UiT The Arctic University of Norway

Faculty of Law

Bias and Discrimination in Clinical Decision Support Systems Based on Artificial Intelligence

Mathias Karlsen Hauglid

A dissertation for the degree of Philosophiae Doctor (PhD) – November 2023

Table of Contents

PART I: INTRODUCTION AND FOUNDATIONS

1 Introduction

1.1 Subject Matter and Objective

Artificial Intelligence (AI) is a class of computer programs capable of performing tasks that require a certain problem-solving capacity.[1] Amongst the current capabilities of AI systems are various forms of pattern recognition, natural language processing, classification of information, and prediction of future events, outcomes, or behaviours.

One of the most promising utilisations of AI in healthcare is Clinical Decision Support (AI-CDS) systems: AI systems that support clinical assessments and decision-making by producing relevant classifications and predictions that may be relied on by clinicians and patients.[2] At the heart of AI-CDS systems lie complex statistical models; sets of rules or patterns derived from data by Machine Learning (ML) algorithms. ML algorithms are procedures that allow a computer program to learn patterns from data – thus, they can be described as recipes for learning.[3] For example, an ML algorithm might learn that a certain combination of words in an electronic health record is correlated with increased probability of a specific disease or that patients with certain characteristics are more likely than others to benefit from a given treatment.

The subject matter of this thesis is bias and discrimination in AI-CDS systems. AI is becoming increasingly recognised for its potential to improve healthcare by facilitating accurate, personalised, rapid, and efficient clinical decisions.[4] However, an oft-cited concern is that AI systems can be 'biased' – a nebulous term often used to describe certain undesirable

---

[1] More on how the technology works and what it can do in section 1.4.3.

[2] Thomas Davenport and Ravi Kalakota, "The Potential for Artificial Intelligence in Healthcare," *Future healthcare journal* 6, no. 2 (2019), https://doi.org/10.7861/futurehosp.6-2-94; Adrienne Kline et al., "Multimodal Machine Learning in Precision Health: A Scoping Review," *Nature: NPJ Digital Medicine* 5, no. 1 (2022): 1, https://doi.org/https://doi.org/10.1038/s41746-022-00712-8.

[3] Section 1.5.3.

[4] European Parliamentary Research Service: Study Panel for the Future of Science and Technology, *Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts* (June 2022); see also section 1.6.1.

side-effects of AI.[5] While the term ‘bias’ is explored in more depth in Part II, it can be understood as a systematic tendency to treat certain individuals or groups differently compared to others. There exists a widespread apprehension that the use of AI to aid decision-making might, due to the presence of ‘bias,’ produce results that amount to discrimination.[6]

---

[5] Batya Friedman and Helen Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems* 14, no. 3 (1996); Moritz Hardt, "How Big Data Is Unfair," *Medium*, *Medium*, 26 September, 2014, https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de; Andrew Guthrie Ferguson, "Policing Predictive Policing," *Washington University Law Review* 94 (2016): 1146, etc.; Pauline T Kim, "Data-Driven Discrimination at Work," *William & Mary Law Review* 58, 3 (2016); David Danks and Alex John London, "Algorithmic Bias in Autonomous Systems," *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence* 17 (August 2017); Monique Mann and Tobias Matzner, "Challenging Algorithmic Profiling: The Limits of Data Protection and Anti-Discrimination in Responding to Emergent Discrimination," *Big Data & Society* 6, no. 2 (2019), https://doi.org/10.1177/2053951719895805; Trishan Panch, Heather Mattie, and Rifat Atun, "Artificial Intelligence and Algorithmic Bias: Implications for Health Systems," *Journal of Global Health* 9, no. 02318 (2019), https://doi.org/10.7189/jogh.09.020318; Mireille Hildebrandt, "The Issue of Bias: The Framing Powers of Machine Learning (Draft Version/Pre-Print)," in *Machines We Trust: Perspectives on Dependable AI*, ed. Marcello Pelillo and Teresa Scantamburlo (The MIT Press, 2021).

[6] Oscar H Gandy, "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems," *Ethics and Information Technology* 12, no. 1 (2010), https://doi.org/10.1007/s10676-009-9198-6; Devin G Pope and Justin R Sydnor, "Implementing Anti-Discrimination Policies in Statistical Profiling Models," *American Economic Journal: Economic Policy* 3, no. 3 (2011), https://doi.org/10.1257/pol.3.3.206; Executive Office of the President and John Podesta, *Big Data: Seizing Opportunities, Preserving Values*, United States: White House (2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf; Ferguson (2016) 1148; Bryce Goodman, "Discrimination, Data Sanitisation and Auditing in the European Union's General Data Protection Regulation," *European Data Protection Law Review* 2, no. 4 (2016); Kim (2016); Michael Veale and Reuben Binns, "Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data," *Big Data & Society* 4, no. 2 (2017); Stephanie Bornstein, "Antidiscriminatory Algorithms," *Alabama Law Review* 70 (2018); Ignacio N Cofone, "Algorithmic Discrimination Is an Information Problem," *Hastings Law Journal* 70, no. 6 (2018): 1394; David Jacobus Dalenberg, "Preventing Discrimination in the Automated Targeting of Job Advertisements," *Computer Law & Security Review* 34, no. 3 (2018), https://doi.org/10.1016/j.clsr.2017.11.009; Philipp Hacker, "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under Eu Law," *Common Market Law*

While ‘discrimination’ is a word that tends to be used in general societal discourse to describe differential treatment perceived to be unfair, this thesis is concerned with ‘discrimination’ as understood within the EU legal order. In this latter sense, ‘discrimination’ refers to decisions, actions or practices that violate the fundamental right to non-discrimination, also referred to as the ‘non-discrimination principle’ or ‘equal treatment principle.’ This principle is enshrined in Article 21 of the Charter of Fundamental Rights of the European Union (the ‘Charter’), which has constitutional status within the EU legal order. It is also expressed in secondary EU legislation pertaining to certain sectors, including healthcare.[7]

With the aim of ensuring the effective protection of the safety and fundamental rights of EU citizens, including the right to non-discrimination, the EU legislature has proposed a common European regulatory framework for AI systems – the Artificial Intelligence Act (‘AI Act’/’AIA’). At the time of submitting this thesis, in the fall of 2023, the final details of the AIA have yet to be decided on a political level. The European Parliament is, as of the time of

---

Review 55, no. 4 (2018); Selena Silva and Martin Kenney, "Algorithms, Platforms, and Ethnic Bias: An Integrative Essay," *Phylon* 55, no. 1 & 2 (2018), https://doi.org/10.2307/26545017; Betsy Anne Williams, Catherine F. Brooks, and Yotam Shmargad, "How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications," *Journal of Information Policy* 8 (2018), https://doi.org/10.5325/jinfopoli.8.2018.0078; Frederik Zuiderveen Borgesius, *Council of Europe Study on Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*, Council of Europe: Anti-Discrimination Department (2018), https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73; Maddalena Favaretto, Eva De Clercq, and Bernice Simone Elger, "Big Data and Discrimination: Perils, Promises and Solutions. A Systematic Review," *Journal of Big Data* 6, no. 12 (2019), https://doi.org/10.1186/s40537-019-0177-4; Anna Lauren Hoffmann, "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse," *Information, Communication & Society* 22, no. 7 (2019); Robin Allen and Dee Masters, "Artificial Intelligence: The Right to Protection from Discrimination Caused by Algorithms, Machine Learning and Automated Decision-Making," *ERA Forum* 20, no. 4 (2020), https://doi.org/10.1007/s12027-019-00582-w; Janneke Gerards and Raphaële Xenidis, *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law*, European Commission Directorate-General for Justice and Consumers (Brussels: Publications Office, 2021), https://op.europa.eu/en/publication-detail/-/publication/082f1dbc-821d-11eb-9ac9-01aa75ed71a1; European Union Agency for Fundamental Rights (FRA), *Bias in Algorithms: Artificial Intelligence and Discrimination* (Vienna, 2022 2022), https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

[7] The applicability of EU non-discrimination law to AI-CDS systems is discussed in chapter 6.

writing, negotiating with the European Commission and the EU Council, to (hopefully) arrive at a consolidated version of the AIA.[8] If enacted, the AIA will require that certain preventive measures must be taken to ensure the effective protection of safety and fundamental rights, as a prerequisite for the deployment of an AI system.[9] Although the exact scope of these preventive measures is not yet determined, they are likely to entail that AI developers and/or users must conduct various assessments before deployment. In the context of this thesis, the term 'deployment' refers to placing an AI system on the EU market,[10] or putting it into service for the first time.[11] Assessments conducted before any of these events, for the purpose of determining whether an AI system may be deployed, are generally referred to in this thesis as 'pre-deployment assessments.' Pre-deployment assessments may take various forms and rely on various methodologies. For example, one category of pre-deployment assessments is those that are based on risk assessment methodologies (risk assessments).[12]

The point of departure for this thesis is the assumption that a pre-deployment assessment, in whatever form, may encompass aspects of discrimination, which implies an application of the non-discrimination principle in EU law. If discrimination, in this sense, is included in a pre-deployment assessment, the part of such an assessment that encompasses discrimination is referred to herein as a 'pre-deployment discrimination assessment.' One way for the AI Act to pursue its objective of effective fundamental rights protection could be to require that pre-

---

[8] The fifth and final trilogue meeting, for political negotiations, is set to 6 December 2023. If not finalised, there could be a delay in the adoption of the AI Act until after the upcoming EU election in June 2024, which may in turn lead to different positions on the draft regulation.

[9] This focus on preventive compliance measures is ingrained in the overarching regulatory approach taken in the AIA proposal from the European Commission. In their respective negotiation versions, the EU Council and the European Parliament have not deviated from this overarching regulatory approach.

[10] "Placing on the market" is a defined term in the AIA, which refers to "the first making available of an AI system on the Union market:" Article 3(9) AIA (EP). The phrase "making available on the market" is defined as the "supply of an AI system for distribution or use on the Union market in the course of a commercial activity, whether in return for payment or free of charge:" Article 3(10) AIA (EP).

[11] "Putting into service" is a defined term in the AIA, which refers to "the supply of an AI system for first use directly to the deployer or for own use on the Union market for its intended purpose:" Article 3(11) AIA (EP).

[12] Various types of pre-deployment assessments are discussed in chapter 7.

deployment discrimination assessments must be conducted before an AI system is deployed. The extent to which the proposed AIA entails such requirements is examined in Part III. As Part III elaborates, when risk assessments and other types of pre-deployment assessments are required by law, these requirements are often part of a regulatory strategy called 'meta-regulation,' involving the reliance on regulated entities to conduct certain assessments aimed at ensuring compliance with legal requirements. In the context of this thesis, the relevant assessment requirements are those obliging providers and deployers of AI-CDS systems to carry out certain assessments *before* deployment, for the purpose of preventing discrimination from occurring *after* deployment. In the context of this thesis, the assessor is typically an AI developer aiming to place an AI system on the market (a 'provider' in the context of the AIA) or a healthcare institution contemplating whether to put an AI system into service (a 'deployer' in the context of the AIA). However, a pre-deployment discrimination assessment may also be conducted by third parties, such as external auditors or supervisory authorities.

