG* [2010] ECR I-365, para 22.

[82] Case C-414/16, *Vera Egenberger v Evangelisches Werk für Diakonie und Entwicklung eV* [2018].

[83] Case C-68/17, *IR v JQ* [2018].

[84] Joined Cases C-569/16 and C-570/16 *Stadt Wuppertal v Maria Elisabeth Bauer* (C-569/16) and *Volker Willmeroth, in his capacity as owner of TWI Technische Wartung und Instandsetzung Volker Willmeroth e.K. v Martina Broßonn* [2018], para 76–77.

[85] *Id.*, para 87.

[86] Frantziou (2018): 'The final positive point of significance in this case is what I interpret as a tentative clarification of the existing doctrine on direct effect in horizontal disputes. The omission of an explicit reference to direct effect in paragraph 91 of the ruling might be easy to overlook. However, in my view, the judgment appears to make a careful and accurate procedural refinement to the horizontality case law (one that the reporting judge had herself fervently defended during her academic career): the direct effect of EU law, i.e. its invocability in a dispute before national courts, depends on the mandatory nature of the right. In cases against the state, there is parity between that invocability and the remedy offered. Yet, in horizontal disputes, different legal systems have traditionally incorporated fundamental rights in a variety of ways – say, by imposing the obligation on the employer directly or by requiring the state to step in. Bauer suggests that, as long as the right is offered effectively, some

Apart from the vertical versus horizontal scope, there is also the question on the substantial scope of the legislative instruments. The Charter of Fundamental Rights of the European Union only protects the rights governed by it when the Member States are applying European Union law. Therefore, in order to claim an infringement of the prohibition of discrimination on the grounds of Articles 20 or 21 of the EUCFR, the situation in which discrimination occurs must be stipulated in the laws of the European Union.[87] This is remarkable as such, since fundamental rights have not always been in the centre of the EU law, and the Charter only became a binding legal instrument upon the entry into force of the Lisbon Treaty in 2009.[88] Article 6 of the Treaty on European Union gives to the Charter of Fundamental Rights the same legal value as the treaties. In addition, paragraph 3 of Article 6 of the TEU states that the European Convention of Human Rights shall constitute general principles of the Union's law.[89]

When it comes to the European Convention on Human Rights, Article 14, protection from discrimination applies only in relation to the exercise of another right guaranteed by the Convention.[90] In this case, however, the case can be about applying any laws, national or international. Everyone within the jurisdiction of a member state is protected, both citizens and non-citizens. Also, beyond the national territory, in areas under the effective control of the state, claims of infringement of Article 14 can be brought against the state.[91] What makes the protection granted by the ECHR even wider is that Article 1 of Protocol No. 12 covers not only discrimination related to other rights of the Convention, but also in relation to the 'enjoyment of any right set forth by law' and 'by any public authority'. The wording therefore refers primarily to public

space is starting to be carved out for this additional constitutional complexity of horizontality to be accommodated.'

[87] European Union Agency for Fundamental Rights and Council of Europe (2018), p. 29.

[88] Rosas (2015), p. 12. In addition, the binding nature of the EUCFR has been recognised in the case law of the ECJ, Case C-617/10 *Åklagaren v Hans Åkerberg Fransson* [2013], para. 21.

[89] Article 2 of the Treaty on European Union also emphasizes the importance of fundamental rights: 'The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.'

[90] In the aforementioned ECtHR case *X and Y v. the Netherlands* (1985), the applicants contended that the violation of Y's right to respect for private and family life was in relation to discrimination based on the grounds of Y's disability. It can be criticised that in that specific case, the Court did not examine the violation of article 8 in conjunction with Article 14. According to the applicants, Y had not been able to file a complaint herself due to her disability, which was the reason why the complaint was filed by her father X. X's appeal to the decision not to open proceedings against B who had sexually assaulted Y was dismissed because Y had not taken action herself. One could argue that the appeal would have been successful if Y had not been disabled, i.e. Y was discriminated against based on a protected ground.

[91] European Union Agency for Fundamental Rights and Council of Europe (2018), p. 27.

authorities, which means that protection by the Protocol in private relationships would need to be confirmed by the ECtHR. The Protocol does not prevent states from taking positive action, provided that there is an objective and reasonable justification for those measures. This is separately mentioned in the in the preamble to the Protocol that refers to measures in order to promote full and effective equality being allowed on the aforementioned conditions.[92]

3.3. The General Principle of Equal Treatment and Prohibition of Discrimination on Specific Protected Grounds

3.3.1. Introduction to the Principles

Both the principle of equality and the prohibition of discrimination are central human rights, so fundamental that they are placed at the very beginning of the Universal Declaration of Human Rights, Articles 1 and 2.[93] International non-discrimination legislation thus sets forth the general principle of equal treatment on the one hand, and prohibition of discrimination based on specific protected grounds on the other hand. These are stipulated, for example, in the Charter of Fundamental Rights of the European Union, where Article 20 covers the general principle of equal treatment, and Article 21 contains an open list of grounds based on which discrimination is prohibited. It is important to note that the list of grounds is non-exhaustive:

> 'Article 20
> Everyone is equal before the law.
>
> Article 21
> 1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.
> 2. Within the scope of application of the Treaty establishing the European Community and of the Treaty on European Union, and without prejudice to the special provisions of those Treaties, any discrimination on grounds of nationality shall be prohibited.'

[92] Ojanen – Scheinin (2011), part III, chapter 2, section 'Syrjinnän kielto' (Prohibition of discrimination), subsection 'Syrjintäkiellot ihmisoikeussopimuksissa' (Prohibition of discrimination in human rights conventions).

[93] Article 1 sets forth the general principle of equality: 'All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.', whereas Article 2 announces the prohibition of discrimination on specific protected grounds: 'Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty.' UN General Assembly, Paris 1948. Universal declaration of human rights (217 [III] A). Available at http://www.un.org/en/universal-declaration-human-rights/.

The European Convention of Human Rights lacks a provision on equal treatment and contains two different articles on non-discrimination based on protected grounds, Article 14 (prohibition of discrimination), and Protocol No. 12, Article 1 (general prohibition of discrimination). Despite the absence of an explicit equal treatment article in the ECHR, the European Court of Human Rights has referred to the principle of equal treatment in its case law, in the *Belgian Linguistic Case* dated 23 July 1968.[94]

> 'Article 14 Prohibition of discrimination
> The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.
>
> Protocol No. 12, Article 1 General prohibition of discrimination
> 1. The enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.
> 2. No one shall be discriminated against by any public authority on any ground such as those mentioned in paragraph 1.'

Equal treatment means that two people in the same situation shall be treated in the same way, and on the contrary, two people in a different situation shall be treated in a different way. Any deviation from this rule needs to happen on objective and reasonable grounds, otherwise the treatment shall be considered discrimination.[95] An example in the context of social benefits would be if a woman and a man apply for parental allowance and they both fulfil the requirements set forth in the law in order to be granted the allowance, they are in the same situation and shall both receive the allowance. The gender of the applicant shall not affect the decision of the authority on whether to grant the allowance or not.

Even so-called *de jure* equality, i.e. equality in the legislation, is not always considered enough. *De facto* equality refers to a situation in which all individuals would be equal in practice, and not only by the book. In some situations, the state may have a positive obligation to act in order to eliminate inequality in the society, even if such inequality was related to private entities' practices.[96] In order to achieve *de facto* equality, positive action might be required. Neither the

---

[94] *Case 'relating to certain aspects of the laws on the use of languages in education in Belgium' v. Belgium* (1967), B. Interpretation adopted by the Court, para 10. See also Ojanen – Scheinin (2011), part III, chapter 2, section 'Johdanto' (Introduction).

[95] Ojanen – Scheinin (2011), part III, chapter 2, section 'Johdanto' (Introduction). See also Romei – Ruggieri (2013), p. 112.

[96] Ojanen – Scheinin (2011), part III, chapter 2, section 'Johdanto' (Introduction).

general principle of equal treatment nor the prohibition of discrimination based on specific grounds forbid positive action.[97]

The principle of prohibition of discrimination based on specific protected grounds can be applied in the context of machine learning. An example of a discriminatory situation is where two individuals have the same characteristic relevant to the decision making, and they differ only in the sensitive attribute (e.g. gender or race), but an automated decision-making model results in different decisions.[98] This kind of a situation is discrimination on specific protected grounds as stipulated in Article 21 of the EUCFR and Article 14 of the ECHR, provided that the sensitive attribute can be included in the scope of the articles. As mentioned above, the lists of protected grounds in the articles in question are non-exhaustive. Therefore, it is likely that a sensitive attribute would be in the scope of the Charter and the Convention.

What is important to note is that in some cases, there are grounds by which it is acceptable to treat individuals differently, for example the freedom of contract. In many cases, private sector is allowed to select their customers when offering goods and services. For example, an insurance company could decide not to grant insurances in a certain geographical area, e.g. a certain neighbourhood in a city. However, if this decision was made based on a sensitive attribute discovered in the area in question, anti-discrimination laws may prohibit such operation.[99] In the European Union, an important piece of legislation in this context is the so-called Gender Goods and Services Directive, which will be studied further in chapter 3.3.3.

3.3.2. The Asymmetrical Scope of Protected Grounds

In addition to the human rights and fundamental rights regulation, there are numerous provisions on lower levels of normative hierarchy on non-discrimination. The Treaty on the Functioning of the European Union refers to prohibition of discrimination several times. The treaty does not separate the specific grounds of discrimination but includes sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation in all articles related to discrimination. The central provisions in the treaty in this regard are Article 10 (combatting discrimination in defining and implementing the Union's policies and activities), Article 18 (prohibition of

[97] This interpretation has been used at least in Finland, see Government proposal 309/1993, p. 44. Positive action will be explained in more detail in chapter 3.4.

[98] Calders – Žliobaitė (2013), p. 45 and Kamiran – Žliobaitė (2013), p. 156.

[99] Custers (2013), p. 10.

discrimination on grounds of nationality), and Article 19 (authorization of the Council, with the consent of the European Parliament, to take appropriate action to combat discrimination). Also the provisions related to the Citizenship of the Union (Articles 20-25) have relevance in terms of equality.

There are also four EU directives on discrimination, namely the Employment Equality Directive (2000/78/EC), the Racial Equality Directive (2000/43/EC), the Gender Goods and Services Directive (2004/113/EC), and the Gender Equality Directive (recast) (2006/54/EC). In these directives, the scope of protection for different protected groups is not symmetrical. The hierarchy of grounds for non-discrimination from the widest level of protection until the narrowest scope is the following: 1. race and ethnicity; 2. sex; and 3. sexual orientation, disability, religion or belief, and age. Prohibition of discrimination in access to employment, welfare systems, and goods and services are all regulated in EU directives for people of different race and ethnicity. However, sexual minorities, disabled people, people of different religions and ages are only protected by non-discrimination legislation in terms of access to employment.[100] This means that in a case in which, for example, a person is denied access to healthcare based on their sexual orientation, the legal grounds for protection of that individual against discrimination trace back to the EU treaties and the Charter of Fundamental Rights because the directives are not enough to show that such treatment is illegal. Nevertheless, it is possible for Member States to provide wider protection than the original directives in the national laws implementing the EU directives.

In 2008, the European Commission issued a proposal for a fifth EU directive related to discrimination, the proposal for Council directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation (11531/08). This proposal is an attempt to create a better position for the least favoured groups in the hierarchy of protection from discrimination on the European Union level. More specifically, the proposal aims to extend the protection against discrimination on the grounds of religion or belief, disability, age or sexual orientation to areas outside employment, namely social protection, including social security and healthcare; education; and access to goods and services, including housing. The proposed directive would secure horizontal equal treatment in the mentioned areas.[101] However, the proposal has not reached unanimity in the Council until 2019,

[100] European Union Agency for Fundamental Rights and Council of Europe (2018), p. 34.

[101] Council of the European Union (2019), p. 1.

hence it has not been approved. The proposal has, however, been discussed in the Council several times during these eleven years. The latest report is from May 2019, in which the doubts raised by some Member States concern the proposal infringing national competence for certain issues and conflicting with the principles of subsidiarity and proportionality. Moreover, some Member States question the inclusion of social protection and education within the scope, and some have requested clarifications and expressed concerns relating to the lack of legal certainty, the division of competences, and the practical, financial and legal impact of the proposal.[102] To conclude, the Presidency reports: 'Despite the broad support for the objectives of the proposed Directive, technical work and further political discussions are needed before the required unanimity can be reached in the Council.'[103]

The inability of the Council to pass such legislation is problematic. Taking advantage of the asymmetry in the directives to decline healthcare or education from people with a certain religion, disability or sexual orientation, as a few examples, is contradictory to the international human rights instruments, including the EUCFR, since they prohibit any discrimination based on those grounds. It is even possible that a case will be brought before the European Court of Justice to find that the existing non-discrimination directives are invalid under Articles 20 and 21 of the Charter of Fundamental Rights of the European Union, as far as discrimination on the grounds of religion or belief, disability, age or sexual orientation is made possible due to the asymmetrical scope of the directives.[104]

For example, in Finland the Employment Equality Directive and the Racial Equality Directive have been implemented by the Non-Discrimination Act (21/2004)[105] that was later superseded by a new act with the same name, Non-Discrimination Act (1325/2014). The reason why a new legal act was drafted was exactly in the asymmetries in the directives. The Finnish legislator's will was to make the legal act uniform so that identical judicial remedies and sanctions would apply regardless of the grounds for discrimination.[106]

---

[102] Council of the European Union (2019), p. 2.