How to conduct a pre-deployment discrimination assessment, regardless of which underlying assessment methodology one applies (e.g., risk assessment), represents a considerable uncertainty for assessors. While the core of the non-discrimination principle has gained clarity over time, it is not evident what discrimination means in the context of AI-CDS systems,[13] or *how* discrimination can be assessed, particularly *before* an AI-CDS system is deployed. Non-discrimination law is typically applied 'ex post,'[14] and on a case-by-case basis. This implies an urgent need for development of a system of methods and principles, i.e., a 'methodology,' for assessing discrimination in AI-CDS systems.[15] In practice, discrimination is one of several

[13] On the ambiguity of discrimination as a legal concept, see: Goodman (2016) 506. Moreover, Zuiderveen Borgesius sees the ambiguity of non-discrimination law as weakness in the context of AI-based decision-making: Zuiderveen Borgesius (2018): 19. However, it should be noted that fundamental rights are purposefully articulated in a broad manner, to cover the wide array of situations where the need for protection may arise: Nathalie A Smuha, "Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea," *Philosophy & Technology* 34, no. 1 (May 2020): 9, https://doi.org/10.1007/s13347-020-00403-w.

[14] The term 'ex post' is used here to denote that the law is applied after an incident has occurred, after a claim has been made, or after a product has been placed on the market.

[15] A 'methodology' can be understood as "a system of methods and principles for doing something": Collins Dictionary, s.v. "Methodology," accessed October 6, 2023, https://www.collinsdictionary.com/dictionary/english/methodology; Cambridge Dictionary, s.v.

issues that assessors may consider as part of a broader pre-deployment assessment based on an underlying assessment methodology, such as a risk assessment methodology. Consequently, whichever underlying assessment methodology one applies, the part of a pre-deployment assessment that considers discrimination needs to be informed by methodological elements based on the non-discrimination principle in EU law. This thesis aims to contribute to the development of such methodological elements.

As already alluded to, a 'methodology' is understood as a system of methods and principles. Such a system of methods and principles typically encompasses the procedural steps of how something is done, as well as the underlying strategies and rationales of doing something, and of doing it in a certain way.[16] In accordance with this understanding of what a methodology is, this thesis aims to develop considerations that should be included when assessing discrimination in AI-CDS systems before deployment, as well as the principles and criteria guiding such an assessment. The 'considerations' developed in this thesis are essentially questions that an assessor should raise and consider during a pre-deployment discrimination assessment. The implications of the answer to these questions depend on the relevant principles and criteria. These principles and criteria are developed through an interpretation and contextual adaptation of EU non-discrimination law.[17] Moreover, where specific methods are identified which may be used to conduct the proposed considerations in practice, such methods are highlighted.

Even though the objective of the thesis relates to the application of non-discrimination law in a specific, practical setting, the thesis does not aim to develop a complete system of methods or principles (methodology) or a readily applicable, practical assessment framework. Rather, the methodological elements developed in this thesis could serve as a basis for further development, refinement, and integration into broader assessment methodologies. For

---

"Methodology," accessed 6 October, 2023, https://dictionary.cambridge.org/dictionary/english/methodology (defining a methodology as "a system of ways of doing, teaching, or studying something").

[16] The word 'methodology' consists of *method* and *-ology*. While the first part of the word refers to a way of doing something ('method'), the latter part refers to the knowledge or science of doing something.

[17] Chapter 2 further describes the method of developing methodological elements in this thesis.

example, the thesis does not aim to explore in-depth how the methodological elements developed here can be operationalised within a risk assessment methodology. This is a recommended question for further research.

Moreover, it is important to note that this thesis offers methodological elements based on an interpretation of the non-discrimination principle in EU law. Thus, the methodological elements are developed with the ambition that they should represent a proper legal interpretation that is loyal to the jurisprudence of the Court of Justice of the European Union (CJEU).[18] They are not primarily developed with a view to maximising practical feasibility or efficiency. From a practical perspective, the emphasis on CJEU jurisprudence could be seen as a limitation. However, it is argued in this thesis that relevant pre-deployment discrimination assessment requirements indeed refer to the non-discrimination principle in EU law. Consequently, it is relevant to develop methodological elements for an assessment based on this principle. In future discourse on pre-deployment assessments for AI-CDS systems, practical challenges of operationalising the contributions of this thesis could be addressed and serve as a basis for policy developments and improved assessment methodologies.

The objective of this thesis is illustrated by the following figure, which uses the pre-deployment assessment requirements identified in Part III as examples. The figure illustrates how the thesis aims to develop methodological elements that can be operationalised across different underlying assessment methodologies:

[18] The method is further explained in chapter 2.

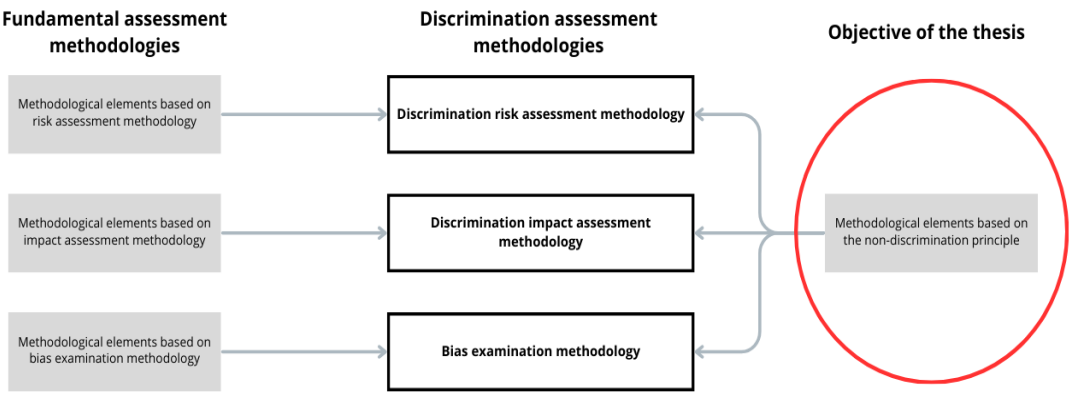


Figure 1: The objective of the thesis is to develop methodological elements based on the non-discrimination principle. These methodological elements are not adapted to one specific type of pre-deployment assessment, but are meant to be operationalised within any pre-deployment assessment context where discrimination is to be considered.

Importantly, this thesis does not aim to specify standards or thresholds for determining whether an AI-CDS system should be deployed. The decision to deploy an AI-CDS system ultimately rests on the assessor and depends on the underlying assessment methodology applied in each case. For example, if risk assessment is the underlying assessment methodology, the decision to deploy an AI-CDS system will depend on whether the assessor deems the risks posed by the system acceptable.[19] While the law may require that a risk assessment must be conducted before deployment, it does not define the threshold for acceptable risk. The implications of different underlying assessment methodologies for the decision to deploy an AI-CDS system, are discussed in Part III.

Furthermore, it is important to note that the subject matter of this thesis – bias and discrimination in AI-CDS systems – is rapidly evolving and remains a relatively nascent research topic. It is to be expected that there may be gaps between the scrutinization of an AI-CDS system that is required according to an interpretation of EU law and the technical methods of scrutinization that currently exist. As regards existing technical methods for testing and scrutinization of AI systems, it is also worth noting that this thesis does not carry out a systematic review aimed at mapping such methods. While this thesis gives examples of potentially relevant technical methods, it must be assumed that many more exist. There is an

[19] As further discussed in section 7.6.

increasing body of technical literature concerning methods of testing that can be used to detect and, thus, assess bias and potential discrimination in AI systems.[20] A systematic mapping and investigation of technical methods for assessing discrimination in AI-CDS systems lies outside of this thesis's scope, but the thesis provides certain points for future research efforts in this direction.

While this thesis develops methodological elements aimed specifically at a pre-deployment context, discrimination assessments should also be conducted reiteratively during the operational life of an AI-CDS system. Indeed, AI providers are obliged under the AI Act to establish systems aimed at monitoring an AI system's behaviour after deployment.[21] To some extent, the methodological elements developed in this thesis can be utilised also in the context of such monitoring systems. Also, to some (lesser) extent, these elements may be relevant in the context of ex post litigation, where one needs to assess whether an individual has been discriminated against. The thesis does not develop methodological elements tailored to the ex post context, but it provides generalisable insights into the interpretation and adaptation of the non-discrimination principle to the specific context of AI-CDS systems.

1.2 Research Questions

To develop methodological elements of a pre-deployment discrimination assessment, this thesis considers several research questions. First, a conceptual and technical understanding of the bias problem, including how and why this problem occurs in AI-CDS systems, is

---

[20] Reva Schwartz et al., "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence," *NIST Special Publication* 1270 (2022): 14-15, https://doi.org/10.6028/NIST.SP.1270; Florian Tramer et al., "Fairtest: Discovering Unwarranted Associations in Data-Driven Applications" (paper presented at the 2017 IEEE European Symposium on Security and Privacy (EuroS&P), 2017); Rico Angell et al., "Themis: Automatically Testing Software for Discrimination" (Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Lake Buena Vista, FL, USA, Association for Computing Machinery, 2018); Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou, "Fairness Testing: Testing Software for Discrimination" (Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, Paderborn, Germany, Association for Computing Machinery, 2017); Sakshi Udeshi, Pryanshu Arora, and Sudipta Chattopadhyay, "Automated Directed Fairness Testing" (Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, Montpellier, France, Association for Computing Machinery, 2018).

[21] Article 61(1) AIA (EP).

established. The underlying rationale for requiring a pre-deployment discrimination assessment is the assumption that AI systems might cause discrimination, due to the presence of biases. Consequently, a basic conceptual and technical understanding of the bias issue is foundational to the objective of this thesis. The following research questions are pursued in the interest of developing a conceptual and technical understanding of the bias problem:

- How can 'bias' be conceptualised and defined in the context of AI-CDS systems?
- Through which mechanisms may relevant biases, from an equality perspective, occur in AI-CDS systems?

For the purpose of establishing the need for the methodological elements developed in this thesis and identifying the legal context in which this need arises, the following questions are addressed:

- To what extent are pre-deployment discrimination assessments required in the legal framework applicable to AI-CDS systems? Relatedly, what are the different types of discrimination assessments that are required, and what determines the decision to deploy an AI-CDS system according to these assessment types?

To develop methodological elements of assessing discrimination based on the non-discrimination principle in EU law, the thesis addresses the following research questions:

- What considerations should be included when assessing discrimination in an AI-CDS system before deployment?
- What principles or criteria should inform those considerations and, thus, the decision on deployment of an AI-CDS system?

Recognising that this thesis merely represents an early effort towards developing pre-deployment discrimination assessment methodologies for AI-CDS systems, the thesis also identifies recommendable directions for future research into both legal and technical issues related to such assessments.

1.3 Content and Structure of the Thesis

Part I of this thesis introduces the subject matter and explains the essentials of AI-CDS systems, the technology they rely on, and how they are developed and used in practice. Part I also describes the research method applied in the thesis.[22]

Part II delves into the problem that forms the backdrop for the thesis, and for the introduction of pre-deployment discrimination assessment requirements in the AI Act: the concern that AI-CDS systems might violate the non-discrimination principle due to the presence of biases. Part II comprises chapters 3-5. Chapter 3 aims to establish an understanding of the concept of 'bias' by drawing on insights from multidisciplinary literature. It also constructs a working definition of 'bias' which is relied on throughout the subsequent parts of the thesis. Building upon the conceptual understanding established in chapter 3, chapter 4 employs the notion of 'equality' to sharpen the focus of thesis on the specific types of biases that may be potential causes of discrimination. Furthermore, chapter 4 describes the key mechanisms that can give rise to such equality-related biases in AI-CDS systems. Chapter 5 introduces two distinct case studies illustrating the issue of bias and potential discrimination in AI-CDS systems.