[103] *Id.*, p. 5.

[104] This kind of an application could be inspired by the famous Joined Cases C-293/12 and C-594/12, *Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources and Others and Kärntner Landesregierung and Others* [2014], in which it was found that Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC was invalid under Articles 7 and 8 of the EUCFR.

[105] Government proposal 44/2003, p. 1.

[106] Parliamentary reply 95/2003, section 1.

3.3.3. The Context of Machine Learning and Case Law

It has been said that, compared to data protection legislation, anti-discrimination laws are scattered due to the regulation being split into multiple, specialised directives.[107] It seems that in the society, questions of equality are still not agreed upon, which can be seen in the difficulties faced when drafting and negotiating legislation securing the rights of minorities. It is not very long ago when, for example, only white men had voting rights, or slavery was still accepted based on a person's ethnicity.[108] Therefore, it is not surprising that algorithmic bias has been detected in the models developed for automated decision-making – it is almost natural that the training data contains discriminating factors. The more problems there have been in the society related to inequal treatment of people, the more important it is to compensate when developing models to be used in the future for automated decision-making, even by utilising methods of positive action.

The directives mentioned above can be directly applied to situations of automated decision-making, even in private relationships. For example, the Gender Goods and Services Directive guarantees that everyone has equal access to goods and services regardless of their gender. If, for example, the terms of a loan offered by a bank are different to women and men, and the only reason behind the different treatment is the gender of the subject, this kind of conduct is against the directive. However, the directive is restricted to anti-discrimination between genders, not, for example, nationalities or sexual minorities. Furthermore, 'gender' is not defined in the directive, and as only the female and male genders are mentioned in the text of the directive, it is unclear whether other genders, such as transgender individuals, would be protected.[109]

One of the most controversial articles in the directive is Article 5(2) related to the insurance industry. Previously, it was considered that gender is an objective ground for different insurance

[107] Gellert et al. (2013), p. 63.

[108] These are a couple of examples mentioned by Ojanen and Scheinin on the historical changes in the society that affect the development of anti-discrimination laws. Ojanen – Scheinin (2011), part III, chapter 2, section 'Syrjinnän kielto' (Prohibition of Discrimination).

[109] An interesting study paper on the protection of non-binary genders in the European Union non-discrimination law was published in April 2019, see Espinhaço Gomes (2019). Gender Goods and Services Directive is examined on pages 24, 44–45, and 75.

payments, for instance because statistically, women live longer than men.[110] The directive states in Article 5(1) that 'the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits'. Article 5(2), however, makes an exception to the foregoing by giving the chance to Member States to permit proportionate differences in individuals' premiums 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. The Court of Justice of the European Union decided in the *Test-Achats* case that Article 5(2) is invalid due to the fact that it is not in line with the fundamental right of non-discrimination.[111]

The *Test-Achats* case was brought against the Federal Council of Ministers of Belgium (*Conseil des ministres*) by Test-Achats, an association defending the rights of the consumers in Belgium (Association belge des Consommateurs Test-Achats ASBL), as well as two individuals. The applicants wanted to annul a Belgian legal act that transposed the Gender Goods and Services Directive into Belgian law. That was because the Belgian legislator had included in the law the above-mentioned exception for using gender as a basis for insurance fees when statistical data is available. According to the applicants, the derogation was contrary to the principle of equality between men and women. As said, the CJEU ruled Article 5(2) of the Gender Goods and Services Directive invalid with effect from 21 December 2012. It would be interesting to know what would happen if someone challenged the directive based on the fact that it only protects equality between men and women, not, for example, equality between people of different races or social origin. For example, there might be statistics noting that people from a certain ethnic group are more likely to be exposed to a certain disease, and insurance companies might be tempted to use this statistic as a basis for calculating insurance premiums.

Following the judgment, the European Commission published Guidelines on the application of Council Directive 2004/113/EC to insurance in 2011. The aftermath of the case left the situation unclear for insurance industry for years, and the European Parliament issued a report in 2013[112] in order to put pressure on the Commission to publish their report on the matter, a report that was supposed to be published in 2010 in accordance with Article 16 of the Gender Goods and

[110] Report on transposition and application of Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services (2010/2043(INI)), p. 9.

[111] Case C-236/09 *Association Belge des Consommateurs Test-Achats ASBL and Others v Conseil des ministres* [2011] ECR I-00773, para 36.

[112] European Parliament (2013), p. 4.

Services Directive. The Commission finally published the report on the application of Council Directive 2004/113/EC implementing the principle of equal treatment between men and women in the access to and supply of goods and services in 2015.[113]

Even though the *Test-Achats* case did not concern automated decision-making, the takeaway that even statistical differences between groups does not justify discrimination can be applied in the context of automated decision-making models. This is the case in Europe, but there are differences internationally. In the United States, the Supreme Court of Wisconsin has decided a case related to discrimination on specific protected grounds in 2016. The case, *State v. Loomis*, was related to a convict whose motion for post-conviction relief requesting a new sentencing hearing was denied. Loomis filed the motion because he had been convicted based on an assessment made by an algorithmic tool called COMPAS.[114] The assessment concluded that there was a general likelihood that the applicant would reoffend, based on a comparison to others with a similar history of offending.[115] The question evaluated by the supreme was whether the use of COMPAS in sentencing violated the defendant's right to due process for any of the three reasons presented by Loomis: 1) the proprietary nature of COMPAS prevents defendants from challenging the COMPAS assessment's scientific validity; 2) it violates a defendant's right to an individualised sentence; or 3) COMPAS assessments take gender into account.[116] Therefore, the right to due process was the main question in this case. Since the conflict between the intellectual property protection of companies providing automated decision-making tools and the individual's right to due process and transparency in the decision-making are outside the scope of this study, only the question number three on whether the tool was directly discriminating based on a protected ground, i.e. the gender of the subject, is of interest.

Unfortunately for the purposes of this thesis, even the question of discrimination was only assessed from the point of view of due process because Loomis did not bring an equal protection challenge in the case.[117] Interestingly, the supreme court found that use of gender in the assessment served the interests of the justice system rather than a discriminatory purpose because of the statistical facts that men, on average, have higher recidivism and violent crime rates compared to women. In addition, Loomis had failed to show a causal link between his gender and

[113] European Commission (2015).

[114] COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions.

[115] *State v. Loomis* (2016), para 15.

[116] *Id.*, para 5–6, 34.

[117] *Id.*, para 80.

the sentencing based on the assessment.[118] As examined above in the context of insurance fees, in Europe even the statistical differences between genders do not matter at least when it comes to the offering of goods and services, so it is likely that this argument would not have as much weight. In addition, the supreme court stated that the use of COMPAS for sentencing was acceptable, provided that there were other independent factors supporting the decision, i.e. the COMPAS assessment shall not be the determining factor in the decision.[119] This is relevant also in the European legislation, considering that the GDPR makes a distinction between solely automated decisions and decisions confirmed by a human, which will be explored in more detail in chapter 4.2.1. It is notable that even though Loomis did not mention other grounds for discrimination than gender in his appeal, independent testing done with COMPAS has shown that the assessment tool was discriminatory based on a person's skin colour as well, since black defendants were more likely to receive higher risk ratings.[120]

[118] *State v. Loomis* (2016), para 78–86.

[119] *Id.*, para 8–9, 44, 85.

[120] Harvard Law Review (2017), p. 1534.

3.4. Direct and Indirect Discrimination

3.4.1. Introduction to the Principles

Direct and indirect discrimination are concepts that can reveal that the general principle of equality is breached, or that discrimination occurs based on protected grounds. Direct discrimination occurs when a person is treated in a less favourable way in comparison to another person in a similar situation and this difference is based directly on a forbidden ground. Therefore, people in a similar situation are not treated equally but instead, the person belonging to a protected group is discriminated against. In order to establish whether someone has been directly discriminated, a comparison is made between the allegedly discriminated person and another person without the protected characteristic. There is no remarkable difference in the definition of direct discrimination between the European Union law and the European Convention on Human Rights. However, procedurally speaking, in order to complain to the European Court of Human Rights, the applicant must show that they were directly affected, whereas in the European Court of Justice, even general complaints not directly affecting a victim can be made, such as in the *Feryn* case, Case C-54/07.[121] Therefore, a company using an automated decision-making model that is suspected to be discriminating could be sued based on the EU law even if there is no individual applicant claiming to be a victim of discrimination.

To elaborate further on the definition of direct discrimination, firstly, what counts as 'less favourable way' is usually quite obvious, for instance receiving a lower salary. Secondly, the comparison to another person needs to be made through identifying a 'comparator', i.e. a person in materially similar circumstances, with the main difference between the two persons being the 'protected ground'. However, this is not always needed, for example in a situation in which the discrimination is systematic against a certain protected group, such as in the *Feryn* case mentioned above. Thirdly, there needs to be a causal link between the less favourable treatment and the protected grounds. It should be asked whether the alleged victim would have been treated less favourably if the protected ground was not there, e.g. if they had been of a different sex.[122]

[121] European Union Agency for Fundamental Rights and Council of Europe (2018), pp. 42–44. Case C-54/07 *Centrum voor gelijkheid van kansen en voor racismebestrijding v Firma Feryn NV* [2008] ECR I-05187, para 28. The case was regarding an employer declaring publicly that it will not recruit employees of a certain ethnic or racial origin, which was constituted to be direct discrimination even if there was no specific applicant of the mentioned minority.

[122] European Union Agency for Fundamental Rights and Council of Europe (2018), pp. 44–50.

Indirect discrimination occurs when apparently neutral provisions, criteria or practices have the side effect of discriminating against one of the specific forbidden grounds. The concept of indirect discrimination is built on the idea that discrimination is not only treating people in a similar situation in a different way (direct discrimination), but also treating people in a different situation in a similar way. In the case of indirect discrimination, the European Union law definition somewhat differ from that of the ECHR.[123] The EU directives define indirect discrimination to occur when an apparently neutral provision, criterion or practice puts an individual at a particular disadvantage compared to other persons.[124] The Convention itself does not have a wording regarding indirect discrimination, but the European Court of Human Rights has referred to indirect discrimination, as opposed to direct discrimination, in the following way: 'The Court has so far considered that the right under Article 14 not to be discriminated against in the enjoyment of the rights guaranteed under the Convention is violated when States treat *differently persons in analogous situations* without providing an objective and reasonable justification. However, the Court considers that this is not the only facet of the prohibition of discrimination in Article 14. The right not to be discriminated against in the enjoyment of the rights guaranteed under the Convention is also violated when States without an objective and reasonable justification *fail to treat differently persons whose situations are significantly different*.' (emphasis added).[125] The ECtHR has further stated in its judgments that 'a difference in treatment may take the form of disproportionately prejudicial effects of a general policy or measure which, though couched in neutral terms, discriminates against a group'.[126]

Going into detail in the definition of indirect discrimination, firstly, 'a neutral rule, criterion or practice' refers to a general requirement or process applied to everyone regardless of their qualities.[127] Sometimes, however, it is reasonable to adjust the requirement to the subject. As a concrete example, in the ECtHR case *D.H. and Others v. the Czech Republic*, a test was used to determine which pupils should be placed into special schools for special educational needs.

[123] European Union Agency for Fundamental Rights and Council of Europe (2018), p. 42, 53.

[124] A similar wording can be found in the Racial Equality Directive, Art. 2(2)(b), the Employment Equality Directive, Art. 2(2)(b); the Gender Equality Directive (recast), Art. 2(1)(b); and the Gender Goods and Services Directive, Art. 2(b).

[125] *Thlimmenos v. Greece* (2000), para. 44. However, the ECtHR notes in this connection that States enjoy a margin of appreciation in assessing whether and to what extent differences in otherwise similar situations justify a different treatment in law.

[126] The ECtHR has used this definition in its judgments in cases *Biao v. Denmark* (2016), para. 103, and *D.H. and Others v. the Czech Republic* (2007), para. 184, for example.

[127] European Union Agency for Fundamental Rights and Council of Europe (2018), p. 54.

The test was similar to everyone despite the fact that some of the children, especially of the Roma ethnic group, had not gone to preschool and did not speak the Czech language. The practice was found to be indirect discrimination by the ECtHR.[128] Secondly, 'significantly more negative in its effects on a protected group' refers to the requirement that the neutral rule needs to put a protected group in an unfavourable situation. The similar treatment shall therefore have a different effect on different groups. Thirdly, as is the case with direct discrimination, also indirect discrimination needs to be proven by a comparator, a person in a similar situation who is not part of the protected group.[129]

3.4.2. In the Context of Machine Learning

When it comes to direct discrimination in the context of machine learning, the first two conditions, i.e. treatment in a less favourable way and finding a comparator without the protected quality may be easily found when examining the rationale behind an automated decision. The decision made about a certain individual may be less favourable compared to another individual who was also subject to automated decision-making. However, it can be challenging to find the causal link between the protected ground and the less favourable decision by the automated decision-making model if the logic behind the machine learning model is obscure and not necessarily understandable to humans. This, in turn makes it more difficult for the individuals to show that discrimination has occurred and to file and application in the courts.

It can be even more difficult to prove that discrimination occurs indirectly. It seems to be indisputable that the value of someone's apartment is a reasonable ground to use in the assessment of someone's credit score. Even if a majority of people living in a certain area represent a certain ethnic background, knowing for a fact that the reason why they get less favourable loan offers is due to the fact that they are from a certain ethnic group would require a comparison to others belonging to a different ethnic group living in the same neighbourhood. Otherwise, the reason could be justified, i.e. that their property is less valuable. Also, it appears to be a general construction in many societies that women make less money than men. Given that income rate is an acceptable attribute for deciding whether someone gets a loan or not, it should not be surprising that women get rejected more often. Whether the salary rate is due to discrimination in the workplace or not is a different question. This is also a topic touched upon in literature, the

[128] *D.H. and Others v. the Czech Republic* (2007), especially para. 25, 44, and 195.

[129] European Union Agency for Fundamental Rights and Council of Europe (2018), pp. 56–58.

fact that explainable differences in decisions need to be accepted, and it is only the illegal discrimination that needs to be eliminated.[130]

The problem with saying that it is not an issue to discriminate based on a fact that is statistically true, for example, that women make less money, is that in the automated decision models, the actual attributes of the subjects to the decisions are not necessarily taken into account. Instead, even if a new female loan applicant has a high income, they will get their application rejected more often than males because the model has used historical data when training, and that historical data has taught the model that females more often have lower income and therefore fail to pay back their loan more often.[131] The position of females, just to give an example, in the society is rapidly changing and if we keep using historical data to train decision-making models, we will end up discriminating those previously underrepresented groups. The issue is how to find the attributes that objectively explain different treatment and distinguish those from the illegally discriminating attributes. For example, if women statistically work less hours than men, an observation that women earn less money can be the result of less working hours, which is an objective reason.[132]

An important term with relation to indirect discrimination is redlining. In a data set, there are sensitive attributes, such as ethnicity, and objective attributes, such as postal code. If an objective attribute correlates with a sensitive attribute, it is possible to deduce the sensitive attribute from the objective attribute and to indirectly use the sensitive attribute as a basis for the decision-making even though technically, the sensitive attribute has been deleted from the data set. This is called redlining.[133] In other words, the objective attribute serves as a neutral 'proxy' that places a protected group at a disadvantage, such as a ZIP code of a neighbourhood revealing one's ethnicity due to the high number of people from a certain ethnic group living in the same area.[134] The proxies are often information that is valuable in the training of an algorithm or in the decision-making, hence eliminating the proxy may lead to insufficient information in order to make a decision, whereas including the proxy may lead to indirect discrimination.[135]

---

[130] Kamiran – Žliobaitė (2013), pp. 155–169.

[131] Verwer – Calders (2013), pp. 261–262.

[132] Kamiran – Žliobaitė (2013), pp. 157–158.

[133] Pedreschi – Ruggieri – Turini (2013), p. 92. See also O'Neil (2017), pp. 162–163.

[134] See e.g. Gellert et al. (2013), p. 65, Romei – Ruggieri (2013), pp. 121–122, and Calders – Žliobaitė (2013), p. 49, Kamiran – Žliobaitė (2013), p. 156, Hajian – Domingo-Ferrer (2013), p. 243, Verwer – Calders (2013), p. 262, Bayamlioglu (2018) p. 442.

[135] Kroll et al. (2017), p. 681.

The attributes that appear objective but correlate with sensitive attributes are also called explanatory attributes.[136] Is using explanatory attributes discrimination if they are correct? For instance, if there are more females than males with occupation 'nurse', and the occupation is used as a basis for decision-making with the aim to decide whether a person is likely to have a high income or not, it should not be discriminatory to come to the conclusion that more males than females are likely to have a high income. Therefore, it can be criticised if the target of a model is an equal number of positive decisions between the allegedly discriminated group and the allegedly privileged group. For example, a model could aim at giving the same amount of positive loan decisions to both females and males, regardless of their income rate, by modifying the training data or the model itself so that some of the allegedly discriminated females get a positive decision although based on the data, they should have gotten a negative decision, and vice versa, some of the allegedly privileged males will get a negative decision although the data suggests to give them a positive decision.[137] Having the exact same amount of positive and negative results for each group should not necessarily be the objective, but rather trying harder to identify those cases that actually are discriminating, e.g. the data about a female with high income who historically did not receive a loan would be modified in the training data as if she had gotten the loan. However, the equal amount of positive loan decisions can be argued for in terms of positive action.

Explanatory attributes could also lead to discrimination by association, i.e. a situation in which the person being discriminated against does not themselves belong to a protected group.[138] Discrimination by association has been found to happen, for example, in a case where a mother of a disabled child was discriminated against at her workplace due to her child's disability[139] and in a case in which a natural father was discriminated against on the basis of his fatherhood[140], i.e. he was not eligible for maternity benefits even though adoptive male parents were. In comparison, a person may be denied loan because they live in an area with majority of the inhabitants belonging to a certain ethnic origin and the decision not to grant loan is based on redlining, i.e. it was falsely deduced from their ZIP code that they would also belong to that protected group. On top of discrimination against the people with the protected attribute, e.g. ethnic

[136] Kamiran – Žliobaitė (2013), p. 159, and Verwer – Calders (2013) p. 262.

[137] Verwer – Calders (2013), p. 269.

[138] European Union Agency for Fundamental Rights and Council of Europe (2018), p. 51.

[139] Case C-303/06 *S. Coleman v Attridge Law and Steve Law* [2008] ECR I-05603, para 27 and 63.

[140] *Weller v. Hungary* (2009), para 33–35.

origin, people living in that area without the protected attribute have been discriminated against based on discrimination by association.

In order to demonstrate the issue of discrimination, *Žliobaitė*'s hypothetical example of a dataset and what a machine learning model would learn if it was trained with the data set is presented.[141] The following data is available on the salaries of different employees:

Ethnicity	**Experience**	**Salary**	**Ethnicity**	**Experience**	**Salary**
1	1	600	0	1	1100
1	2	700	0	3	1300
1	3	800	0	5	1500
1	4	900	0	7	1700
1	10	1500	0	10	2000

The aim is to develop a way to calculate the salary for a new employee. On the paper, it could have been decided that a person's salary in a company is decided with a simple formula:

Salary = 1000 (base salary) + 100x (x = experience years)

However, when looking at the data above, it can be noticed that people of a certain ethnic origin have lower salary rates than their colleagues of another ethnic origin. If the historical data on the salaries is used to train a machine learning model with the task of calculating the salary for a new employee, the machine will learn the following formula:

Salary = 1000 + 100x - 500y (y = ethnicity)

The resulting model would be directly discriminating towards people of certain ethnic origin because the historical salary decisions in the company were discriminating. Even if it would be noticed that such a bias exists and the model was edited so that it would not take into consideration the employee's ethnicity, we would end up with a formula in which the base salary would be lower, and many years of experience would be rewarded even higher. That is because in the data, it can be seen that the employees with the highest salaries have proportionally more experience than the employees with the lowest salaries. The machine learning model would not

[141] This example is from Žliobaitė's presentation Fairness-aware Machine Intelligence in May 2019.

know that the difference between the employees with a lower salary and a higher salary would be their ethnicity, instead it would learn that the salary should get higher progressively if the employees have several years of experience. Because minorities often have less experience and no access to higher education, even this model taught without the sensitive attribute of ethnicity is punishing the minorities and favouring the privileged group.[142] This is a concrete, yet hypothetical, example of indirect discrimination, where the apparently neutral criteria is in the end discriminating against a group with a protected charasteristic.

3.4.3. Case Law on Direct Discrimination in Automated Decision-Making

Possibly the first legal case regarding automated decision-making and discrimination was decided by Finland's National Non-Discrimination and Equality Tribunal. The decision 216/2017 is dated 21 March 2018. The circumstances of the case were that A had applied to be granted credit for an online purchase, after which the retailer contacted Svea Ekonomi, a credit company that took care of the credit applications for the retailer, but the request to grant credit for the applicant was rejected. The applicant contacted the credit company in order to find out the reasoning for the rejection. The tribunal describes the response that the applicant received in the following way: 'The decision had been based on a credit rating made by credit surveillance services using statistical methods, which do not take the solvency of individual credit applicants into account and which may differ significantly from the profile of the credit applicant and may seem unfair to the credit applicant.' The applicant did not have payment defaults, and the credit company had not requested nor investigated the applicant's financial situation.[143]

The credit rating was based on e.g. the place of residence, gender, first language, and age of the applicant. The scoring system used the percentage of people with a bad credit history in each group (female, male, Swedish-speaking, Finnish-speaking) and calculated a score for each

[142] In the Supreme Court of Wisconsin case *State v. Loomis* (2016), the supreme court evaluates the accuracy of the COMPAS tool used to predict the likelihood of convicts reoffending. Interestingly, the court mentions that 'risk assessment tools may disproportionately classify minority offenders as higher risk often due to factors that may be outside their control such as familial background and education'. It seems that the court recognised the fact that it may be necessary to even 'overcompensate' this kind of tools when it comes to decisions on individuals from minorities due to the fact that there is inequality in the society, as a form of positive action. Another interesting remark by the court in the case was that '[o]ther state studies indicate that COMPAS is more predictive of recidivism among white offenders than black offenders.' This seems to point at the fact that the capability to make accurate decisions on individuals is tied to the amount of data available on comparable cases, i.e. since the majority of the population of the United States are white, there is more data available on white offenders, thus the prediction tool was more accurate with regard to white population. This is related to the conflict between data minimisation principle and non-discrimination studied in more detail in chapter 4.4. More on the *State v. Loomis* case in chapter 3.3.3.

[143] Finland's National Non-Discrimination and Equality Tribunal 216/2017, pp. 1–3.

applicant depending on which group(s) they belong to. For example, men with the first language Finnish were awarded less points because on average, they had more payment defaults than Swedish-speaking women.[144] The National Non-Discrimination and Equality Tribunal found that direct discrimination based on protected grounds had occurred in the case. The tribunal also refers to the *Test-Achats* case to the extent that the applicant's gender was used as one of the grounds for deciding on the creditworthiness.[145] Regarding the use of statistical information based on protected grounds, the tribunal states the following:

> 'Based on the information it has received, the National Non-Discrimination and Equality Tribunal considers that the scoring assessment used by Svea Ekonomi AB focused on statistical information on and the credit history of other people, based on which assumptions on the creditworthiness of A were made. With prohibited grounds of discrimination related to the person, such as gender, first language, age and place of residence, Svea Ekonomi AB has assumed the creditworthiness of A to be less than it would have been with other characteristics. At the same, Svea Ekonomi AB has ignored the individualised information regarding A's credit behaviour and financial standing even though these factors would have favoured extending credit to A. Disregarding such information about A by using formal and abstract statistical payment default information created from the credit behaviour of others, without performing an individual assessment of A's financial standing, is disproportionate and therefore not acceptable as intended by section 11 of the Non-Discrimination Act.'[146]

What is interesting is that the Non-Discrimination Ombudsman suggests in their petition to the Non-Discrimination and Equality Tribunal that the credit company would only be required to investigate the financial status of an applicant individually in cases in which the automatic scoring system finds the applicant not creditworthy.[147] This could result in the procedure being discriminatory in reverse, i.e. that for a positive decision, the applicant could be treated solely based on the average in their group without individual assessment. In that situation, two individuals in different situations would be treated in the same way, against the principle of equal treatment. There would be legal grounds for treating the individuals in a different way. For example, a woman with low income, a factor that should affect the credit decision, would be granted an equal amount of credit as a woman with high income, since the relevant factors such as income rate are not taken into account, but only grounds such as gender, language and place of residence. Like the above example of progressive salary raise for people with more experience, allowing positive automated decisions and only adding human involvement to negative

[144] Finland's National Non-Discrimination and Equality Tribunal 216/2017, p. 3.

[145] *Id.*, p. 6.

[146] *Id.*, p. 20.

[147] *Id.*, p. 6.

decisions could lead to discriminating protected groups.[148] For example, if first language is used as grounds for deciding the credit applications and the automated model is more likely to give a positive decision to Finnish-speaking applicants, they will be granted credit without human involvement in the decision-making and they may get higher amounts of credit than applicants with first language other than Finnish, even if they are in a better financial situation.[149] Whereas, as the model is more likely to give negative decisions to e.g. English-speaking applicants, most of them would end up having their applications decided by humans instead of the machine. Therefore, Finnish speakers are favoured at the expense of English speakers whose applications are reviewed more carefully. What is more, the non-Finnish speakers got under the magnifying glass because it was noticed that on average, they are in a less favourable financial situation to begin with. Therefore, by default Finnish speakers are granted larger amounts of credit than English speakers, even if they were in the same financial position, and even if the Finnish speaker's financial situation was worse than the English speaker's.

Working Party 29 also comments on using comparison to others as a basis for decision-making, concluding that this could lead both to punishing decent individuals based on the actions of others, and to rewarding individuals that do not merit such treatment: 'Hypothetically, a credit card company might reduce a customer's card limit, based not on that customer's own repayment history, but on non-traditional credit criteria, such as an analysis of other customers living in the same area who shop at the same stores. This could mean that someone is deprived of opportunities based on the actions of others. In a different context using these types of characteristics might have the advantage of extending credit to those without a conventional credit history, who would otherwise have been denied.'[150] The Information Commissioner's Office of the United Kingdom shares this view, referring to evidence from the United States where

[148] The Non-Discrimination Ombudsman specifically notes in their petition that in the system used in this case, ethnic minorities with an official first language other than Finnish or Swedish were put in an unfavourable position in the granting of credit. Swedish speakers received the largest amount of points, Finnish speakers came second, and those speaking any other language as their first language received least points. The Non-Discrimination Ombudsman found that the scoring of the official first language in the extension of credit will result, *de facto*, in the segregation on ethnic lines. Finland's National Non-Discrimination and Equality Tribunal 216/2017, pp. 5–6. On the other hand, this could also lead to expats with a high income not being granted credit just because their native language is not Finnish. This kind of a system therefore does not necessarily even serve the business purpose.