Part III describes the legal framework of EU law that governs clinical decision-making and AI-CDS systems (chapter 6), focussing especially on the extent to which the different laws within this framework are concerned with non-discrimination. Chapter 7 then examines more closely the provisions in the relevant framework that can be interpreted as necessitating a pre-deployment discrimination assessment. Hence, it highlights the specific legal contexts in which the contributions from this thesis might be utilised. It also discusses the implications of the different types of pre-deployment discrimination requirements that are identified, in relation to the decision on whether to deploy an AI-CDS system.

Part IV does the heaviest lifting in terms of pursuing the objective of the thesis. It analyses the prohibitions on direct and indirect discrimination in EU law and adapts them as needed to develop methodological elements (considerations, principles, criteria, and methods) for the assessment of discrimination in an AI-CDS system before deployment. This development is based on legal interpretation informed by multidisciplinary knowledge.[23]

---

[22] Chapter 2.

[23] The method of development is further described in section 2.2.

Part V concludes by summarising the main findings of the thesis, discussing practical implications of the proposed methodological elements, and pointing out directions for future research.

1.4 Contributions of the Thesis in Relation to Prior Legal Research and Societal Discourse

1.4.1 Contributions in the Light of Prior Legal Research

Bias and discrimination in AI systems – often referred to as 'algorithmic discrimination' – has garnered increased attention in legal academic literature in the time during and prior to the writing of this thesis. Particularly, there exists a considerable body of literature pertaining to the US legal system and US anti-discrimination law.[24] Although European legal literature on this subject initially lagged behind its US counterpart, the literature discussing AI bias and discrimination in a European legal context has caught up in recent years. Several research papers have discussed the application of EU non-discrimination law to biased AI systems,[25]

---

[24] Favaretto suggests that the interest for the topic in US legal scholarship can be attributed particularly to a 2014 White House report: Favaretto, De Clercq, and Elger (2019).: 19; Executive Office of the President and Podesta (2014); Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact Essay," *California Law Review* 104, no. 3 (June 2016), https://doi.org/10.15779/Z38BG31; Kim (2016); Mark MacCarthy, "Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms," *Cumberland Law Review* 48, no. 1 (2017); Bornstein (2018); Cofone (2018); Jon Kleinberg et al., "Discrimination in the Age of Algorithms," *Journal of Legal Analysis* 10 (2018), https://doi.org/10.1093/jla/laz001; McKenzie Raub, "Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices," *Arkansas Law Review* 71, no. 2 (2018); Matthew U Scherer, Allan G King, and Marko J Mrkonich, "Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms," *South Carolina Law Review* 71, no. 2 (2019); Lydia X.Z. Brown et al., *Challenging the Use of Algorithm-Driven Decision-Making in Benefits Determinations for People with Disabilities*, Center for Democracy and Technology (October 2020), https://cdt.org/insights/report-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/; Thomas B Nachbar, "Algorithmic Fairness, Algorithmic Discrimination," *Florida State University Law Review* 48, no. 2 (2020); Pauline T Kim, "Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action," *California Law Review* 110, no. 5 (October 2022), https://doi.org/10.15779/Z387P8TF1W.

[25] Dalenberg (2018); Hacker (2018); Allen and Masters (2020); Frederik Zuiderveen Borgesius, "Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law," *European Business Law Review* 31, no. 3 (2020); Sandra Wachter, "Affinity Profiling and

while some contributions have discussed biased AI in relation to the right to non-discrimination in Article 14 ECHR.[26]

Based on the existing EU law-oriented literature, it is common ground that the use of AI systems in decision-making processes generally has the potential to result in discrimination in violation of non-discrimination law. Moreover, there is a recurring concern that non-discrimination law might not be able to function according to its intentions when faced with biased AI systems, due to the anticipated challenges associated with enforcing non-

---

Discrimination by Association in Online Behavioural Advertising," *Berkeley Technology Law Journal* 35, no. 2 (2020), https://doi.org/10.15779/Z38JS9H82M; Raphaële Xenidis, "Tuning Eu Equality Law to Algorithmic Discrimination: Three Pathways to Resilience," *Maastricht Journal of European and Comparative Law* 27, no. 6 (2020), https://doi.org/10.1177/1023263X20982173; Ljupcho Grozdanovski, "In Search of Effectiveness and Fairness in Proving Algorithmic Discrimination in Eu Law," *Common Market Law Review* 58, no. 1 (2021); Anastasiya Kiseleva and Paul Quinn, "Are You AI's Favorite? Eu Legal Implications of Biased AI Systems in Clinical Genetics and Genomics," *European Pharmaceutical Law Review* 5, no. 4 (2021), https://doi.org/10.21552/eplr/2021/4/4; Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Bias Preservation in Machine Learning: The Legality of Fairness Metrics under Eu Non-Discrimination Law," *West Virginia Law Review, Forthcoming* (2021); Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Why Fairness Cannot Be Automated: Bridging the Gap between Eu Non-Discrimination Law and AI," *Computer Law & Security Review* 41 (2021), https://doi.org/10.1016/j.clsr.2021.105567; Pablo Martínez-Ramil, "Discriminatory Algorithms. A Proportionate Means of Achieving a Legitimate Aim?," *Journal of Ethics and Legal Technologies* 4, no. 1 (2022); Daniel Vale, Ali El-Sharif, and Muhammed Ali, "Explainable Artificial Intelligence (Xai) Post-Hoc Explainability Methods: Risks and Limitations in Non-Discrimination Law," *AI and Ethics*, no. 2 (2022), https://doi.org/10.1007/s43681-022-00142-y; Malwina Anna Wójcik, "Algorithmic Discrimination in Health Care: An Eu Law Perspective," *Health and human rights* 24, no. 1 (June 2022); Jeremias Adams-Prassl, Reuben Binns, and Aislinn Kelly-Lyth, "Directly Discriminatory Algorithms," *The Modern Law Review* 86, no. 1 (2023), https://doi.org/10.1111/1468-2230.12759.

[26] Frederik J. Zuiderveen Borgesius, "Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence," research-article, *The International Journal of Human Rights* 24, no. 10 (2020), https://doi.org/10.1080/13642987.2020.1743976; Janneke Gerards and Frederik Zuiderveen Borgesius, "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence Articles and Essays," *Colo. Tech. L.J.* 20 (2022); Audrey Lebret, "Allocating Organs through Algorithms and Equitable Access to Transplantation—a European Human Rights Law Approach," *Journal of Law and the Biosciences* 10, no. 1 (2023), https://doi.org/10.1093/jlb/lsad004.

discrimination law in this context.[27] These anticipated challenges stem from the fact that non-discrimination law is created with 'human' discrimination in mind and the belief that algorithmic discrimination is more complex, subtle, and less easily discernible compared to cases where AI is not involved.[28] While some prior works have discussed the application of EU law to biased AI systems in specific contexts such as employment,[29] advertisement,[30] and the pricing of services,[31] the lion's share of the existing literature discusses AI bias in a more general manner. Building on the existing academic literature in the field, this thesis contributes by exploring the issues of bias and discrimination specifically in relation to AI-CDS systems, thus providing sector- and application-specific insights.

This thesis concentrates on issues relating to the *assessment* of discrimination in AI-CDS systems and, thus, the application of the non-discrimination principle *before a system is deployed*. Consequently, the thesis emphasises a different context than most prior works that have addressed the application of non-discrimination law to AI-based decision-making. However, Wachter, Mittelstadt, and Russell have authored a scholarly contribution that does examine the application of EU non-discrimination law to AI systems before deployment.[32] In contrast to the present thesis, their article primarily addresses the feasibility of 'automating' the non-discrimination principle by building this principle into an AI system, thereby ensuring non-discrimination automatically.[33] The authors argue that such automation is challenging due to the contextual nature of EU non-discrimination law.[34] However, they also highlight the need to develop "a more coherent and consistent set of assessment procedures" for assessing discrimination in AI systems.[35] The present thesis focusses on the *assessment* of an AI system rather than the automation of non-discrimination law as a matter of system design. The thesis

---

[27] Borgesius (2020 A); Grozdanovski (2021). Enforcement challenges in non-discrimination law are further discussed in section 7.2.

[28] Hacker (2018) 1148-11459.

[29] Martínez-Ramil (2022).

[30] Dalenberg (2018); Wachter (2020).

[31] Borgesius (2020 B).

[32] Wachter, Mittelstadt, and Russell (2021 B).

[33] Wachter, Mittelstadt, and Russell (2021 B) 1.

[34] Wachter, Mittelstadt, and Russell (2021 B) 3.

[35] Wachter, Mittelstadt, and Russell (2021 B) 4.

is concerned with how the non-discrimination principle can be applied as part of an assessment aimed at determining whether an AI-CDS system should be deployed.[36]

Given that the application of EU non-discrimination law is a matter of contextual judicial interpretation,[37] this thesis asserts the importance of developing an assessment methodology that properly incorporates the non-discrimination principle while considering the specific circumstances in which an AI system is intended to be used. The need to develop an assessment methodology that accommodates the specific circumstances of use, is the reason why this thesis particularly considers four types of clinical decisions.[38] By taking this approach, the thesis provides insights into types and sources of bias occurring in a particular context, as well as the application of EU non-discrimination law in this context. Considering the high level of investment in research and development of AI for the health sector, the application of EU non-discrimination law to AI-CDS systems is currently an understudied issue.

1.4.2 Contributions in the Light of Broader Societal Discourse and Technical Literature

Given that biases are assumed to be inherent in many AI systems, the importance of determining an acceptable level of bias in AI systems, as well as how to scrutinise AI systems for biases, has become recognised in the broader societal discourse on AI technologies.[39] What is 'acceptable' is a question that can be addressed from many different perspectives, applying diverse criteria. For example, in non-legal literature, Landers and Berhend propose a

---

[36] This attention towards assessment rather than development of an AI-CDS system also distinguishes the perspective of the thesis from the Norwegian Equality and Anti-Discrimination Ombud's guidance on embedded protection against discrimination *(innebygd diskrimineringsvern*), which is further discussed in section 7.5.2.

[37] Chapter 2 further elaborates methodical aspects of applying non-discrimination law to pursue the objective of this thesis.

[38] Section 1.7.

[39] Wachter, Mittelstadt, and Russell (2021 B) 26 (asking "how much bias is acceptable, and where a threshold should be set for illegal disparity"); Mirja Mittermaier, Marium M. Raza, and Joseph C. Kvedar, "Bias in AI-Based Models for Medical Applications: Challenges and Mitigation Strategies," *NPJ Digital Medicine* 6, no. 113 (2023/06/14 2023): 1, https://doi.org/10.1038/s41746-023-00858-z. ("Looking into the future, one question that will most definitely arise is what level of bias is acceptable for an AI algorithm").