[149] Finland's National Non-Discrimination and Equality Tribunal 216/2017, p. 3: 'The credit company had not investigated the applicant's income or financial situation, and neither was this information required on the credit application.' The Finnish Consumer Protection Act (38/1978) even contains a provision, Chapter 7, Section 14, that sets an obligation to creditors to investigate the consumer's credit rating and financial status. It is specifically stated that the evaluation must be done based on sufficient information on the consumer's income and other financial circumstances. This provision is based on Article 8(1) of the Directive 2008/48/EC of the European Parliament and of the Council of 23 April 2008 on credit agreements for consumers and repealing Council Directive 87/102/EEC, hence similar obligations are in force in all Member States.

[150] WP Guidelines (2017a), p. 22.

people's credit limits were lowered because of payment defaults by other people shopping in the same stores.[151]

3.5. Positive Action

Positive action refers to measures which are necessary to ensure full equality in practice, in situations in which formal equality is not enough to reach factual equality between individuals.[152] Outside of Europe, e.g. in the United States, the concept goes by the name affirmative action.[153] The European law sources refer to positive action, which is the term used in this study as well. In the European Union, Member States are not obliged to permit positive action.[154]

In the European Union law, the Charter of Fundamental Rights of the European Union recognises positive action in the prevention of discrimination between men and women: 'The principle of equality shall not prevent the maintenance or adoption of measures providing for specific advantages in favour of the under-represented sex.' In addition, three of the non-discrimination directives use almost an identical wording with regard to positive action, including protected grounds other than gender in the legislation. Article 5 of the Racial Equality Directive, Article 7(1) of the Employment Equality Directive, and Article 6 of the Gender Goods Directive all state: 'With a view to ensuring full equality in practice, the principle of equal treatment shall not prevent any Member State from maintaining or adopting specific measures to prevent or compensate for disadvantages linked to [a protected ground]'. The protected grounds mentioned are, respectively, racial or ethnic origin; religion or belief, disability, age or sexual orientation as regards employment and occupation; and sex. The Gender Equality Directive (recast) uses a slightly different definition: 'Member States may maintain or adopt measures within the meaning of Article 141(4) of the Treaty with a view to ensuring full equality in practice between men and women in working life.' As the directive was drafted prior to the Lisbon Treaty, Article 141(4) of the Treaty refers to the Treaty establishing the European Community: 'With a view to ensuring full equality in practice between men and women in working life, the principle of equal treatment shall not prevent any Member State from maintaining or adopting measures providing for specific advantages in order to make it easier for the underrepresented

[151] Information Commissioner's Office (2017), pp. 20–21.

[152] European Commission (2018), p. 81.

[153] See, for example, European Commission (2012), pp. 63–68.

[154] European Commission (2018), p. 81.

sex to pursue a vocational activity or to prevent or compensate for disadvantages in professional careers.'

The directives provide only high-level guidance on what sort of measures would be considered positive action and therefore not discrimination, which is why it is better to examine the scope of positive action through case law.[155] When assessing such cases, the courts apply different regimes, such as whether the measures are justified, permitted and proportionate,[156] whether there is a legitimate aim[157], and whether the measures are planned and temporary.[158]

In practice, the measures taken are often quotas for underrepresented groups in different occasions.[159] Having quotas for different groups in the training data for machine learning models could be a justified way to ensure full equality in automated decision-making. This topic will be applied to the situations of automated decision-making in chapter 5.1 in the context of data massaging, reweighing and resampling.

It is interesting that most European countries specify only certain of the protected grounds where positive action is used in practice, i.e. positive action can be taken with regard to e.g. disabled people or Roma people, but not people representing sexual minorities or ethnic minorities other than Roma people. Disability and ethnic origin seem to be the grounds based on which positive action is implemented most. In addition to specifying the grounds, many countries limit positive action to certain fields, such as employment or education.[160] This approach

[155] European Commission (2018), p. 89.

[156] In several cases related to proportionate representation of members of ethnic minorities in the state administration, the judiciary and local authority bodies and administrations, the Constitutional Court of Croatia defined positive action measures not to be discriminatory as long as they are justified, permitted and proportionate. See more in European Commission (2018), p. 89.

[157] The requirement of legitimate aim in positive action was recognised by the ECtHR in *Posti and Rahko v. Finland* (2002), para 83.

[158] The Finnish Ministry of Justice published a tool for the assessment of equality upon the entry into force of the new national Non-Discrimination Act (1325/2014) in Finland. The requirements related to positive action are listed as follows on the website: 'Positive discrimination must have an acceptable objective in terms of fundamental and human rights, and must be planned, proportionate and temporary.' See more at http://yhdenvertaisuus.finlex.fi/en/. The website refers incorrectly to term 'positive discrimination', although the translation of the Finnish Non-Discrimination Act uses the correct term 'positive action', which is the name of section 9 of the act: 'Proportionate different treatment that aims to promote de facto equality, or to prevent or remove the disadvantages attributable to discrimination, does not constitute discrimination.'

[159] European Commission (2018), pp. 90–91. For example, at least one person with disability should be appointed to the boards of certain state entities in Malta, whereas in Norway, a moderate quota system in favour of non-ethnic Norwegians was introduced in 12 state-owned companies.

[160] European Commission (2018), pp. 92–93, contains a table of the main grounds and fields where positive action is used in practice in different European countries.

is in line with the aforementioned difficulties related to harmonising the directives related to non-discrimination in terms of protected grounds and areas of applying the directives.

The European Court of Human Rights has stated that differential treatment, whether or not caused by positive action on the part of the State or by a failure to ensure non-discrimination, needs to pursue a legitimate aim, and that there must be a reasonable relationship of proportionality between the aim sought to be realised and the means employed to that end.[161] Therefore, in order to show that positive action does not constitute discrimination, the act needs to have a legitimate aim and be proportional.

[161] *Posti and Rahko v. Finland* (2002), para 83.

4. Data Protection

4.1. General about Data Protection Legislation

Compared to anti-discrimination legislation, one could say that the legislation concerning data protection is more advanced. The level of protection for the processing of personal data has been high in the European Union since the coming into force of the Data Protection Directive in 1995, and its implementation in 1998. The directive secured the most important principles related to the use of personal data. The most recent milestone in data protection was reached on 25 May 2018 when the General Data Protection Regulation came into force. One of the issues with the preceding directive was that it had been implemented in different ways in different Member States. Since regulations, unlike directives, have a direct effect in the European Union law, meaning that the regulations can be invoked before national courts, the level of protection has now been harmonised in the European Union.[162]

The GDPR establishes many obligations to the processors[163] and controllers[164] of personal data, and many rights to the data subjects[165] whose personal data is being processed. As a starting point, any processing of personal data, including the use of personal data as training data in machine learning or using the machine learning model in automated decision-making, must be lawful. In order for the processing of personal data to be lawful, the processing must fulfil all requirements of Article 5 of the GDPR and be based on one of the grounds for lawfulness listed in Article 6.[166] In addition, the definition of lawful contains the requirement for the processing to be in compliance with all other applicable laws, including anti-discrimination laws.[167]

This study will focus on three of the mentioned data protection principles listed in Article 5 of the GDPR, namely the principles of purpose limitation, data minimisation, and data accuracy.

---

[162] Reinisch (2012), p. 63.

[163] Article 4(8) of the GDPR defines processor as a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.

[164] The definition of controller in Article 4(7) of the GDPR, is the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.

[165] Data subject is an identified or identifiable natural person according to Article 4(1) of the GDPR.

[166] de Hert – Papakonstantinou (2016), p. 187.

[167] WP29 Opinion 03/2013, p. 20.

These three principles are perhaps the most relevant in the creation of machine learning applications that can be used in automated decision-making. This is because the data that could be used in the training of such models may have been collected for a purpose other than training an automated decision-making model, which leads to the fact that the right to use the data set for this purpose needs to be obtained from the data subjects. In addition, it has been claimed that in order to build a reliable, non-biased model, large sets of training data are needed, which may go against the principle of data minimisation.[168] Lastly, discriminatory decisions are often made due to inaccurate data, especially when data regarding individuals other than the subject to the decision-making is generalised.

In addition, the GDPR contains provisions related to automated decision-making as such. Article 22 of the GDPR titled 'Automated individual decision-making, including profiling' sets forth the main rule according to which a data subject has the right not to be subject to automated decision-making, as well as exceptions to this main rule. Automated decision-making described in Article 22 GDPR is explained further in chapter 4.2. Provisions related to automated decision-making in the GDPR have been criticised because it is not unambiguous what kind of obligations it sets to the controllers that use automated decision-making methods. These arguments will be elaborated in chapter 4.2.2.

In addition to the data processing principles and specific provisions on automated decision-making, the GDPR provides some remedies for deficient data, the use of which in training of a machine learning model may lead to discrimination or other mistakes. These remedies are right to rectification, right to erasure, and right to restriction of processing.[169] In short, the right to rectification refers to the data subject's right to rectify inaccurate personal data concerning them, as well as to have incomplete personal data completed. The right to erasure, also known as the right to be forgotten, provides the data subject a right to erasure of personal data concerning them, in addition to which the controller has an obligation to erase personal data in certain circumstances. Rectification, completion or erasure may have an effect on the outcome of an automated decision if inaccurate personal data was used in order to reach that decision.[170]

[168] Žliobaitė – Custers (2016), especially pp. 16–17.

[169] The rights to rectification, erasure and restriction of processing are established in Article 16, Article 17, and Article 18 of the GDPR, respectively.

[170] Working Party 29 uses an example of rectification in a situation in which a data subject has been profiled into a group that is most likely to get heart disease. The data subject would have the right to rectify this data by providing more accurate health records, even if it would be statistically correct that they are more likely than the individuals initially compared to them to get heart disease. Therefore, the right to rectification applies in a wide range of situations. An analogy could be drawn between WP29's example and an automated decision-making situation

The right to restriction of processing guarantees to the data subject the right to restrict the controller's processing of their personal data, for example processing in order to make an automated decision, if the accuracy of the personal data is contested by the data subject. Subsequently, the controller needs to verify the accuracy of the personal data. Due to the limited scope of this work, these remedies will not be studied in more detail, and the study will focus on the three above-mentioned data protection principles and automated decision-making. These topics are, however, interesting with relation to automated decision-making, so further research is recommended.

The Convention for the Protection of Individuals with Regard to the Processing of Personal Data (hereinafter 'Convention 108') by the Council of Europe is a remarkable instrument aside the General Data Protection Regulation. It is to be noted that the Convention 108 opened for signature already in 1981, even before the Data Protection Directive. The Convention 108 was modernised in 2018, and the contents of the modernised Convention 108 are similar but not identical to the GDPR. One interesting difference is that Article 6(2) of the modernised Convention 108 states '[appropriate] safeguards shall guard against the risks that the processing of sensitive data may present for the interests, rights and fundamental freedoms of the data subject, notably a risk of discrimination.' This means that the Convention 108 recognises the connection between the processing of sensitive data and discrimination. If it is concluded that sensitive data should be processed in the building phase of a machine learning model in order to prevent discrimination in automated decision-making, this provision may provide guidance on the use of appropriate safeguards on such sensitive data if it is processed for the prevention of discrimination.

where the initial results would show that the data subject is not creditworthy, but the data subject would have the chance to complement their data to prove that their loan application ought to be successful. See WP29 Guidelines (2017a), p. 18. On the other hand, as Koskinen (2018) points out, if data subjects make large amounts of erasure requests regarding data that has been used to train an algorithmic model, erasure of such data may lead to the model becoming less accurate and even discriminative. It is, nevertheless, unclear whether such data related to an individual can even be erased from a machine learning model. As mentioned in chapter Koskinen (2018), p. 243.

4.2. Automated Decision-Making in the General Data Protection Regulation

4.2.1. The Right Not to Be Subject to Automated Decision-Making

The General Data Protection Regulation sets certain restrictions to the use of automated decision-making. It is generally interpreted that the main rule set forth in Article 22(1) of the GDPR is that the data subject shall have the right not to be subject to automated decision-making.[171] However, some academics argue that the wording provides rather a right to object than a prohibition of automated decisions.[172] The conditions for this right to object to automated decisions are that

1) the decision is based solely on automated processing; and
2) the decision produces legal effects concerning the data subject or similarly significantly affects the data subject.

Both of the conditions need to apply simultaneously. However, there are three exceptions to this general rule, situations in which the data subject has no right to object automated decision-making. These are listed in Article 22(2). Automated decision-making is allowed if the decision

a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
c) is based on the data subject's explicit consent.

The exception to these exceptions, GDPR Article 22(4), is that an automated decision shall never be based on special categories of personal data.

The GDPR therefore sets forth a general prohibition of solely automated decisions with a legal or similarly significant effect to the data subject, and the exceptions listed above are the only

[171] For general discussion on Article 22 of the GDPR, see e.g. WP29 Guidelines (2017a). On the right to object to automated decision-making, including profiling, see pp. 18–19.