"framework for evaluating fairness and bias in high stakes AI predictive models."[40] Moreover, Paulus and Kent propose "a practical framework for evaluating algorithmic bias and fairness in clinical decision-making and prediction in healthcare."[41] While efforts like these underscore the need for an assessment methodology or evaluation framework aimed at bias in AI-CDS systems, the aforementioned works do not rely on non-discrimination law. Arguably, in the context of deploying AI-CDS systems in the EU, there is an urgent need for assessment methodologies that are based on the non-discrimination principle in EU law. This knowledge could also enhance the broader societal discourse on the benefits and risks of AI implementation, where concern is often expressed about the potential for discrimination resulting from biases in AI systems. Such concerns are frequently grounded on an unspecified understanding of 'discrimination.'[42] Proper legal interpretation of the non-discrimination principle is not easily available, because it requires analysis of case law and contextual interpretation tailored to specific situations.[43]

Another general issue associated with the implementation of AI systems is what kinds of tests should be conducted, and what kind of information about a system should be produced, before an AI system is deployed. This question, too, has different answers depending on the perspective from which it is addressed. This thesis enhances the comprehension of the components of AI-CDS systems that must be tested and documented before deployment, given that discrimination in the systems is to be assessed. Thus, the thesis provides a type of

---

[40] Richard N Landers and Tara S Behrend, "Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models," *American Psychologist* 78, no. 1 (2023), https://doi.org/10.1037/amp0000972.

[41] Jessica K. Paulus and David M. Kent, "Predictably Unequal: Understanding and Addressing Concerns That Algorithmic Clinical Prediction May Increase Health Disparities," *NPJ Digital Medicine* 3, no. 99 (2020/07/30 2020), https://doi.org/10.1038/s41746-020-0304-9.

[42] Favaretto, De Clercq, and Elger (2019) 21.

[43] Goodman (2016) 506; Wachter, Mittelstadt, and Russell (2021 B). Moreover, Dalenberg calls the lack of clarification of the scope and meaning of 'discrimination' in EU secondary law a "defect": Dalenberg (2018) 619. While one may discuss whether it is justified to call it a "defect," it is true that the definition of 'discrimination' in EU law cannot be understood without reference to case law. Smith notes that "the law says 'do not discriminate,' but does little to help people work out what discrimination is and how not to do it": Belinda Smith, "How Might Information Bolster Anti-Discrimination Laws to Promote More Family-Friendly Workplaces?," *Journal of Industrial Relations* 56, no. 4 (2014): 553, https://doi.org/10.1177/0022185614540128.

knowledge that facilitates legally informed testing of AI-CDS systems before they are deployed. Knowledge of how AI-CDS systems should be tested and assessed could also enable developers to make non-discriminatory AI systems from the beginning. Hence, while the thesis primarily emphasises *assessment*, the insights produced here may also have implications for the various stages of *developing* an AI-CDS system.[44]

In technical literature on biases in AI systems, the issue of potentially discriminatory biases in is often discussed in relation to various technical 'fairness metrics,' i.e., mathematical formula according to which the distribution of outputs from an AI system can be deemed 'fair.' Although fairness metrics for AI systems are often seen as contributing to non-discrimination, their compatibility with legal definitions of discrimination or the legislative aims of non-discrimination law are rarely discussed. This has been pointed out by MacCarthy, who links technical fairness metrics with US anti-discrimination law.[45] The relationship between technical fairness metrics and EU non-discrimination law has been addressed by Wachter, Mittelstadt and Russell,[46] who focus on the legislative aims underpinning EU non-discrimination law.[47] They argue that certain fairness metrics are more reconcilable than others with the notion of 'substantive equality.'[48] This thesis focuses on applying the non-discrimination principle – particularly, the prohibitions on direct and indirect discrimination. The methodological elements of discrimination assessment that this thesis develops are based on these rules and the CJEU's jurisprudence in relation to these rules. The relationship between these rules and various fairness metrics proposed in technical literature is not considered. This relationship and, particularly, how to apply fairness metrics in a pre-deployment assessment context, are worth exploring further. However, a proper analysis would necessitate an interdisciplinary effort beyond the scope of this project. Nevertheless, by developing methodological elements of discrimination assessments aimed at AI-CDS

---

[44] An overview of the process of developing an AI system is provided in section 1.5.9.

[45] MacCarthy (2017).

[46] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Bias Preservation in Machine Learning: The Legality of Fairness Metrics under Eu Non-Discrimination Law," *West Virginia Law Review* 123, no. 3 (Spring 2021).

[47] Wachter, Mittelstadt, and Russell (2021 B) 747, etc.

48 Wachter, Mittelstadt, and Russell (2021 B) 753-754, and 782. The notion of substantive equality is further discussed in chapter 4.

systems, the thesis provides a frame within which technical fairness metrics can perhaps be further developed and evaluated in the future.

1.4.3 Related Issues Outside of Scope

There are several topics related to bias and discrimination in AI systems that have received attention in recent scholarship, which are not addressed in this thesis. For instance, existing literature contends that the scope of application of EU law is too limited to properly protect various groups that may be affected by bias in AI systems and are deserving of protection.[49] EU non-discrimination law only applies to discrimination based on certain characteristics. Moreover, it is argued in the literature that the law offers weak protection against intersectional discrimination (i.e., where someone is discriminated against based on a combination of protected characteristics).[50] It has also been argued that medical AI potentially disadvantages patients with rare medical conditions and that these patients do not have proper legal protection.[51] This thesis does not discuss these issues, which are related to how protected groups should be defined, which characteristics should be protected, or whether new groups deserve protection from algorithmic discrimination. This thesis takes for granted the current definition of discrimination in EU law, and the analyses remain within the scope of currently protected characteristics. While this thesis contemplates discrimination based on sex or ethnicity, rather than other protected characteristics, its primary interest lies in the methodological elements of assessing discrimination in AI-CDS systems based on EU law. These elements are relevant to the application of the non-discrimination principle regardless of the protected characteristic at issue and can be applied to new groups that may become protected by EU non-discrimination law in the future.

Moreover, due to the focus on assessing discrimination in a pre-deployment context, there are certain topics often discussed in treatises in the field of non-discrimination law that are not

---

[49] Mann and Matzner (2019); Allen and Masters (2020); Borgesius (2020 A) Gerards and Borgesius (2022); Wachter (2020); Xenidis (2020).. However, Allen & Dee Masters are more optimistic about EU non-discrimination law generally being “well-equipped to challenge discriminatory technology”: Allen and Masters (2020) 598.

[50] Xenidis (2020).

[51] Fruzsina Molnár-Gábor, "Artificial Intelligence in Healthcare: Doctors, Patients and Liabilities," in *Regulating Artificial Intelligence*, ed. Thomas Wischmeyer and Timo Rademacher (Springer Nature Switzerland AG, 2020), 345.

included in this thesis. Notably, in relation to ex post litigation contexts, the rules concerning the reversed burden of proof are often discussed in scholarly works.[52] The reversed burden of proof entails that a claimant alleging that discrimination has occurred only needs to establish "facts from which it may be presumed that there has been direct or indirect discrimination."[53] As long as such a presumption of discrimination is established, the burden of proof shifts to the respondent, who gets the chance to rebut the presumption by proving that there is no discrimination. When discrimination is assessed in a pre-deployment setting, the burden of proof and other evidentiary rules applicable to ex post enforcement contexts may not be relevant.

1.5 AI-Based Clinical Decision Support Systems: Technology and Terminology

1.5.1 Introduction

This section provides a brief introduction to AI technology and explains certain technical terms that are used throughout the thesis. The most important terms are explained in some detail in the following sections.

1.5.2 Artificial Intelligence (AI)

Above all, the technologies addressed by this thesis are based on machine learning (ML), which is normally framed as a form of artificial intelligence (AI). AI is in fact an umbrella term that covers a range of technologies, depending on how one defines AI. Definitions of AI vary in scope, and it is not important for the purposes of this thesis to define the exact boundaries of the term. Most contemporary definitions of AI include the type of systems addressed in this thesis, which are computer programs that utilise models developed through machine learning or similar techniques. These systems are capable of performing certain tasks

[52] Kristin Henrard, "The Effective Protection against Discrimination and the Burden of Proof: Evaluating the Cjeu's Guidance through the Lens of Race," in *Eu Anti-Discrimination Law Beyond Gender*, ed. Uladzislau Belavusau and Kristin Henrard (Bloomsbury Collections: Hart Publishing, 2019), 96.

[53] Article 8 RED; Article 9 GSED.

based on automated interpretation of input information.[54] For instance, the World Health Organization explains the term AI as follows:

> "Artificial intelligence" generally refers to the performance by computer programs of tasks that are commonly associated with intelligent beings. The basis of AI is algorithms, which are translated into computer code that carries instructions for rapid analysis and transformation of data into conclusions, information or other outputs. Enormous quantities of data and the capacity to analyse such data rapidly fuel AI.[55]

The forthcoming AI Act will provide the first legal definition of AI in EU law. The definition set forth in the initial proposal from the European Commission refers to certain techniques, including machine learning, for developing AI systems. Whether the final definition will refer to certain techniques and if so, which techniques, remains to be seen. Based on the negotiation positions of the three legislative bodies of the EU before commencing the tripartite negotiations, it seems likely that the definition of an AI system will include the following components:[56]

- the ability to engage with its environment and with specific input information;
- the ability to pursue an objective;

---

[54] A glossary provided by Gartner defines AI as the application of "advanced analysis and logic-based techniques, including machine learning, to interpret events, support and automate decisions, and take actions": Gartner Glossary, s.v. "Artificial Intelligence (AI)," accessed November 6, 2023, https://www.gartner.com/en/information-technology/glossary/artificial-intelligence; A similar but lengthier definition is provided by the European Union's High-Level Expert Group on Artificial Intelligence (AI-HLEG): "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.": High-Level Expert Group on Artificial Intelligence, *A Definition of AI: Main Capabilities and Disciplines* (European Commission, 8 April 2019).

[55] World Health Organization, "Ethics and Governance of Artificial Intelligence for Health: WHO Guidance" (2021): 4.

[56] Article 3(1) AIA (EC); Article 3(1) AIA (EP).

- the ability of generating outputs consisting of various content, predictions, recommendations, decisions, etc.;
- a certain degree of autonomy in the generation of outputs.

What exactly the reference to 'autonomy in the generation of outputs' means, is not obvious. However, it is likely that the inclusion of such an element of 'autonomy' in the definition stems from an algorithm's ability to learn patterns from data and act in accordance with those self-learned patterns, rather than being explicitly programmed to handle a specific input in a certain way. Technically speaking, the autonomy of machine learning algorithms lies more in the learning process than in the generation of outputs. However, there is no doubt that systems that are developed based on machine learning techniques are meant to be covered by the AIA's definition of AI.[57] Besides, if a higher degree of autonomy in the generation of outputs were required, there would not currently exist any technologies to which the AIA would be applicable. If the requirement of autonomy in the generation of outputs ends up in an adopted version of the AIA, it must therefore be interpreted rather leniently, to stand any chance of achieving its objectives as a regulation.

What sets AI apart from regular (hard-coded) computer software is arguably its ability to generate meaningful outputs in pursuit of an objective, even without being explicitly programmed to handle a specific environment. A traditional software program consists of code that dictates which actions to take when certain inputs are encountered. Therefore, the distinction between AI and traditional software can be pinpointed to AI's ability to act based on patterns learned from data and its ability to pursue principles or objectives rather than rigid rules.[58]

---

[57] European Commission, *Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Brussels, 21 April 2021).

[58] Tobias Fahse, Viktoria Huber, and Benjamin van Giffen, "Managing Bias in Machine Learning Projects," *Innovation Through Information Systems* II (2021), 95-96; Cennydd Bowles, *What Is AI? 2*, podcast audio, Machine Ethics podcast, 24:30 at 6: "Artificial intelligence is technology which is able to take decisions on the basis of principles rather than rules." The vision of a technology displaying this capability dates back to early attempts of teaching computers to play the game of checkers, see Arthur L Samuel, "Some Studies in Machine Learning Using the Game of Checkers,"

1.5.3 Machine Learning, Algorithms, Models and Training Data

'Machine learning' (ML) refers to a set of techniques in which a computer program discovers patterns in data using a machine learning 'algorithm,' These machine learning algorithms are software-implemented procedures designed to generalise the patterns discovered in training data. 'Training data' refers to the data an algorithm is exposed to for the purposes of learning.[59] Before being exposed to data, machine learning algorithms are considered general learning procedures or learning algorithms. Once the learning algorithm is executed on a dataset, it learns patterns which it then uses to modify itself. This way, the algorithm constructs a model that can be used to classify or predict the content of unknown information or future events.

There are various techniques in machine learning. Some of these techniques involve learning from examples.[60] For instance, a learning algorithm can be exposed to the medical records of patients who have undergone spine surgery. Each record is labelled as either 'successful' or 'not successful,' The algorithm can then search for patterns and correlations between the information in the medical records and the outcomes of the surgeries.[61] This type of learning is called 'supervised learning.'

Another common category of machine learning techniques is 'unsupervised learning,' where ML algorithms discover patterns in data without the guidance of predefined labels. Unsupervised learning is especially useful for finding patterns in data when one does not

---

IBM Journal of research and development 3, no. 3 (1959). AI sometimes also refer to attempts at reproducing a level and type of intelligence that specifically resembles human intelligence – often labelled as AGI ('Artificial General Intelligence'): Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence* (Knopf Publishing Group, 2017). In some contexts, AI may refer to systems that generally surpasses human intelligence ('superintelligence'): e.g., Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, 1st ed. (Oxford University Press, 3 September, 2014).

[59] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning* (fairmlbook.org, 2019), Introduction, 5. http://www.fairmlbook.org; Matteo Pasquinelli, "How a Machine Learns and Fails," *Spheres: Journal for Digital Cultures*, no. 5 (2019): 5.

[60] Tom M Mitchell, *Machine Learning*, vol. 1 (McGraw-hill New York, March 1, 1997), 2; Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (New York, NY: Springer New York, 2017), 29.

[61] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of Machine Learning*, 2nd ed. (Cambridge, Massachusetts: The MIT Press, 2018), 9.

know what one is looking for. For example, if applied to a dataset of patients with cardiovascular disease, unsupervised learning can be used to identify subgroups of those patients.[62] These subgroups may reveal differences in response to certain types of treatments. There is also a middle ground between supervised and unsupervised learning, known as 'semi-supervised learning.' This approach combines elements of both supervised and unsupervised learning techniques.

'Reinforcement learning' refers to a learning algorithm that engages with an environment through a trial-and-error process. This involves simulating rewards for 'good' outputs and penalties for 'bad' outputs based on the desired objective set by the developer.[63] While reinforcement learning is not the most commonly used type of algorithm in medicine, it can be used to create models intended for treatment selection based on individual patient characteristics.[64]

Finally, there is 'deep learning,' which is further explained in the next section.

Within these categories of ML techniques, there are various specific types of algorithms that developers can use. In practice, developers often try to build models using a few different algorithms, to determine which one gives the best results. AI-CDS systems can incorporate any type of algorithm. The specific algorithm used in an AI-CDS system does not significantly impact the analyses conducted in this thesis.[65]

As mentioned, the data used by learning algorithms to find patterns is called 'training data.' The process of running a learning algorithm on training data is known as 'training.' The

---

[62] Kipp W Johnson et al., "Artificial Intelligence in Cardiology," *Journal of the American College of Cardiology* 71, no. 23 (2018): 2677, https://doi.org/https://doi.org/10.1016/j.jacc.2018.03.521; On identifying subgroups in an intensive care unit: Kelly C. Vranas et al., "Identifying Distinct Subgroups of Icu Patients: A Machine Learning Approach*," *Critical Care Medicine* 45, no. 10 (2017), https://doi.org/10.1097/ccm.0000000000002548.

[63] Oliver Theobald, *Machine Learning for Absolute Beginners: A Plain English Introduction*, 3rd edition, Kindle edition ed. (Scatterplot Press, 31 December, 2020), 15, 16, and 24; Johnson et al. (2018) 2677.

[64] Johnson et al. (2018) 2678.

[65] For examples of medical applications of different types of algorithms, see Johnson et al. (2018) 2673.

objective of the training process is for the algorithm to continuously adjust itself to improve its accuracy in performing the task it is being trained for. During supervised learning, it relies on patterns found between the input data it observes and the labels that indicate the correct output. For example, the algorithm may be given data from patients who have been diagnosed with COVID-19 as well as data from healthy patients. During training, the algorithm examines the data concerning each patient and the labels of 'COVID-19' or 'healthy.' ML algorithms autonomously update themselves as they learn from the data they are exposed to. This means that the model they rely on, e.g., for diagnosing patients, is modified with each training iteration.[66] What occurs can be illustrated by comparison with human learning and cognitive reasoning. One could say that the algorithm modifies its way of 'reasoning' through the process of learning. However, in practice, the generation of the model is the result of mathematical calculation relying on statistical inference.[67] The algorithm infers patterns in the data it observes.[68] As it learns more about the patterns inferred from training data, the algorithm automatically modifies the weight it assigns to the features observed in the data.[69]

The learning process produces an updated algorithm – an algorithm that is no longer a general learning algorithm, as it now reflects the specific data used for training.[70] Such a trained,

---

[66] Brent Daniel Mittelstadt et al., "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* 3, no. 2 (2016): 3, https://doi.org/10.1177/2053951716679679.

[67] Pasquinelli (2019) 10.

[68] Houssem Ben Braiek and Foutse Khomh, "On Testing Machine Learning Programs," *Journal of Systems and Software* 164, no. 110542 (2020): 1, https://doi.org/https://doi.org/10.1016/j.jss.2020.110542.

[69] Steven M. Appel and Cary Coglianese, "Algorithmic Governance and Administrative Law," in *Camrbdige Handbook on the Law of Algorithms*, ed. Woodrow Barfield (Cambridge University Press, 2020), 164. ("Rather than humans making key specifications, the algorithms automatically select the "best" variables (and variable combinations) and the "best" mathematical functional form that analyses the variables – with "best" defined in terms of which variable combinations and functional forms optimize an end goal or "objective function" that has been defined by the human analyst.")

[70] Peter Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data* (Cambridge University Press, 2012), 13.

specific algorithm is often called a 'model,'[71] and this is how the term 'model' is used throughout this thesis. The model is the result of inferences from the training data; it is the accumulated set of relationships that are discovered during training,[72] and it constitutes a recipe for producing an output based on those relationships.[73] Any output that the model generates is determined by the contents of training data and input data.[74] Therefore, the properties of the training data are of paramount importance to how the model will work. The goal of the training is in practice to achieve a model that provides outputs that are as accurate as possible when applied to the inputs that will be encountered in practice.[75] It is therefore important that the training data are as representative as possible of the world in which a model will be deployed. For this reason, machine learning normally makes use of authentic, historical datasets as training data.

The training process is complete once the model reaches an accuracy level that the developer is satisfied with, after testing the model on a separate dataset.[76] A trained model can be used to classify or predict the content of new data (input), i.e. data which the learning algorithm was never exposed to during training. It can be implemented into an AI-CDS system and applied to make decisions concerning new patients, in which case each new patient would represent an unknown input.

An AI-CDS system can incorporate one or more AI models. AI-CDS systems with several models might combine the outputs from each model to finally produce a singular output that is presented to the user of the system, or the system might display several outputs. To avoid overcomplicating the discussions in this thesis, it is generally assumed here that an AI-CDS

---

[71] Some authors explicitly distinguish between the learning algorithm and the model which is the result of applying the learning algorithm to training data, while others do not make this distinction: Kleinberg et al. (2018) 132-33.

[72] Barocas and Selbst (2016) 677.

[73] Theobald (2020) 19.

[74] Theobald (2020) 10.

[75] Hastie, Tibshirani, and Friedman (2017) 29.

[76] Pasquinelli (2019) 8; Elizamary de Souza Nascimento et al., "Understanding Development Process of Machine Learning Systems: Challenges and Solutions" (paper presented at the 2019 ACM/IEEE international symposium on empirical software engineering and measurement (ESEM), 2019), 3.

system produces a singular output, such as one binary classification to diagnose the patient or one numerical value indicating the probability of successful surgery. In practice, when assessing discrimination in an AI-CDS system that generates multiple outputs, discrimination must be assessed in each individual model within the system. Despite this, the essential methodological elements of the assessment remain the same.

1.5.4 Deep Learning and Neural Networks

'Deep learning' is a category of ML techniques characterised by complex algorithms capable of learning from large amounts of raw data with minimal human guidance or pre-processing.[77] Deep learning may be supervised or unsupervised.[78] The algorithms used in deep learning are called 'neural networks.' Structurally and conceptually, the idea of neural networks is inspired by the human brain and its ability to learn from experience.[79] For the purposes of this thesis, a basic and simplified introduction to deep learning and neural networks is considered adequate.

A neural network has several layers of processing units called 'nodes' or 'neurons' (interchangeably) which are autonomously modified by the neural network itself during training. The layers are the reason why deep learning techniques are called 'deep.'[80] There are at least three layers of nodes. The first layer is the 'input layer,' where input data is initially received and processed. The nodes in the input layer have sensors allowing them to receive

---

[77] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521 (28 May 2015): 436, https://doi.org/10.1038/nature14539; Curtis E. A. Karnow, "The Opinion of Machines," in *The Cambridge Handbook of the Law of Algorithms*, ed. Woodrow Barfield, Cambridge Law Handbooks (Cambridge: Cambridge University Press, 2020), 20.

[78] LeCun, Bengio, and Hinton (2015) 436..

[79] Johnson et al. (2018); Hussein Abdel-Jaber et al., "A Review of Deep Learning Algorithms and Their Applications in Healthcare," *Algorithms* 15, no. 71 (2022): 1-2, https://doi.org/10.3390/a15020071; Theobald (2020); AD Dongare, RR Kharde, and Amit D Kachare, "Introduction to Artificial Neural Network," *International Journal of Engineering and Innovative Technology (IJEIT)* 2, no. 1 (July 2012): 190.

[80] Lefteris Koumakis, "Deep Learning Models in Genomics; Are We There Yet?," *Computational and Structural Biotechnology Journal* 18 (2020): 1467, https://doi.org/https://doi.org/10.1016/j.csbj.2020.06.017.

input data, before forwarding the information to the next layer.[81] The layer beyond the input layer is called a 'hidden layer' – not because it cannot be observed, but because its relationship with inputs and outputs are obfuscated by the autonomously adjusted nodes and channels connecting the layers ('weights' – see below).[82] There are typically several hidden layers in a neural network.

The layers in a neural network are linked through weighted connections – called 'weights'.[83] Like the nodes, the weights are autonomously modified during training. In visual representations of neural networks, the weights typically appear as lines from the nodes in one layer to the nodes in the next layer. Weights are mathematical functions that help the nodes in one layer transform the data they receive and forward it to the next layer. They determine the strength of the connections between nodes in different layers. In turn, the strength of the connection determines which nodes are activated in the next layer.[84] Consider, for example, a neural network for speech recognition. In a simplified manner, one could say that a recording of a voice saying the word "go" would activate nodes in the input layer related to sounds associated with the letters G and O. The nodes and weights are adjusted during training to recognise these letters. Once all hidden layers have had their say about the data they have received, a neural network finally produces an output when the signals from previous layers reach and are processed by the nodes in the final layer – the 'output layer.'