[172] Wachter – Mittelstadt – Floridi (2017), p. 78.

exceptions to the prohibition.[173] The phrasing '[t]he data subject shall have the right not to be subject to a decision based solely on automated processing' aims at a prohibition, given that it is separate from the right to object to automated decision-making (Article 21), and therefore a prohibition rather than a right to object.[174] Recital 71 implies that in general, automated decision-making is prohibited by stating that it is allowed only in a situation of one of the exceptions. However, it must be noted here that the recitals of EU legislation are not legally binding. They can nevertheless be used in support of the interpretation of the legislation, in addition to which the recitals play a role in creating the reasonable expectations of individuals.[175]

It would supposedly be easy to circumvent the restrictions that Article 22 of the GDPR sets on entities using automated decision-making by involving a human in the loop of the automated decisions, given that the article only applies to solely automated decisions.[176] However, in practice this would not be as attractive to business, since getting rid of the need for a human to take actions is exactly the aim of automating processes.[177] In addition, it has been suggested that minor human involvement, such as pressing a button to confirm decisions made by an algorithm, would not count as a non-solely automated process. Instead, the human involvement should be meaningful, i.e. the human in the loop needs to have the authority to influence the decision.[178]

When it comes to 'legal' or 'similarly significant' effects, there is no definition in the GDPR for either. However, the Working Party 29 has elaborated on the requirement, stating that the requirement for the applicability of Article 22 is that the decision impacts the data subject's legal rights, such as the freedom to vote in an election. A legal effect may also affect a person's legal status or their rights under a contract.[179] The 'legal effects' are mostly related to the public sector, which makes the 'similarly significant effects' more relevant for this study. Recital 71 of the GDPR specifically mentions automatic refusal of an online credit application or e-recruiting practices without any human intervention as examples of decisions with similarly

[173] WP29 Guidelines (2017a), p. 19–20 and 23. See also Zarsky (2016), p. 1015.

[174] WP29 Guidelines (2017a), p. 19.

[175] Wachter – Mittelstadt – Floridi (2017), p. 80.

[176] Zarsky (2016), p. 1016. See also Viljanen (2018), p. 1076 and Information Commissioner's Office (2017), p. 54.

[177] WP29 also recognises this issue: 'Routine human involvement can sometimes be impractical or impossible due to the sheer quantity of data being processed.' WP29 Guidelines (2017a), p. 23.

[178] WP29 Guidelines (2017a), p. 21.

[179] *Ibid.*

significant effects on an individual. WP29 states that the threshold for significance must be similar to that of a decision producing a legal effect. The guidelines list certain effects that the decision may have in order for it to be considered significant, and among these effects, exclusion or discrimination of individuals is mentioned as the most extreme effect. Some examples of significant decisions mentioned include financial decisions, access to health services, denying an employment opportunity, and access to education, all of which have been used as examples in this study as well.[180]

As mentioned, the main rule is that it is prohibited to use automated decision-making if the criteria above are fulfilled, i.e. the decision is based solely on automated processing and produces legal or similarly significant effects to the data subject. Next, the three exceptions to this main rule will be elaborated in more detail.

Firstly, automated decision-making is permitted if it is necessary for entering into, or performance of, a contract between the data subject and a data controller. In order for automated decision-making to be necessary, it must be shown that there was no method available that would be less privacy-intrusive and still effective.[181] Another exception to the prohibition of automated decision-making is a decision based on the data subject's explicit consent. Explicit consent is not defined in the GDPR, but given that automated decisions that affect the data subject's legal rights pose significant data protection risks, a high level of individual control over personal data is deemed appropriate.[182] The data subject should therefore understand what they are consenting to and what the legal or similarly significant effects of the automated decision they are subject to may be. It is likely that consent will be used as the legal grounds for processing in a situation of automated decision-making since the other two exceptions that allow automated decision-making are quite limited. It should be noted that the main rule with the principle of purpose limitation is that personal data can only be used for one purpose under the

[180] WP29 Guidelines (2017a), p. 21–22.

[181] WP29 Guidelines (2017a), p. 23, and European Data Protection Supervisor (2017), pp. 17–18. Working Party 29 provides an example of a situation in which it deems the use of automated decision-making necessary, i.e. a recruitment process in which the company recruiting receives tens of thousands of applications. According to WP29, it would be permitted to shortlist these applications by automated means. In practice, this is indeed the only feasible way to process such applications. It would, however, be justifiable to have a mechanism to understand based on which criteria these automated shortlists are created. It is possible that the algorithm used would ignore applications by people with a certain ethnic origin, or even people from a certain neighbourhood in which the majority of inhabitants would be from a certain ethnic origin (redlining) if the algorithmic model is not trained appropriately.

[182] WP29 Guidelines (2017a), p. 24. See more about consent in WP29 Guidelines (2017b). The data subject should give an express statement of consent, and it may even appropriate that the controller require the data subject to sign such consent statement.

same legal grounds, hence the consent would be needed separately for using personal data in the training of a machine learning model and for making automated decisions regarding the data subject with the said model. Even if automated decision-making may be based on performance of a contract between the controller and the data subject, consent would likely be needed at least for the use of personal data as training data.[183]

The third scenario in which automated decisions are permitted is when Union or Member State law authorises the use of automated decision-making as stipulated in Article 22(2)(b) of the GDPR. Recital 71 of the GDPR states as examples the use of automated decision-making for monitoring and preventing fraud and tax-evasion, or to ensure the security and reliability of a service provided by the controller. The relevant Union or Member State law must also lay down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests. It is therefore the data subject's legitimate interests that need to be taken into consideration, not the processor's. It is interesting that the list of situations in which the data subject cannot object to automated decisions does not contain 'necessary for the purposes of the legitimate interests pursued by the controller or by a third party' even though it is one of the grounds by which processing of personal data in general can be lawful.[184] Therefore, legitimate interests of the controller do not justify an automated decision but only processing of personal data, and even processing of personal data on these grounds requires balancing between these legitimate interests and the fundamental rights and freedoms of the data subject. In practice, the threshold for the controller's legitimate interests weighing more than the data subject's right to privacy may be high.

Recital 71 further states that in case of any of the three exceptions, automated decision-making should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express their point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. In addition, recital 71 takes a stand on the issue of inaccuracies in data that may lead to discrimination:

[183] Koskinen (2018), p. 242.

[184] The grounds for lawful processing of personal data are listed in Article 6 of the GDPR. There are exceptions to this legitimate interests ground, namely that it does not apply to processing carried out by public authorities in the performance of their tasks. See also Koskinen (2018), p. 242.

> '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. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.'

It is unclear what type of technical and organisational measures would be sufficient to ensure data accuracy, probably even to the developers of such technologies. Even though this creates uncertainty, the unspecified language is in accordance with the principle of technology neutrality in regulation, meaning that the regulator should only describe the result to be achieved and leave companies free to adopt whatever technology is most appropriate to achieve the result.[185] Some insights to what sort of technical measures could be used are provided in chapter 5. It is, however, remarkable that the GDPR recognises the threat of discrimination in automated decision-making, and somewhat regrettable that this threat is only brought up in a recital and not the legally binding provisions. On the other hand, it could be argued that the controllers and processors are bound by Article 24(1) and Article 28(1) to implement any technical and organisational measures to ensure the protection of the rights of the data subject, including the measures to prevent discrimination in automated decision-making.

An exception to these exceptions is that if the decision-making involves special categories of personal data, the controller must also ensure that they can meet the requirements of Article 22(4) of the GDPR. Basically, this means that primarily, automated decisions are not permitted even in these exceptional situations if any sensitive data is processed in the decision-making. This restriction as such is already remarkable from the point of view of discrimination, since the special categories of personal data include racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, data concerning health, and data concerning a natural person's sex life or sexual orientation.[186] These are all attributes protected by anti-discrimination legislation, and compliance with such legislation requires that these attributes do not affect decisions made on individuals. The GDPR is therefore in line with anti-discrimination laws on this matter. Exceptionally, automated decisions based on sensitive data are

[185] Reed (2007), p. 264.

[186] The special categories of personal data are listed in Article 9 of the GDPR.

permitted if either the data subject has given explicit consent; or the processing is necessary for reasons of substantial public interest, and in addition to one of those requirements, suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.[187] As mentioned above, it may not be sufficient to delete the sensitive data from the machine learning model because there may be proxies in the data based on which it is possible for the model to deduce the sensitive attributes of data subjects. It has been claimed that it would be crucial that the training data contain also special categories of personal data in order for the machine learning model to learn that the protected grounds shall not have a bearing on the decision regarding an individual.[188]

4.2.2. Transparency and Automated Decision-Making in the GDPR

Article 5(1)(a) of the GDPR requires that personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject. This is relevant especially in the context of automated decision-making. Automated decision-making and especially profiling are often invisible to the consumer, which is not compliant with the GDPR, especially when new personal data is derived from the collected data by profiling the data subjects.[189] Transparency also relates to the data subjects' right to be informed[190], which includes the right to receive meaningful information about the logic involved in the automated decision-making tool, as well as the significance and the envisaged consequences of such automated processing for the data subject. The controller is obligated to provide such information upon their own initiative as soon as they collect personal data, either directly from the data subject or from other sources. In addition, the information needs to be provided to the data subject upon request. These obligations are stipulated in Articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR.[191] In addition to

[187] The requirements can be found in full in Article 9(2) of the GDPR, points (a) and (g).

[188] Žliobaitė – Custers (2016) and Žliobaitė (2017), p. 1068.

[189] WP29 Guidelines (2017a), p. 9.

[190] *Id.*, p. 24–25.

[191] WP29 Guidelines (2017a), p. 25–26. The Working Party gives a concrete example on how transparency could be achieved in a sufficient way: 'An insurance company uses an automated decision making process to set motor insurance premiums based on monitoring customers' driving behaviour. To illustrate the significance and envisaged consequences of the processing it explains that dangerous driving may result in higher insurance payments and provides an app comparing fictional drivers, including one with dangerous driving habits such as fast acceleration and last-minute braking. It uses graphics to give tips on how to improve these habits and consequently how to lower insurance premiums.' Another example of transparency by the legislator can be found in the modernised Convention 108 by the Council of Europe (Convention for the Protection of Individuals with Regard to the Processing of Personal Data), recital 63: 'Where possible, the controller should be able to provide remote access to a secure system which would provide the data subject with direct access to his or her personal data.' Given that the amount of data collected on individuals is massive and the definition of personal data broad, it is likely that creating such a system would be nearly impossible.

transparency, fairness in data processing is a principle that supports non-discrimination in automated decision-making. Working Party 29 highlights that profiling as a form of automated decision-making may be unfair and create discrimination, for example by denying people access to employment opportunities, credit or insurance, or targeting them with excessively risky or costly financial products.[192]

Regarding the obligation to explain the logic involved in automated decision-making, there are various interpretations as to what it means in practice. Some authors simply state that the data subjects have a right to obtain an explanation of the decision reached.[193] Others even disagree that such a right to an explanation exists, arguing that the language of the GDPR is too vague and provides merely a right to be informed about the general logic behind an automated decision-making tool instead of the rationale behind a specific decision reached. In addition, they emphasize that even if there was a right to an explanation, it would only apply when the decision-making is considered solely automatic and having significant effects, which requirements some authors claim to be easy to circumvent, as discussed in chapter 4.2.1.[194] There may also be practical difficulties in providing such an explanation, given that humans may not understand the logic behind the model, especially in case of a layered model such as neural networks.[195]

Gellert et al. conducted a comparative analysis of anti-discrimination and data protection legislation, and came to an interesting conclusion that the protection of these two human rights could be achieved in an almost fully harmonised way.[196] In practice, this could be done by granting the individuals the right to be informed that they are a subject to automated decision-making, in the same way as they have the right to be informed that their data is being processed in accordance with Article 13 of the GDPR. Since the GDPR requires that the data subject is informed about cases of automated decision-making, similarly, individuals could be granted the right to access the decision-making system in the same way that they have the right to access their data when it is being processed, pursuant to Article 15 of the GDPR. The problem with this requirement is the question of trade secrets, i.e. whether the decision-making model can be opened for the public. From a more practical point of view, access to the decision-making tool

[192] WP29 Guidelines (2017a), p. 10.

[193] Voigt – von dem Bussche (2017), p. 61 and 184, and Information Commissioner's Office (2017), p. 54.

[194] Wachter – Mittelstadt – Floridi (2017), pp. 78. See also Edwards – Veale (2017), p. 22, who consider it to be unclear what kind of right to an explanation the GDPR provides.

[195] Goodman – Flaxman (2016), pp. 6–7.

[196] Gellert et al. (2013), pp. 69–71.

might not even be useful to the individual in case they are not knowledgeable enough to investigate the functioning of the tool. According to the comparative analysis, individuals should also have right to object to or restrict the use of an automated decision-making tool, in the same way that they have the right to restriction of processing of their personal data, as well as the right to object to processing of personal data. As shown above, it is generally viewed that Article 22 of the GDPR already grants such a right to data subjects.