In medicine, deep learning techniques are particularly applied in medical image analysis and speech analysis, but the potential is undoubtedly much wider and some claim that neural networks based on deep learning could be among the most important types of AI in medicine in the future.[85] Deep learning can be applied to analyse the natural language found in medical

---

[81] Jürgen Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural networks* 61 (2015): 86, https://doi.org/http://dx.doi.org/10.1016/j.neunet.2014.09.003.

[82] Barry Solaiman and Mark G Bloom, "AI, Explainability, and Safeguarding Patient Safety in Europe," in *The Future of Medical Device Regulation*, ed. I. Glenn Cohen et al. (Cambridge University Press, 2022), 93.

[83] Dongare, Kharde, and Kachare (2012) 189.

[84] Dongare, Kharde, and Kachare (2012) 189; Solaiman and Bloom (2022) 93.

[85] Johnson et al. (2018) 2676. ("Clinicians should understand that deep learning models are quickly becoming the state-of-the-art method and will enable the coming future applications of AI".)

records, as well as datasets that contain mixed types of data.[86] Neural networks developed from such data can serve as decision support, such as predicting a patient's risk of developing certain diseases.[87] One advantage of deep learning systems is that they can achieve higher levels of accuracy in contexts where the input data contains many features.[88] For example, electronic health records often include free text notes from doctors and nurses, resulting a vast number of feature variables. Due to the high complexity of this unstructured data, advanced algorithms are necessary to utilise the information.

Issues of bias and potential discrimination can arise in relation to AI systems based on deep learning, similar to how these issues may arise in AI systems based on other ML techniques[89] Therefore, this thesis does not distinguish categorically between neural networks and other ML-based models. However, while AI-CDS systems based on deep learning are not fundamentally different from other AI-CDS systems, deep learning systems are associated with certain particular challenges. First and foremost, it is important to note that deep learning produces more complex and opaque decision-making models. This complexity arises from the presence of hidden layers within a neural network, which are adjusted automatically during the training process. As a result, the relationship between inputs and outputs, which is non-linear in nature, becomes obscured and difficult for a human observer to comprehend.[90] Moreover, the lack of human control with how the hidden layers adjust themselves during training is likely to render neural networks more unpredictable than other AI-CDS systems.

1.5.5 Classifications and Regressions

At the level of trained models, there is one distinction that should be noted because it is important in certain parts of this thesis. In the context of AI-CDS systems, two types of

---

[86] This is called 'mixed modality data' and 'multimodal machine learning: Alvin Rajkomar et al., "Scalable and Accurate Deep Learning with Electronic Health Records," *NPJ Digital Medicine*, no. 18 (2018): 1, https://doi.org/10.1038/s41746-018-0029-1; Kline et al. (2022) 1.

[87] e.g., Davide Placido et al., "A Deep Learning Algorithm to Predict Risk of Pancreatic Cancer from Disease Trajectories," *Nature Medicine* 29, no. 5 (2023), https://doi.org/10.1038/s41591-023-02332-5.

[88] Rajkomar et al. (2018) 1.

[89] Barocas, Hardt, and Narayanan (2019) Appendix: Technical Background, 5-6.

[90] Hildebrandt (2021) 6.

outputs are of particular importance: classifications and regressions.[91] A classification model is a model that predicts which class the input data belongs to. In other words, it categorises input into predefined classes. The classification is sometimes binary, such as when the model predicts whether a patient is suffering from an ischemic stroke or not. In other situations, there may be multiple classes. For example, the model predicts whether a patient would benefit more from treatment A, B or C. In practice, the output from a classification model is often presented as the probability that the input data belongs to each one of the different class labels.[92]

The term 'regression model' is used in this thesis to denote a model that outputs a numerical value on a continuous scale – a regression.[93] This category includes, for example, a model that predicts how much pain a patient would experience during an invasive procedure, represented as a value between 1 and 10.

1.5.6 Generative AI and Large Language Models

Following the release of OpenAI's famous conversational agent, ChatGPT, in 2022, the terms 'generative AI' and 'Large Language Models' (LLMs) suddenly leapt to the forefront of the legal and societal discourse on AI systems.[94] ChatGPT, which is an LLM trained on a very large body of text, displays a level of conversational capability that was previously unheard of in an AI system. Similar capabilities have since been demonstrated by several AI-based conversational agents. When prompted with an input, current LLMs can respond in a style that often perfectly mimics natural conversational language used by humans. These capabilities are enabled by recent advances in Natural Language Processing (NLP), which is a

---

[91] To avoid confusion, it is worth noting that there is an algorithm called 'logistic regression' which, despite the name, is used for classification tasks. Linear regression is, in contrast, one of many algorithms that are used for tasks that require a numerical output along a continuous scale: Jason Brownlee, "Difference between Classification and Regression in Machine Learning," *Machine Learning Mastery*, 22 May, 2019, https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/.

[92] Brownlee (2019).

[93] Hastie, Tibshirani, and Friedman (2017) 9-10.

[94] See, for example, the literature review conducted by Sallam: Malik Sallam, "ChatGPT Utility in Health Care Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns," *Healthcare* 11, no. 887 (2023): 5-11, https://doi.org/doi.org/10.3390/healthcare11060887.

family of techniques for analysis, learning, comprehension and production of natural language,[95] i.e., the language that humans use when they communicate with each other.

Many NLP techniques are based on machine learning and/or deep learning.[96] LLMs, specifically, are based on deep learning techniques. While NLP and LLMs are specific to AI systems intended to analyse and generate text or speech, the term 'generative AI' refers more generally to AI systems that generate content based on the patterns they have learned during training. While conversational agents generate text, generative AI can also generate art, music, images, drawings, speech, video, and more.

There are several possible applications of generative AI and LLMs in healthcare,[97] including the utilisation of such models in AI-CDS systems.[98] LLMs could rapidly process new information provided by a patient during a medical consultation and simultaneously take into account existing information in the patient's health record as well as scientific research or other sources of medical information.[99] However, there is currently little research on the use

---

[95] Neguine Rezaii, Phillip Wolff, and Bruce H Price, "Natural Language Processing in Psychiatry: The Promises and Perils of a Transformative Approach," *The British Journal of Psychiatry*, no. 220 (2022): 251, https://doi.org/10.1192/bjp.2021.188.

[96] Uday Kamath, John Liu, and James Whitaker, *Deep Learning for Nlp and Speech Recognition* (Cham, Switzerland: Springer Nature Switzerland AG, 2019), 11-15; Tom Young et al., "Recent Trends in Deep Learning Based Natural Language Processing," *IEEE Computational Intelligence Magazine* 13, no. 3 (2018), https://doi.org/10.1109/MCI.2018.2840738.

[97] Mathias Karlsen Hauglid and Tobias Mahler, "Doctor Chatbot: The Eu's Regulatory Prescription for Generative Medical AI," *Oslo Law Review* 10, no. 1 (2023), https://doi.org/10.18261/olr.10.1.1; Catherine Diaz-Asper et al., "A Framework for Language Technologies in Behavioral Research and Clinical Applications: Ethical Challenges, Implications and Solutions (Preprint)," https://doi.org/10.1037/amp0001195; Peter Lee, Sebastien Bubeck, and Joseph Petro, "Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine," *New England Journal of Medicine* 388, no. 13 (2023): 1234-35.

[98] e.g., Mathias K Hauglid, "What's That Noise? Interpreting Algorithmic Interpretation of Human Speech as a Legal and Ethical Challenge," *Schizophrenia Bulletin* 48, no. 5 (2022), https://doi.org/10.1093/schbul/sbac008; Felipe C Kitamura, "ChatGPT Is Shaping the Future of Medical Writing but Still Requires Human Judgment," *Radiology* 307, no. 2 (2023), https://doi.org/https://doi.org/10.1148/radiol.230171; Diaz-Asper et al. (2023).

[99] Peter Lee, Carey Goldberg, and Isaac Kohane, *The AI Revolution in Medicine: GPT-4 and Beyond* (Pearson, 2023).

of LLMs as clinical decision support and there is a lot of uncertainty surrounding the reliability of current LLMs and how to assess their performance.[100] AI systems trained on natural language data are known to display biases that are ingrained in language.[101]

This thesis primarily refers to AI systems producing pre-defined classifications or regressions rather than generative outputs. The analyses and discussions in this thesis are nonetheless relevant to LLMs. In many regards, LLMs for clinical decision support are comparable to other AI-CDS systems. For instance, even though an LLM can present a diagnosis or treatment recommendation in a natural language, the most important output is the diagnosis or treatment recommendation itself. An assessment of discrimination in an LLM should largely apply the same methodologies as discrimination assessments of other AI-CDS systems. However, considering the unique nature of LLMs, future research should nonetheless look more specifically and comprehensively into how to assess discrimination in LLMs, as this thesis primarily focuses on other types of models.

1.5.7 Feature Variables and Target Variables

A 'feature variable,' also known as an 'input variable' or 'independent variable,' represents a specific measurable or observable quantity used as an input during a model's training or operation. These variables play an important role in both supervised and unsupervised learning. When an AI-CDS system is used in clinical practice, this means that the feature variables are the factors that the system will consider in the process of generating an output. For example, a model intended for the prediction of spine surgery outcomes might include the patient's age as a feature variable, in which case the patient's age must be provided as input each time the model is used. A model may include any number of feature variables. Neural networks based on deep learning typically consist of a large number of feature variables, each

---

[100] Sallam (2023) 13-14; Lee, Bubeck, and Petro (2023) 1234.

[101] Dirk Hovy and Shrimai Prabhumoye, "Five Sources of Bias in Natural Language Processing," *Language and Linguistics Compass* 15, no. 8 (2021), https://doi.org/https://doi.org/10.1111/lnc3.12432; Deven Shah, H Andrew Schwartz, and Dirk Hovy, "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview," *arXiv preprint arXiv:1912.11078* (2019); Tony Sun et al., "Mitigating Gender Bias in Natural Language Processing: Literature Review," *arXiv preprint arXiv:1906.08976* (2019); Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356, no. 6334 (2017), https://doi.org/10.1126/science.aal4230.

corresponding to a 'node.'[102] For example, in AI systems that interpret and classify medical images, each pixel in an image is considered a feature variable.

The 'target variable,' also known as the 'dependent variable' because it depends on the feature variables, is the feature that a model based on supervised learning is intended to predict. In other words, the target variable defines what kind of *output* a model should produce. For example, a target variable may be the diagnosis that a patient is most likely to have, or the number of days before a patient is fully recovered from an injury. In LLMs, the target variable is typically the sequence of words that would be most appropriate in the given context. In classification models, the target variable is a pre-defined category whereas in regression models it is a numerical value along a continuous scale.

1.5.8 AI-CDS Systems and Clinical Decision-Making

'Clinical decision support' (CDS) is an established term in medical jargon. While CDS can refer to any system or tool that can be used by health personnel when preparing for or making a clinical decision (e.g., written clinical guidelines),[103] it often refers to a computer program intended for this purpose.[104] Rule-based computer programs have been used for CDS purposes for decades, and the idea of CDS computer programs has existed much longer .[105]

---

[102] Section 1.5.4.