There is also some case law on transparency with regard to automated decision-making. In addition to the complaint to the Finnish Non-Discrimination and Equality Tribunal regarding the credit company's automated decision-making, there were two cases against the same credit company in the Office of the Data Protection Ombudsman in Finland. The Non-Discrimination and Equality Tribunal's competence is to evaluate cases from the point of view of the Non-Discrimination Act, whereas it is in the Data Protection Ombudsman's authority to take a stand on the data protection legislation, such as the requirements of the GDPR. There were two different cases examined in this connection. The first case was raised by the data subject, and it concerned the right of access by the data subject in accordance with Article 15 of the GDPR more specifically the data subject's right to inspect the personal data used to assess their creditworthiness, as well as the lawfulness of that data.[197] The second case was initiated by the Office of the Data Protection Ombudsman and it concerned the company's notification practices, namely the information to be provided where personal data are collected from the data subject, specifically with regard to the obligation to provide the data subject with the information on the existence of automated decision-making, including profiling, and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. The latter case was thus an assessment of compliance with Article 13(2)(f) of the credit company.[198] The applicant to the Data Protection Ombudsman's Office was different than the one who filed a complaint with the Non-Discrimination Ombudsman, and this time the applicant claimed to have been discriminated against especially based on their age.[199]

In both cases, the Data Protection Ombudsman used their corrective powers to order the credit company to bring their processing operations into compliance with the GDPR, in accordance

[197] Office of the Data Protection Ombudsman (2019a), p. 5.

[198] Office of the Data Protection Ombudsman (2019b), p. 2.

[199] Office of the Data Protection Ombudsman (2019a), p. 3.

with Article 58 of the GDPR.[200] The first case was divided into two questions. Firstly, the ombudsman found that the use of a categorical upper age limit in assessing creditworthiness is not acceptable. In this decision, the ombudsman applied the Finnish Credit Information Act, which requires that information used as a basis to assess a person's creditworthiness must be related to the person's solvency. The ombudsman also refers to Article 5(1)(a) of the GDPR, stating that personal data must be processed lawfully, fairly and in a transparent manner in relation to the data subject, which is not the case if the processing leads to discrimination. The Data Protection Ombudsman concluded that the age of a credit applicant does not describe their solvency, willingness to pay or ability to deal with their commitments and that in the case at hand, the credit applicant's financial position had not been taken into consideration at all in the automatic processing of the credit application, which is why the credit company shall change their conduct in order to be in compliance with the relevant legislation, including the GDPR.[201]

Secondly, the Data Protection Ombudsman found that the credit company had failed to fulfil the obligations of Article 15(1)(h) of the GDPR, i.e. they had not provided to the data subject meaningful information about the logic involved in the automated decision-making, as well as its consequences for the credit applicant, upon request. In this connection, the ombudsman considered that the company's online credit decision service was automatic decision-making as referred to in Article 22 of the GDPR, and that the grounds by which the company was allowed to use automated decision-making were that the decision was essential in order to conclude or implement an agreement between the company and the credit applicant.[202] Any company planning to implement measures of automated decision-making should therefore be prepared to share information regarding their automated decision-making tool to data subjects.

Apart from the obligation to provide information upon request, controllers have a general obligation to provide information when collecting personal data from a data subject. The second case decided by the Office of the Data Protection Ombudsman was related to this general obligation, as stipulated in Article 13 of the GDPR, specifically section 2(f) on information to disclose when using automated decision-making. Again, the ombudsman ordered the credit company to change their practice. The wording of Article 13(2)(f) is identical to the wording of Article 15(1)(h) in terms of what information is to be provided to the data subject. The

[200] Office of the Data Protection Ombudsman (2019a), p. 6 and 8; and Office of the Data Protection Ombudsman (2019b), p. 2.

[201] Office of the Data Protection Ombudsman (2019a), pp. 6–8.

[202] *Id.*, pp. 8–10.

ombudsman specifies, referring to WP29 guidelines, that there is no need to disclose the algorithm used as a whole but instead, explaining the most important factors taken into account for the decision-making process, the source of the information, and the effect of these factors to the decision. The ombudsman emphasizes that the information should be sufficient for the data subject to understand the reasoning behind the decision affecting them.[203]

4.3. Principle of Purpose Limitation

4.3.1. Compatibility of the Processing for Statistical Purposes

Pursuant to Article 5(1)(b) of the GDPR, controllers and processors need to have a specified, explicit and legitimate purpose for the processing of personal data. In addition to the GDPR, the purpose limitation principle can be found in several other legislative instruments, such as the EUCFR, Convention 108 and the ECHR.[204] However, in the era of big data it often happens that while data is collected for one purpose, the data proves useful for another purpose as well once it's analysed.[205] One example of such a new use case is training an automated decision-making model. As a main rule, processing of personal data for a new purpose would be considered a new processing activity, which must also fulfil the requirements of lawfulness.[206] In these cases, it becomes relevant to study the legislation in order to see if using the data for a new purpose could be lawful on the grounds of being compatible with the original purpose for the processing.[207]

Article 5(1)(b) of the GDPR allows processing for a new purpose, provided that it is compatible with the original purpose. The article further states that compatible refers to processing 'for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes'. Processing for statistical purposes could apply in the case of using personal data as training data in order to build a machine learning model. Big data applications looking for correlations have been considered use for statistical purposes.[208] Automated decision-making models are doing exactly that, looking for correlations in the data to come up with recommendations.

[203] Office of the Data Protection Ombudsman (2019b), pp. 2–3. See also WP29 Guidelines (2017a), p. 27.

[204] WP29 Opinion 03/2013, pp. 6–9.

[205] Colonna (2014), p. 312–313. Forgó – Hänold – Schütze (2017), p. 17 and 20. Zarsky (2016), pp. 1005–1006.

[206] Article 6 of the GDPR lists the grounds by which processing of personal data can be lawful.

[207] Zarsky (2016), p. 1006.

[208] Forgó – Hänold – Schütze (2017), p. 30.

Recital 162 of the GDPR clarifies that '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'. Therefore, the processing for a new purpose cannot create new personal data, instead it has to be data that cannot be traced back to individuals once it has been re-processed. If a machine learning model has used this data for training purposes, will the resulting model count as aggregate data? The challenge in regulating questions like these is the need for the legislator to understand and consider all of the technical details.

Another problematic part is that the results of the statistical processing shall not be used to take measures or make decisions regarding individuals[209], which is exactly the purpose for creating a machine learning model.[210] In addition, recital 162 requires appropriate measures to safeguard the rights and freedoms of the data subject and for ensuring statistical confidentiality to be applied.[211] If finding general correlations is allowed, does it mean that it is allowed to use the machine learning model to predict e.g. risk in admitting a loan, even if it leads to making a decision about an individual based on this aggregate data? This kind of a use case could also be considered parallel to 'improving users' experience', which is considered too vague to fulfil the requirement of a specific purpose.[212] The safeguards and derogations relating to processing for statistical purposes are further stipulated in Article 89 of the GDPR.

In order to be eligible for the exception of statistical purposes, WP29 states that 'functional separation' is needed. This means that the data subject's authorisation would be necessary in order to take a measure or decision related to the data subject. In order to comply, the controller would need to anonymise the data or use other technical and organisational measures to show that the data has been separated from the data subject, and that the further use cannot affect the data subject in any way, be it negative or positive.[213]

[209] This was added to the GDPR by the European Council, i.e. the European Commission did not originally suggest this wording. Council of the European Union (2016), p. 29 (C 159/29).

[210] See e.g. Zarsky (2016), p. 1008.

[211] About further processing for statistical purposes, see also WP29 Opinion 03/2013, p. 28. Note that the opinion has been written with regard to the Data Protection Directive, when the General Data Protection Regulation was only being drafted.

[212] WP29 Opinion 03/2013, p. 16.

[213] *Id.*, p. 30.

4.3.2. Compatibility of the Processing for Other Purposes

The new purpose could be compatible for other reasons than archiving, research or statistical purposes as well, and the GDPR does not explicitly define what kind of purpose is 'compatible'.[214] Article 6(4) of the GDPR lists what needs to be taken into consideration when assessing whether processing for another purpose is compatible with the purpose for which the personal data are initially collected:

> '(a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;
> (b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;
> (c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;
> (d) the possible consequences of the intended further processing for data subjects;
> (e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.'

Out of the above, 6(4)(a) and (6)(4)(d) could be considered when assessing whether further processing for machine learning purposes in order to build an automated decision-making application would be compatible with the original purpose. If the context of collecting personal data was, for example, processing an application by an individual for a loan, this purpose has a link to the processing of future loan applications. However, the link should be unsurprising. Even though the Working Party 29 has proposed that the reasonable expectations of the data subjects as to the further use of personal data would need to be taken into consideration when assessing the compatibility of further use[215], the reasonable expectations have been left out in the wording of Article 6(4) of the GDPR. Reasonable expectations were also not mentioned in the Data Protection Directive. However, recital 50 of the GDPR does mention data subject's reasonable expectations as one prerequisite for compatibility. It has been argued that the reason why reasonable expectations have been placed in the recital instead of the article means that the assessment should rather be objective than subjective.[216] Is it reasonably expectable that

[214] Colonna (2014), p. 303.

[215] WP29 Opinion 03/2013, p. 3. Also the United Kingdom's Information Commissioner's Office notes that the use of personal data in big data applications shall be within people's reasonable expectations. It is interesting that the report focuses on the reasonable expectations relatively much considering that it is something mentioned only in a recital of the GDPR, and also not included in the UK's Data Protection Act 1998 or Data Protection Act 2018. See Information Commissioner's Office (2017), p. 19–20 and 22–27.

[216] Moerel – Prins (2016), pp. 52–53.

one's loan application will be further processed to train a machine learning model with the objective to make the future loan decisions automatically?

If one of the prospects of whether further processing is compatible or not is whether the further processing is surprising to the data subject or not, for example receiving advertisement of products that one browsed through in the internet should not be surprising. Neither should it surprising that the insurance fee offered to a person would be higher because the insurance company has used personal data about the applicant to profile them. For example, the person may have searched on the internet about certain medical conditions, for which reason the insurance company will pump up the fees of the insurance, thinking that the person is likely to be ill. In the machine learning context, the individuals are compared to the individuals in the training data set, or aggregate data about the individuals.[217]

Furthermore, recital 50 of the GDPR provides some freedom for the Member States to determine what should be considered compatible processing in the situation of processing that is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller. This could provide broader rights for the public sector to further process personal data, compared to the private sector. Perhaps using historical data related to tax decisions in order to automatically calculate the tax percentage of an individual would be more acceptable than using job applications to train a recruitment robot selecting which future candidates to invite for an interview. It has been recorded that different Member States determine compatibility in different ways. Some assess the principle from the point of view of the data subject's reasonable expectations, whereas others approach it through the principle of fairness or various balance tests.[218] This authority and differences in interpretation of the Member States weakens the principle of purpose limitation and creates a threat to the harmonisation of data protection laws in the European Union.[219]

[217] See more about the use of aggregate data in order to classify individuals, even based on e.g. ethnical group, in Bayamlioglu (2018), p. 441. See also Information Commissioner's Office (2017), pp. 20–21. In the above-mentioned case in Finland's National Non-Discrimination and Equality Tribunal, use of statistical data to make assumptions on the data subject was also mentioned and found not acceptable due to the Finnish Non-Discrimination Act. Finland's National Non-Discrimination and Equality Tribunal 216/2017, p. 2 and 5.

[218] European Commission (2010), p. 29.

[219] European Data Protection Supervisor (2012), p. 20.

4.3.3. Context of Machine Learning

Principle of purpose limitation is relevant here because with big data and machine learning available, usually the original purpose for which the data is collected, e.g. providing a service, is not the only interesting use case for the data. For instance, an online store collects information about its customers, such as contact details and purchased products, for the purpose of delivering an order. This data could be analysed via machine learning for another purpose, e.g. targeted marketing. In the long run, the store will have enough information about different customers' buying patterns to create an algorithm that can predict which other products a certain user is likely to purchase after their first purchase, and then use this data to advertise those products to the user. This kind of reutilisation of the personal data of the customers may be against the purpose limitation principle.[220] In this situation, also the requirement of Article 30(1)(b) of the GDPR that the purpose of the processing of personal data must be maintained in the records proves challenging. The purpose should be updated whenever the data is used for a new purpose instead of the original purpose stated upon collection of the data.[221] Furthermore, when using discovery-driven data mining, it may not be known at the time of collecting the data what it will be used for. It is characteristic to discovery-driven data mining that the purpose will be discovered only after the data has been collected and analysed.[222]

If the training data was anonymised before it was used for a new purpose, it would not be a problem from data protection legislation's point of view to use the data to train a model.[223] However, it might be impossible to use anonymised data and remove the bias in the training data. There are two methods for anonymisation, randomisation and generalisation.[224] Randomisation means that some attribute in the data is modified in a way that makes it impossible to recognise the identity of the person.[225] If the modified attribute is valuable in terms of determining whether discrimination occurred or not, modifying the attribute will make it impossible to use the data to eliminate the bias. Generalisation means that the data is treated on a higher

[220] Information Commissioner's Office (2017), p. 11–12, and 37–39, however, in this latter chapter, the Information Commissioner's Office evaluates that it is not obvious that further processing of the personal data in big data applications would be incompatible with the original purpose and therefore against the principle of purpose limitation.

[221] Information Commissioner's Office (2017), p. 51.

[222] Colonna (2014), p. 312–313. See also Forgó – Hänold – Schütze (2017), p. 17 and 20.

[223] Recital 21 of the GDPR.

[224] WP29 Opinion 05/2014, p. 3 and 10.

[225] *Id.*, p. 12.

level in order to make it impossible to single out individuals, for example, using the data subject's nationality instead of city of residence.[226] This method may not provide wanted results in terms of removing the bias because even if some attribute related to the data subjects is generalised, it is possible that discriminating factors remain in the data set through proxies.