[103] A. T. M. Wasylewicz and A. M. J. W. Scheepers-Hoeks, "Clinical Decision Support Systems," in *Fundamentals of Clinical Data Science*, ed. Pieter Kubben, Michel Dumontier, and Andre Dekker (Cham, Switzerland: Springer Nature Switzerland AG, 2019), 153.

[104] Guidance from the Medical Device Coordination Group (MDCG) defines decision support software as "computer based tools which combine general medical information databases and algorithms with patient-specific data. They are intended to provide healthcare professionals and/or users with recommendations for diagnosis, prognosis, monitoring and treatment of individual patients.": Medical Device Coordination Group, *MDCG 2019-11 Guidance on Qualification and Classification of Software in Regulation (Eu) 2017/745 - MDR and Regulation (EU) 2017/746 - IVDR* (October 2019), section 9.

[105] Robert S Ledley and Lee B Lusted, "Reasoning Foundations of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason," *Science* 130, no. 3366 (1959); Paul D Clayton and George Hripcsak, "Decision Support in Healthcare," *International journal of bio-medical computing* 39 (1995); Dereck L Hunt et al., "Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient

The term ‘clinical decision’ refers to decisions that are taken in the course of providing healthcare to a patient. Building on Hugaas Ofstad’s doctoral thesis on medical decision-making, a clinical decision can be defined as a

> (…) statement committing to a particular course of clinically relevant action and/or statement concerning the patient’s health that carries meaning and weight because it is said by a medical expert.[106]

According to this definition, a clinical decision is a statement that comes from a medical expert and is legitimized by the decision-maker’s medical expertise. A CDS system, then, can be defined as a system that is used by a medical expert in the course of arriving at a “statement committing to a particular course of clinically relevant action and/or statement concerning the patient’s health.”[107] This includes, for example, deciding the diagnosis for a patient, deciding which treatment to recommend to the patient and deciding whether or not an action should be taken in respect of the patient. This understanding of the term ‘clinical decision’ does not encompass strictly administrative decisions, such as decisions regarding internal resource planning within an institution.[108] AI systems intended for such purposes are therefore excluded from the scope of the thesis. Moreover, the focus on clinical decision-making excludes AI systems intended to perform tasks that are not decision-making, such as surgical robots or other applications where AI is used to automate a work task.

In many health systems, shared decision making is considered the best practice for clinical decision-making, meaning that patients are encouraged and entitled to participate in decision-

---

Outcomes: A Systematic Review," *JAMA* 280, no. 15 (1998); See, with further references Malcolm A. Gleser and Morris F. Collen, "Towards Automated Medical Decisions," *Computers and Biomedical Research* 5, no. 2 (1972): 180-81, https://doi.org/https://doi.org/10.1016/0010-4809(72)90080-8; Davenport and Kalakota (2019).

[106] Eirik Hugaas Ofstad, "Medical Decisions in 372 Hospital Encounters" (University of Oslo, 2015), 50.

[107] Ibid.

[108] The Norwegian Directorate of eHealth lists some examples of how AI can be used to assist administrative decision-making in the health sector: *Forprosjekt. Utredning Om Bruk Av Kunstig Intelligens I Helsesektoren*, Norwegian Directorate of eHealth (December 2019), 82-83, https://www.ehelse.no/publikasjoner/utredning-om-bruk-av-kunstig-intelligens-i-helsesektoren.

making concerning their own health.[109] In most cases, decisions on which actions to take are ultimately decided by the patient. However, there are important limitations to a patient's decisional autonomy. A patient cannot decide to have treatment that would contradict a clinician's duty of care (e.g., a medical procedure likely to cause more harm than good). Depending on the health system and jurisdiction, budgetary confines are also likely to limit patients' freedom to choose which examinations and procedures they would like to have carried out.

The abovementioned definition of a clinical decision from Hugaas Ofstad reflects the importance of a clinician's assessment even though patients often have the final say in shared decision-making contexts. Because statements by clinicians regularly constitute the basis for a patient's choice of action, it makes sense to consider certain statements by clinicians as clinical decisions on account of their weight in determining health-related actions or conclusions.

An AI-CDS system is a computer program (software) designed to support clinical decision-making, which relies on a model developed using machine learning. AI-CDS systems can be distinguished from fully automated decision-making. With AI-CDS systems, the software produces an output intended primarily for a clinician. The clinician then decides to rely on the output or carries out further assessments before a decision is made with real implications for a patient. However, depending on the definitions one relies on, the line between 'decision support' and 'full automation' can be blurry. An AI-CDS system can produce a diagnosis without the involvement of a clinician. In other words, the diagnosis is produced in a fully automated manner. In practice, what determines whether such a diagnosis constitutes decision support or a clinical decision in and of itself, is the subsequent involvement of a clinician before the diagnosis is authoritatively communicated to the patient. One could discuss whether it should be called 'decision support' if clinicians in practice rely heavily on the outputs from an AI system without further assessment. However, for the aims of this thesis, it is not important to examine this distinction any further. The term 'decision support' is used here to refer to AI systems that produce outputs intended to support clinicians in clinical decision-making contexts. The methodological elements of assessing discrimination that this

[109] A M Stiggelbout et al., "Shared Decision Making: Really Putting Patients at the Centre of Healthcare," *BMJ* 344, no. e256 (2012): 1, https://doi.org/10.1136/bmj.e256.

thesis develops could to some extent be used to assess systems intended also for fully automated clinical decisions.[110]

To facilitate a nuanced analysis that highlights the implications of different types of clinical decisions when assessing discrimination in an AI-CDS system, this thesis primarily concentrates on a few selected clinical decision categories. These categories are introduced in section 1.7.

1.5.9 Development of an AI-CDS System: Process Overview

The following figure gives an overview of the machine learning process that leads up to the deployment of an AI-CDS system. How biases can enter an AI-CDS system at the various steps of the development process is discussed in section 4.4.

Different names can be used for each step of the process, and process descriptions can have different levels of detail. However, a typical process description includes data collection, pre-processing, training and model development, testing/validation and deployment.[111]

---

[110] Zuiderveen Borgesius notes that many discrimination-related risks are similar for fully and partly automated decisions: Zuiderveen Borgesius (2018): 8.

[111] For examples of process descriptions in the literature, see: I. Glenn Cohen et al., "The Legal and Ethical Concerns That Arise from Using Complex Predictive Analytics in Health Care," *Health Affairs* 33, no. 7 (July 2014), https://doi.org/10.1377/hlthaff.2014.0048; Jennifer Cobbe, Michelle Seng Ah Lee, and Jatinder Singh, "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021): 604, https://doi.org/10.1145/3442188.3445921; Kerstin N. Vokinger, Stefan Feuerriegel, and Aaron S. Kesselheim, "Mitigating Bias in Machine Learning for Medicine," *Communications Medicine* 1, no. 25 (2021): 2, https://doi.org/10.1038/s43856-021-00028-w; Rohan Gupta et al., "New Era of Artificial Intelligence and Machine Learning-Based Detection, Diagnosis, and Therapeutics in Parkinson's Disease," *Ageing Research Reviews* 90, no. 102013 (2023): 8, https://doi.org/10.1016/j.arr.2023.102013; de Souza Nascimento et al. (2019) 3; Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng, "How to Develop Machine Learning Models for Healthcare," *Nature Materials* 18, no. 5 (2019): 410-13, https://doi.org/10.1038/s41563-019-0345-0.

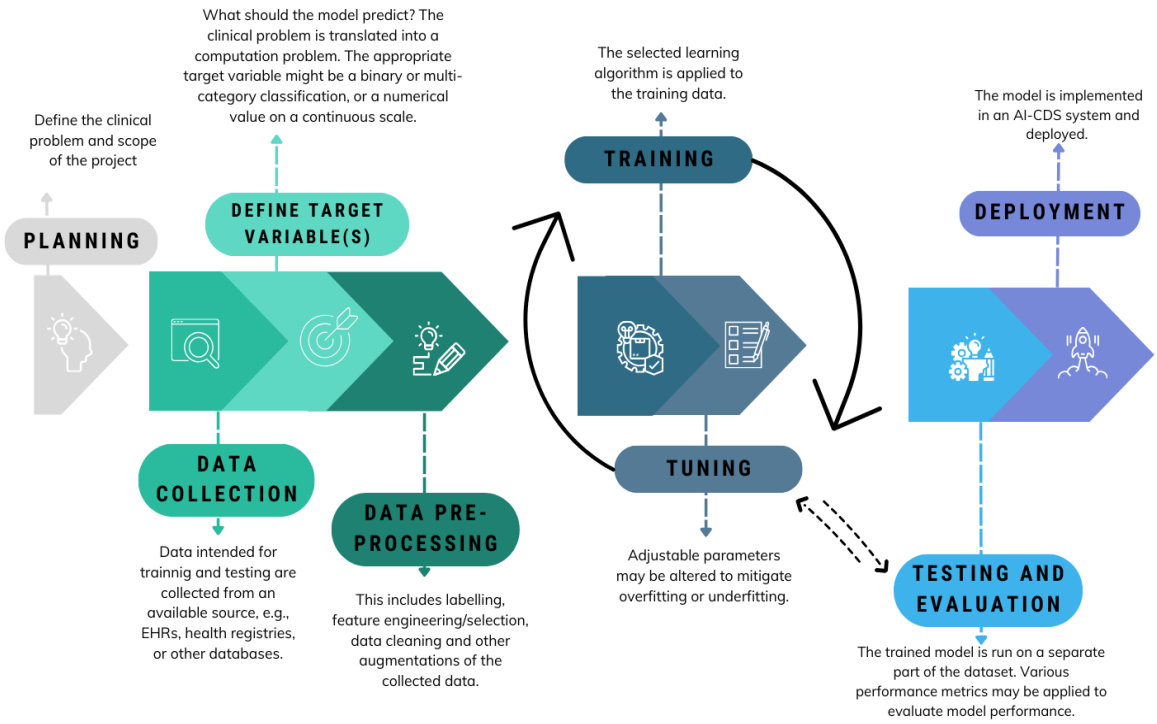


Figure 2: The figure represents the main steps of developing an AI-CDS system based on machine learning.[112]

The contributions of this thesis concern pre-deployment discrimination assessments and therefore relate directly to the component of Figure 2 titled 'Testing and evaluation.' Pre-deployment discrimination assessments can be seen as part of a broader evaluation of a model or system, where several aspects may be considered (e.g., safety and performance). Testing of models is a central aspect of such evaluations. Hence, the methodological elements of pre-deployment discrimination assessments developed in this thesis have implications for the types of tests that ought to be conducted before deployment. More precisely, the analysis in Part IV of the thesis suggests several methods that may be included in discrimination assessments, which require certain testing activities for the purpose of studying a model's behaviour. In addition to these direct implications for testing and evaluation, the thesis

[112] This figure was built using Canva and is included in the thesis in accordance with Canva's standard licensing agreement.

produces insights which may impact other stages of the development process, given that developers aim to develop models that will be assessed positively.[113]

1.5.10 After Deployment, is the Model Locked or Adaptive?

After deployment, the lifecycle of an AI-CDS system continues beyond what is represented by the figure above. A deployed system must be maintained and updated, which may involve further development of deployed models through training on new data. An AI-CDS system can even be designed to learn continuously from the data it observes during its operation.[114] Continuously learning AI systems are called 'adaptive' or 'online' systems.[115]

In practice, continuous learning raises profound safety concerns and regulatory challenges.[116] Adaptive AI-CDS systems that learn continuously are currently challenging due to the importance of controlling what such systems learn and the risks of these models learning undesirable behaviours.[117] Locked models – i.e., models that produce the same output every time they are prompted with the same input – are more predictable.[118] The fact that they are

---

[113] Section 12.2.