4.4. Principle of Data Minimisation

The principle of data minimisation can be found in Article 5(1)(c) of the GDPR: '[Personal data shall be] adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed'. It is mentioned throughout the Regulation in other contexts as well, for example in Article 25 titled data protection by design and by default. The article states that the technical and organisational measures taken by the controller shall be designed to implement data protection principles, such as data minimisation. This could be interpreted in a way that would require any build process of a machine learning model to minimise the amount of training data used. The principle of data minimisation is not mentioned in the European Charter of Fundamental Rights, which provides somewhat greater freedom to the legislator in regulating on data minimisation compared to the purpose limitation principle.[227]

The data minimisation principle requires that both the scope and categories of data initially collected are limited to the strict minimum, in addition to which the duration during which personal data may be retained shall be as short as possible.[228] The justification for data minimisation lies in the prevention of misuse of personal data, i.e. the less personal data is processed, the fewer data breaches may occur, which also applies to the storing of personal data and the possibility of cyber-attacks against databases.[229]

The tension between data minimisation and big data is obvious, given that the very purpose of big data is to collect great volumes of data from a variety of sources. Pursuant to Article 89(1) of the GDPR, the principle of data minimisation does not apply to pseudonymised data.[230] However, use of pseudonymised data for training a machine learning model for automated decision-

[226] WP29 Opinion 05/2014, p. 16.

[227] Zarsky (2016), p. 1009.

[228] Zarsky (2016), p. 1009. Regarding limitation of data storage due to data minimisation, see recital 39 of the GDPR.

[229] Zarsky (2016), p. 1009–1010.

[230] See also recital 156 of the GDPR and Zarsky (2016), p. 1011.

making may not be fruitful, as the data may not be so useful after taking such technical and organisational measures.[231]

Data 'minimummisation' is a concept in which instead of restricting the amount of data and using as little data as possible in order not to infringe the data subject's privacy, it would be obligatory to use also the related metadata.[232] That metadata can give a context to the data, making it non-discriminatory: for example, if it is understood that the police have patrolled in a certain area more often than in another area, the reason why people from the first area have a criminal record more often is perhaps statistical bias instead of that group actually being more prone to criminal activity. Therefore, the data can be treated as less reliable and avoid stigmatisation and discrimination this way.[233]

Something interesting to note is that the principle of data minimisation is connected to the principle of data purpose since Article 5(1)(c) of the GDPR specifically states that it depends on the purpose of the processing what amount of data can be considered necessary. Therefore, if the entity collecting the personal data defined the purpose of processing as creation of a non-biased automated decision-making system, it could be considered lawful to use large amounts of personal data, both in terms of scope and categories of data. In this case, the controller would likely need to be able to show that the quality of the resulting automated decision-making model would be remarkably lower if less data was used. In addition, the question would arise whether such a purpose is specified, explicit and legitimate in accordance with Article 5(1)(b) of the GDPR. According to WP29, specified refers to detailed enough to determine what kind of processing is and is not included within the specified purpose.[234] Therefore, possibly some elaboration on the technology used to build the model would need to be provided. Explicit refers to clearly revealed, explained or expressed in some intelligible form, such as in a notice to the data subjects.[235] With regard to legitimacy, the requirement is broader than just collecting the data for one of the legal grounds provided for in Article 6 of the GDPR. The purposes must be in

[231] Zarsky (2016), p. 1011.

[232] Metadata refers to data about when and where and how the underlying information was generated. Kuner et al. (2012), p. 47.

[233] van der Sloot (2013), pp. 282–284.

[234] WP29 Opinion 03/2013, p. 15.

[235] WP29 Opinion 03/2013, pp. 17–18.

accordance with all applicable laws, including non-discrimination laws.[236] Therefore, a balance between the data subjects' right to non-discrimination and privacy must be found.

4.5. Principle of Data Accuracy

The principle of data accuracy[237] is described in Article 5(1)(d) of the GDPR: '[Personal data shall be] accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay'. Quality of data is of utmost importance when data is used in order to train a machine learning model.[238]

Data accuracy is at least as relevant to the algorithmic bias as purpose limitation and data minimisation, although the two latter principles have been addressed in literature more widely.[239] The data may be recorded accurately, e.g. a history of credit decisions may reflect what kind of decisions were made in reality, but data analytics could draw inaccurate or incorrect conclusions from this data. This is due to the fact that there could be hidden biases in the data, or the sampling of the data could be non-representative.[240] In this case, the principles of purpose limitation or data minimisation would not help to solve the problem of discrimination. Instead, the data subject could use their rights of access to the data that is being processed about them, as well as the rights to rectification and erasure. However, even these measures would not solve the issue of having biased data in the model, i.e. if earlier in time, the data subject's loan application had been rejected by a human based on prejudices towards the data subject, the data would be correct and the model would need to be taught not to take into consideration the discriminating historical data.

The right to ask for the rectification or erasure (Articles 16 and 17 of the GDPR) of the individual's data might help to rectify potential discrimination that the use of the model may lead to in case of incorrect data. If the data used as a basis for decision-making is accurate,

[236] WP29 Opinion 03/2013, pp. 19–20.

[237] See more WP29 Guidelines (2017a), pp. 11–12.

[238] Barocas – Selbst (2016), p. 687.

[239] However, for example Koskinen brings up the accuracy principle alongside purpose limitation and data minimisation, stating that data accuracy supports the principle of non-discrimination. She further acknowledges that despite all data being accurate, too little data or prejudices when defining decisional rules may still lead to discrimination. Koskinen (2018), p. 247.

[240] Information Commissioner's Office (2017), p. 43–45.

discrimination should not occur.[241] However, if the model treats data with bias, even correct data may lead to favouring some groups. In addition, the decisions can be made based on historical data combined with the new data subject's data, in which case one would need to be able to rectify the historical data in addition to their own personal data in order to remove the bias.

[241] Finocchiaro – Ricci (2013), p. 295.

5. Technical Solutions and Legislative Initiatives

5.1. Solutions for Pre-Processing, In-Processing and Post-Processing

When building tools to make automated decisions, the issue of discrimination can be addressed in different phases of the process. Firstly, the bias can be removed from the training data. Secondly, the bias can be eliminated during the training of the model by supervising the learning and 'manipulating' how the model learns. Thirdly, the results offered by the model can be rectified.[242] These can be referred to as pre-processing, in-processing and post-processing approaches.[243]

With regard to the solutions in the phase of pre-processing, methods for removing the bias from training data include massaging, reweighing and resampling.[244] Massaging refers to calculating what kind of results should be achieved without the discriminatory attributes and relabelling the training data according to that, for example by relabelling some of the people with a certain ethnic background as having a positive credit score, and some of the people from the privileged ethnic group as having a negative credit score. The objects selected for relabelling are those closest to being classified to the other class.[245] Reweighing means modifying the weights, usually by assigning higher weights for unsuccessful candidates, and lower weights for successful ones.[246] Resampling is the act of deleting and duplicating some parts of the training material, for example deleting some of the men who got a job offer from the data set, and duplicating some of the women who got the offer.[247] It may be considered that measures such as these are necessary in order to remove direct or indirect discrimination from an automated model. In this case, the assumption is that there was unlawful discrimination in the historical decisions, or that the data was incomplete or inaccurate, in which case the model would potentially become discriminatory even if the historical decisions were not. All three of the above could also be considered forms of positive action from legal point of view. Perhaps positive action would be the closest counterpart in non-discrimination legislation since the techniques could be compared to

[242] Kamiran – Calders – Pechenizkiy (2013), pp. 225–226.

[243] Hajian – Domingo-Ferrer (2013), p. 247.

[244] Kamiran – Calders – Pechenizkiy (2013), p. 226, 234–235.

[245] Kamiran – Žliobaitė (2013), pp. 164–165, and Kamiran – Calders – Pechenizkiy (2013), p. 225, 229–230.

[246] Kamiran – Calders – Pechenizkiy (2013), p. 225, 230–232.

[247] Kamiran – Žliobaitė (2013), pp. 165–166, and Kamiran – Calders – Pechenizkiy (2013), p. 225, 230–232.

quotas, which is the measure often used and justified in positive action. The question is whether these measures fulfil the requirements for positive action, such as legitimate aim and proportionality. Quotas could be addressed at the phase of post-processing, i.e. after the automated decision-making model makes suggestions on which candidates to select, the results would be modified so that a certain number of individuals from a protected group would get a positive decision even though the automated model suggested otherwise. An interesting topic for research would be to understand whether modifying the training data in the pre-processing phase through massaging, preferential sampling or reweighing would, in fact, give the same results as adding quotas post-processing. One way to look at the methods is to consider that when the training data is relabelled or parts of the training material are deleted or duplicated, the quotas are included already in the pre-processing phase. For example, a fixed number of women in the training data would lead to the decision-making model offering the job to more women compared to a situation in which the training data was not modified and contained proportionally more men. Regarding the use of sensitive data in automated decision-making, which is restricted in the GDPR, it is obvious from the description of these technologies that in order to insert quotas in the pre-processing phase, the inclusion of sensitive attributes related to the individuals in the training data is necessary.

In-processing approach includes, for example, two interesting solutions to reach fairer machine learning by *Veale* and *Binns*. Firstly, they present a data storing mechanism hosted by a third party for the purposes of collection of protected characteristics about the data subjects. This third party would then be able to both detect and prevent bias. *Veale* and *Binns* argue that this system would surpass auditing mechanisms proposed by *Pasquale* among others[248] because the material needed for the audit is not always in possession of the target for the audit.[249] Secondly, a model of knowledge sharing in the form of an online platform between organisations that use machine learning algorithms is proposed. The challenges that the authors mention therein are i.e. reluctance of companies to participate in such sharing due to the potentially confidential nature of information, as well as additional costs and resources needed.[250] A third suggestion, exploratory fairness analysis, could be used either in the pre-processing or post-processing phase. This method would build hypotheses about the algorithmic models used, either *ex ante* or *ex post*.[251]

[248] See, for example, Sandvig et al. (2014) and Tutt (2016).

[249] Veale – Binns (2017), pp. 5–8.

[250] *Id.*, pp. 8–10.

[251] *Id.*, pp. 10–12.

One of the post-processing solutions was presented by the aforementioned pioneers *Pedreschi, Ruggieri* and *Turini* in 2013. They have a very interesting hypothesis regarding data mining and discrimination. Data mining, and more precisely machine learning on the data, can be used to categorise and profile people. The hypothesis turns this way of using machine learning the other way around and assumes that data mining could be used to automatically discover the patterns of discrimination that emerge from the available data. The idea is that the sets of rules based on which individuals were classified into a certain group are discovered from the model, after which the rules are assessed both individually and in groups to see whether they contain a discriminatory element. In a simplified example, it would be discovered that a negative credit granting decision was based on two attributes of the data subject, ethnic origin and city of residence. The result of those rules together would then be compared to the result of the rules separately, i.e. if 75% of the people with ethnic origin A and city of residence B are denied loan, but only 25% of people with ethnic origin C and city of residence B are denied the loan, it is likely that the automated decision-making model discriminates against people with ethnic origin A.[252] From legal point of view, already being able to show that someone's ethnic origin was used as a basis for a negative decision would be enough to constitute direct discrimination based on a protected ground.

Another state-of-the-art post-processing approach would be convolutional networks and adversarial neural networks. In addition to being able to evaluate the objectivity of the decisions made by the model, this solution would cut down the amount of data needed to train the system. The first use case for this technology was image processing. In that context, the reason why using simple neural networks does not work is that one layer of the neural network can only recognise a pattern when it is similar enough to the original training data. Adding more layers would solve this but also require a vast amount of training data for each layer to be trained. Solution to this is to create convolutional networks that recognise the pattern even if in different positions, orientations, and sizes, which results in the fact that less training data, images, is needed to train the model. These convolutional networks can then be used not only to process existing images, but also to create new images that look real. However, one convolutional network can be clumsy in this task. Solution to this, in turn, is to create a second convolutional neural network that competes with the first one, its task being to recognise when the first

[252] Pedreschi – Ruggieri – Turini (2013), pp. 93–94.

network has created an image that is too different to the images of the training data. Together, these networks can create more accurate results.[253]

The interesting question is whether the above-mentioned technology could be used to detect bias in algorithmic decision-making. The first convolutional neural network could be trained to make recommendations on loan decisions based on minimal amount of data about the subjects. The adversarial network could then be trained to spot discrimination in the recommendations that the first network has made. This idea has been discussed by *Zhang*, *Lemoine* and *Mitchell*.[254] It should be noted that two of the authors of the article are working for a private company and do not represent academia.

Taking the treatment of bias one step further from the post-processing phase, technological due process solutions could be implemented as a solution. This way individuals could easily challenge AI-based decisions that directly affect them.[255] Due process in the Information Age has been discussed by many academics.[256] In practice, this could mean that AI-based solutions would not be enforced instantly, but a complaint period would be applied. This could work well with relation to credit decisions, for example, but in other cases, retrospective redress may not help to compensate for the damages. Such examples have been presented by *Edwards* and *Veale*.[257]

5.2. Legislative Initiatives

Pasquale argues that in order for the legislator to take into account the recent past and even the current state of the art while regulating, information about the technology used ought to be revealed rather sooner than later. There is a conflict between the companies' interests in keeping their new innovations trade secrets and the legislator's interest in understanding how technology works in order to regulate its use. Moreover, it should be meaningful for the corporations to reveal details about their business. It is not enough that wrongdoings are made public without corrective action from the authorities. It is possible that the consumers will not make their decisions based on the reputation of the company, so mere publicity of non-compliance is not

[253] Elements of AI (2018), chapter 5, section III.

[254] Zhang – Lemoine – Mitchell (2018).

[255] The Independent (2018).

[256] See, for example, Citron (2008), and Crawford – Schultz (2014).

[257] Edwards – Veale (2017), p. 42.

enough of a punishment and will not guide the markets to the right direction in terms of being fair and transparent.[258]

A concrete example about using existing legislation in a better way is auditing big data and the algorithms.[259] The healthcare sector in the United States is already using audits of data practices for entities covered by the Health Insurance Portability and Accountability Act (HIPAA). In addition, the Health Information Technology for Economic and Clinical Health Act (HITECH) gives patients rights such as access to their medical records. *Pasquale's* proposal is that the big data companies be taxed in order to establish a government-run auditing system, and to apply some of the principles from HITECH to data brokers as well. For example, data brokers could be obliged to remove health data from data sets that they use to produce reports to employment and insurance companies.[260] The practical deployment could be done in several ways. The audits could be spontaneous checks aimed at companies randomly or only conducted after someone reports concerns about a certain company. In addition to audits initiated by authorities, there could be a possibility for a contracting party to audit the systems of a vendor of an algorithmic solution to automate decision-making. Another question is whether the audit should be done by a government approved auditor that could even be a state agency itself, or whether it should be allowed for private companies to perform such audits, creating a new business opportunity. The General Data Protection Regulation, Articles 28 (h) and 58 (b), could serve as a model when it comes to both audits done by authorities and audits done by controllers in processor's premises.

The existing legislation related to data protection could also be extended to support algorithmic accountability. The GDPR obliges data controllers to perform a data protection impact assessment if their data processing is likely to result in a high risk to the rights and freedoms of natural persons, including a high risk of discrimination.[261] Automated decision-making can therefore require such an assessment. *Kaminski* and *Malgieri* have published an article on the use of data protection impact assessment to secure transparency in the stages of design, development, and training of algorithmic models, taking into consideration views from both Europe and the

[258] Pasquale (2015), pp. 142–143.

[259] Alternatively, the algorithmic decision-making models could be certified by a public authority. See Edwards – Veale (2018), pp. 10–11.

[260] Pasquale (2015), pp. 150–151.

[261] Data protection impact assessment is regulated by Article 35 of the GDPR.

United States.[262] At the moment, algorithmic impact assessments are being used to support the design and deployment phase of artificial intelligence applications and the aim is to find whether a certain type of decision can be fully automated with the use of algorithms, or whether human involvement would be necessary. So far, the assessments are mostly made by private entities independently, but there is a trend of governmental involvement, i.e. several states have started to use algorithmic impact assessments in deployment of artificial intelligence and the surveillance of algorithmic systems.[263]

In addition to legislative initiatives, there are different communities with the aim to solve the human rights issues related to the use of machine learning algorithms, such as discrimination-aware data mining (DADM)[264], fairness, accountability and transparency in machine learning (FATML)[265] and the International Open Data Charter.[266]

[262] Kaminski – Malgieri (2019), especially pp. 7–13.

[263] Koulu et al., p. 22.

[264] Veale – Binns (2017), p. 1.

[265] See the community's website at http://www.fatml.org/.

[266] International Open Data Charter. See the Open Data Charter's website at https://opendatacharter.net/.

6. Conclusion

The main research question of this study was whether the fundamental rights of non-discrimination and data protection can be protected in a balanced way when using machine learning in automated decision-making. In conclusion, there are several factors to take into consideration as a private entity planning to implement automated decision-making tools, both in the phase of building such a tool and in the phase of deploying the tool. The restrictions derive from non-discrimination laws on the one hand, and from data protection laws on the other hand, and their compatibility is somewhat unclear.

The first research question identified was what kind of situations are protected by prohibition of discrimination, and whether the use of machine learning and automated decision-making is relevant for those situations. The result is that depending on the area in which the automated decision-making occurs, different protected groups enjoy a different level of protection in the light of the current EU law. In employment, the level of protection is wide, and many minorities are specifically mentioned as being protected from discrimination, whereas for example in the offering of goods and services, people of different religions are not mentioned as a protected group. However, treating all minorities equally in the decision-making model should be the starting point regardless of the area of decision-making. Even though the level of protection lacks harmonisation in the EU directives, the international human rights and fundamental rights treaties guarantee the general principle of equality and prohibition of discrimination regardless of the grounds for discrimination. In addition, there is an attempt to harmonise the directives as well through the proposal for Council directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation.

Another sub-question for this research was examining the requirements of non-discrimination when deploying the automated decision-making model. The outcome of this question is that discrimination can occur both because of decisional rules and because of biased data, both of which shall therefore be audited in order to fulfil the requirements of non-discrimination laws. Concerning the decisional rules in the decision-making model, it needs to be ensured that they are not discriminatory, neither directly nor indirectly. Assessing whether a decisional rule is directly discriminatory may be simple, as was seen in the Svea Ekonomi case where the outcome of the automated decision was directly dependent on qualities such as gender and age of the applicant, those qualities being protected grounds. Not being granted credit or being granted less credit than the comparator constituted less favourable treatment. The comparator would be

a person of e.g. different age or gender who did not have payment defaults, i.e. a person is in a similar situation, the only difference between the persons being the protected ground. Even the causal link between the protected ground and the less favourable treatment was clear, given that it could be observed in the model that the outcome of the credit decision would have been more favourable if e.g. the gender of the applicant had been different. The causal link is not always this easily discovered in a more complex machine learning model, which is why in some cases, showing direct discrimination may require running several tests in the automated decision-making model and assessing whether there seems to be a pattern in minorities being treated less favourably, even if not all people belonging to a protected group are systematically discriminated against. This could be done by submitting several applications to the model and evaluating the outcome, for example by categorising the applicants based on their protected grounds and comparing the outcome suggested by the model. For instance, if there were one hundred female and male applicants and the model suggests to grant the credit to 80% of the female applicants but only 50% if the male applicants, direct discrimination based on gender is likely to occur even if it is not possible to interpret the decisional rules as such in an understandable way.

When it comes to showing that indirect discrimination occurs, it may be more complicated. Since the indirectly discriminatory decisional rule will seem neutral, it will not catch the attention of the person auditing the system. It may be necessary to go through all decisional rules and run the automated decision-making model with several individuals as an input, the difference between them being a protected attribute, in order to see whether the decisions are significantly more negative in their effects on one or more of the protected groups. What makes the audit even more complicated is that sometimes, the protected attribute is not e.g. the ethnic origin of the individual but a neutral proxy, such as a ZIP code. Therefore, the difference between the comparators may not be the sensitive attribute itself but something else that is correlated with such a sensitive attribute.

Lastly, in order to avoid discrimination, the data used to train an automated decision-making model needs to be representative, which possibly means that a massive amount of training data, including sensitive attributes, is needed. This leads to the factors to take into account from data protection laws' point of view. It is important to note that all of the conditions for non-discrimination need to be fulfilled in order to meet the requirements of data protection laws because data processing must be lawful, thus non-compliance with anti-discrimination laws when processing personal data is also a breach of the General Data Protection Regulation.

The other requirements based on the data protection laws can be divided into requirements related to building a data-driven automated decision-making model and requirements related to using the resulting model. This separation was also made while defining the research questions, one of which was to identify these requirements. Most of the requirements apply in both situations. In the building phase, the data subjects need to be informed that their personal data is used for the purpose of creating a machine learning model, regardless of whether the data is collected directly from the data subjects or other sources. If it is not possible to inform the data subjects on this, it can be argued that the creation of the model is compatible with the original purpose for which the data was collected, and thus in compliance with the purpose limitation principle. In addition, the model needs to be built with as little data as possible in order to achieve the purpose. The data used must also be accurate and the data subjects must have the right of rectification and erasure already in the building phase, as well as when the model is running.

One of the findings is that the principle of data accuracy is very relevant in terms of discrimination in automated decision-making. As mentioned, apart from discriminatory decisional rules, the other threat to non-discrimination is the quality of data. Nevertheless, many studies made so far on discrimination focus on data minimisation and purpose limitation principle, even though data accuracy may be even more relevant in the attempt to create accountable algorithms. The principle of data accuracy, especially its relation to the right to rectification, right to erasure, and right to restriction of processing, is therefore an interesting field of research for future studies. The hypothesis is that in situations in which automated decisions are based solely on the personal data of the data subject, the risk of discrimination in such decisions is low as long as the aforementioned rights are actively exercised and as a result, the data is accurate. However, there are several questions regarding the feasibility of this solution. Firstly, it may not be fair that the responsibility on the accuracy of the data is on the data subject instead of the processor and the controller, i.e. that it would depend on the activity of the data subject to rectify, erase and contest the accuracy of their personal data, hence restricting the processing. Secondly, often it is not only the personal data of the individual subject to a certain decision that affects the outcome of the decision. Instead, the machine learning model compares the data of several individuals in order to find patterns in the data and come to conclusions on which

decision to make.[267] Therefore, it could be that even if all data regarding the data subject is accurate, mistakes or bias in other individuals' data may lead to discriminatory decisions.

Furthermore, in the deployment phase the most relevant data protection provisions concern automated decision-making as such. First of all, there is, arguably, a general prohibition to make automated decisions. If it can be argued that making automated decisions is allowed in a certain case, the controller has an obligation to inform the data subjects on the fact that they are subject to automated decisions. The controversial obligation to explain the logic behind the automated decisions also becomes applicable in this phase. Even if it would not be needed to explain the rationale behind a certain decision to the subject, the explanation needs to be sufficient to help the data subject contest the decision. In addition to this, the data subject always has a right for manual processing.

All in all, until now the requirements of non-discrimination and those of data protection are often assessed separately, even though they may be contradictory and thus impossible to fulfil at the same time. More research combining the two points of view would be beneficial in order to create more practical legal solutions to the issue at hand. Speaking strictly from the technical point of view, it appears that many solutions both to biased data and discriminatory decisional rules have already been developed. What is lacking is a legal response to the legitimacy of using such methods.

Such a legal response is necessary because as long as there is uncertainty related to how the existing legislation is applied to new technologies and no case-law concerning the various algorithmic solutions for prevention of discrimination, it is not clear which ones of them are legally compliant. It is therefore a risk for the entities building and executing automated decision-making to implement them. No company wants to be the first one sued and found guilty of illegitimate data manipulation, for instance. However, being passive and not developing the systems by trying out e.g. massaging, reweighing or resampling can also be enough to find an entity guilty of discrimination. It is to be noted that in cases that could be brought to the court on the basis of discrimination in automated decision-making, the burden of proof to show that the process was not discriminating lies on the respondent.[268] Therefore, if an applicant brings a

[267] Information Commissioner's Office (2017), pp. 20–21.

[268] Directive 2006/54/EC of the European Parliament and of the Council of 5 July 2006 on the implementation of the principle of equal opportunities and equal treatment of men and women in matters of employment and occupation, article 19.

claim against a company saying that their automated decision-making model is discriminating, the company needs to be able to justify why discrimination did not occur. In this situation, it can be a risk if the company applied one of the algorithmic solutions unsuccessfully and fails to show that what they did eliminated the bias. On the other hand, reasonable effort to aim at compliance can be rewarded in the court, especially in cases where the state of the art is not sufficient to understand which solutions to the problem work and which do not. Being able to show that the company tried to make their data processes unbiased could even be enough to find the respondent innocent.

Regarding the question whether extensive new legislation regarding artificial intelligence and the ethics of algorithms is needed, my personal opinion is that instead, the existing legislation should be applied. In principle, it should not matter whether it is a human or a robot making discriminatory decisions: people can be trained to understand the principle of equality, and it appears that also machines are able to learn not to discriminate against minorities, as long as this is taken into consideration in the training phase. However, the existing legislation does seem to create legal uncertainty due to the fact that it is not clear enough how to comply with the requirements of the laws in practice, especially when using modern technologies.[269] While the approach of technology neutrality is recommended, the legislator should have a practical view in order not to leave all questions on how to apply a certain piece of legislation in real life to the courts to decide. Legislation should be flexible enough to be applied to various technical situations in order to avoid the need to draft new provisions whenever a new innovation has been created.

The final conclusions of this study return to the main research question, i.e. whether the fundamental rights of non-discrimination and data protection can be protected in a balanced way when using machine learning in automated decision-making. There are technical ways to eliminate bias in data, as well as to investigate the discriminatory nature of decisional rules. In addition, it could be argued that neither the purpose limitation principle nor the data minimisation principle prohibit the use of even large data sets in the training of the models, as long as the purpose, i.e. building a non-biased model, is legitimate. The use of sensitive personal data in compliance with the GDPR is very restricted, which may make it difficult to teach the machine learning model to ignore the potential bias in the historical data. It appears that the legislator has recognised that the concepts of discrimination and data protection are closely related

[269] A good example of this is the debate over the 'right to an explanation' over Articles 13, 14 and 15 of the GDPR.

but the practical guidance on how to address these issues is limited. Therefore, in order to find the balance, more research is needed, especially research combining legal expertise on non-discrimination with technical expertise on automated decision-making models and the logic behind them.