[114] Electromedical and Healthcare IT Industry (COCIR) European Coordination Committee of the Radiological, *COCIR Analysis on AI in Medical Device Legislation* (4 September 2020), 6, https://www.cocir.org/fileadmin/Position_Papers_2020/COCIR_Analysis_on_AI_in_medical_Device_Legislation_-_Sept._2020_-_Final_2.pdf; Mittermaier, Raza, and Kvedar (2023) 2.

[115] Kerstin Noëlle Vokinger, Thomas J Hwang, and Aaron S Kesselheim, "Lifecycle Regulation and Evaluation of Artificial Intelligence and Machine Learning-Based Medical Devices," in *The Future of Medical Device Regulation*, ed. I. Glenn Cohen et al. (Cambridge: Cambridge University Press, 2022), 14.

[116] U.S. Food & Drug Administration, *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SAMD) - Discussion Paper and Request for Feedback* (2 April 2019), 2, https://www.fda.gov/media/122535/download?attachment.

[117] The continuous learning that happens in online/adaptive algorithms is typically seen as representing a risk of unpredictability and lack of control with how the algorithm works, including whether it works in accordance with the developer's intention and documentation: Cary Coglianese and David Lehr, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era," *Georgetown Law Journal* 105, no. 5 (2016): 1167. ("… as long as learning algorithms are running, humans are not really controlling how they are combining and comparing data.")

[118] European Coordination Committee of the Radiological (2020): 8; Vokinger, Hwang, and Kesselheim (2022) 14.

locked in this sense, does not mean that they cannot be updated. Such models can be updated and receive further training from time to time, in an offline environment. Currently, locked models are more feasible than adaptive models in the context of clinical decision-making.

1.6 Promises and Concerns

1.6.1 Main Promises

Currently, there are high expectations in the field of medical science and practice regarding the many ways in which AI-CDS systems and other AI applications can enhance medical care. These expectations are so significant that the World Health Organization even emphasises the risks associated with *not* deploying AI:

> Not using the technology could result in avoidable morbidity and mortality, making it blameworthy not to use a certain AI technology, especially if the standard of care is already shifting to its use.[119]

One of the central hopes for the implementation of AI-CDS systems it that they will improve the quality of healthcare by increasing the accuracy of clinical assessments. This is particularly beneficial in situations where assessments can be improved by considering more information than human clinicians can typically digest within a given setting.[120] The ability to consider more information about each individual patient is why healthcare involving AI systems is sometimes referred to as 'personalised medicine.'[121] Once an AI-CDS has been trained, it has learned complex patterns from historical medical data, and can now apply these patterns to information about new patients, instantly.[122]

---

[119] *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, World Health Organization (Geneva, 2021), 33, https://apps.who.int/iris/bitstream/handle/10665/341996/9789240029200-eng.pdf.

[120] M. D. Samad et al., "Predicting Survival from Large echocardiography and Electronic health record Datasets: Optimization with Machine Learning," *JACC Cardiovasc Imaging* 12, no. 4 (April 2019): 685-86, https://doi.org/10.1016/j.jcmg.2018.04.026.

[121] Kline et al. (2022) 1.

[122] W. Nicholson Price, II, "Black-Box Medicine," *Harvard Journal of Law & Technology* 28, no. 2 (2015): 424.

While ML-based models “excel in the analysis of complex signals in data-rich environments,”[123] the human mind has limited capacity to process large amounts of information and understand how different pieces of information relate to each other. One can think of an AI-CDS system as having access to the collective knowledge of many clinicians, instead of relying on just the few who happen to be on duty.[124] AI-CDS systems might therefore improve certain assessments that even experienced and specialised medical professionals find challenging, such as the diagnosis of heart failure.[125] It is important to note however, that the realisation of these promises lies in the future. Current research indicates that ML-based models often do not outperform human domain experts in clinical decision tasks.[126]

Another main argument for deploying AI-CDS systems is their efficiency. AI-CDS systems process input information and produce outputs rapidly. This can potentially reduce the time required for conducting a clinical assessment. As a result, clinicians can either see a greater number of patients or allocate more time to other tasks.[127] Another example is screening programs for presumed healthy persons, such as mammography screening. While there are various arguments for and against such screening programs, budgetary considerations are

---

[123] With further references, Stephanie L. Hyland et al., "Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning," *Nature Medicine* 26, no. 3 (March 2020), https://doi.org/10.1038/s41591-020-0789-4.

[124] Thomas Grote and Philipp Berens, "On the Ethics of Algorithmic Decision-Making in Healthcare," *Journal of medical ethics* 46 (2020): 205, https://doi.org/10.1136/medethics-2019-105586.

[125] Dong-Ju Choi et al., "Artificial Intelligence for the Diagnosis of Heart Failure," *NPJ Digital Medicine* 3, no. 54 (2020), https://doi.org/10.1038/s41746-020-0261-3.

[126] Evangelia Christodoulou et al., "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models," *Journal of clinical epidemiology* 110 (2019), https://doi.org/10.1016/j.jclinepi.2019.02.004; Paolo Fusar-Poli et al., "Clinical-Learning Versus Machine-Learning for Transdiagnostic Prediction of Psychosis Onset in Individuals at-Risk," *Translational psychiatry* 9, no. 259 (2019), https://doi.org/10.1038/s41398-019-0600-9.

[127] Fatima Zulqarnain, S Fisher Rhoads, and Sana Syed, "Machine and Deep Learning in Inflammatory Bowel Disease," *Current Opinion in Gastroenterology* 39, no. 4 (July 2023), https://doi.org/https://doi.org/10.1097%2FMOG.0000000000000945.

sometimes decisive. AI-CDS systems could make it possible to screen and diagnose more patients for less cost.[128]

Moreover, their processing speed potentially makes AI-CDS systems valuable in situations where time is of the essence, such as during triage in intensive care units or when diagnosing patients with diseases for which timely treatment is critical. In these circumstances, especially where there is a lot of information to process, clinicians may struggle to identify the most important information and as a result, may be unable to make the best decision.[129]

Further benefits of AI-CDS systems relate to their consistency and scalability. Human judgement is known to vary. Not only can two different clinicians assess the same patient differently – the same clinician might arrive at different conclusions at different times. In contrast, an AI-CDS system is more consistent and, thus, less subjective.[130] If the input information concerning two patients are similar, the system is likely to assess them similarly. In addition, these consistent systems are – like all computer programs – scalable. At least in principle, the same model can be deployed in any number of healthcare institutions.[131]

Due to this consistency and scalability, there has been some debate over whether AI systems would lead to more fair and less biased decision-making, compared to strictly human decision-making.[132] However, this viewpoint has been less prominent in the recent academic discourse, in which bias and discrimination are typically considered side-effects of AI-supported decision-making.[133]

---

[128] For example, AI-based cancer colon screening has been suggested: Miguel Areia et al., "Cost-Effectiveness of Artificial Intelligence for Screening Colonoscopy: A Modelling Study," *The Lancet Digital Health* (April 2022), https://doi.org/10.1016/ S2589-7500(22)00042-5.

[129] Hyland et al. (2020) 364.

[130] Ethics and Governance of Artificial Intelligence for Health: WHO Guidance (2021): 36. ("AI technologies based on high-quality data can improve the speed and accuracy of diagnosis, improve the quality of care and reduce subjective decision-making.")

[131] However, there is considerable risk that the performance of an AI model might vary across institutions.

[132] George Bouchagiar, "The Long Road toward Tracking the Trackers and De-Biasing: A Consensus on Shaking the Black Box and Freeing from Bias," *Review of European Studies* 11, no. 1 (2019): 29, https://doi.org/10.5539/res.v11n1p27.

[133] Kleinberg et al. (2018).

1.6.2 Concerns

When discussing the challenges of AI-CDS systems, the issue of bias and potential discrimination has become one of the most frequently highlighted side-effects of this technology in recent years.[134] However, there are also other important concerns about undesirable consequences or limitations of AI-CDS systems.[135] The purpose of outlining these concerns is to give the reader a balanced perspective on the potential of AI-CDS systems and to situate the issue of bias and discrimination within a wider context of AI-related challenges and limitations.

Some of the concerns about AI systems are direct flipsides to the advantages mentioned in the previous section. While scalability is on the list of advantages associated with AI systems, scalability also means that systematic errors in an AI-CDS system are likely to cause harm to a larger number of patients. This point underscores the importance of subjecting these systems to preventive measures, including pre-deployment discrimination assessments. If these systems display discriminatory behaviour, there could easily be a systemic discrimination that could affect many patients.[136]

Some challenges are particularly, if not exclusively, associated with AI systems relying on neural networks based on deep learning. One such challenge is the potential for unpredictable behaviour in AI systems.[137] This is due to the autonomous elements of the learning process, which is particularly strong in deep learning systems, and the limited human oversight with what the algorithm learns, as noted in section 1.5.4. One of the most widely discussed concerns about AI systems, alongside the bias issue, is the lack of transparency, interpretability and explainability associated with opaque and inscrutable systems. Systems

---

[134] e.g., Fahse, Huber, and van Giffen (2021) 95-96.

[135] For a relatively early overview: Cohen et al. (2014).

[136] *White Paper on Artificial Intelligence - a European Approach to Excellence and Trust*, European Commission (Brussels, 19 February 2020), 11, https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf ; World Health Organization (2021) 55.

[137] Scherer notes their ability to find unforeseen and creative solutions to the problems they are intended to solve: Matthew U Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies," *Harvard Journal of Law & Technology* 29, no. 2 (Spring 2015): 363.

with these properties are popularly called ‘black boxes’ – and the use of them in medical contexts ‘black box medicine.’[138] In contrast, easily interpretable AI systems are sometimes called ‘white boxes’[139] or ‘glass boxes.’[140] In practice, however, AI systems exist on a continuum from completely inscrutable to more or less interpretable and explainable systems. Sometimes, the decision to deploy an opaque system will be the result of prioritising between objectives that cannot be equally fulfilled at the same time. For example, AI developers might have to prioritise between maximising the accuracy of a model and its interpretability,[141] or between a model’s interpretability and processing speed.[142]

While the challenges mentioned above apply to the use of AI for decision support in general, there are specific concerns within the context of healthcare. For example, concerns have been raised regarding the impact of AI on the professional autonomy of healthcare personnel, [143] as

---

[138] Price (2015).

[139] From the perspective of critical race theory, the term ‘black box’ can be seen as demeaning, especially when contrasted with the notion of a “white box” as something more preferable. For an introduction, see: Richard Delgado, Jean Stefancic, and Angela Harris, *Critical Race Theory (Third Edition): An Introduction* (New York: New York: NYU Press, 2017).

[140] e.g., Anna Markella Antoniadi et al., "Current Challenges and Future Opportunities for Xai in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review," *Applied Sciences* 11, no. 5088 (2021): 7, https://doi.org/10.3390/ app11115088.

[141] World Health Organization (2021) 27. (… “a possible trade-off between full explainability of an algorithm (at the cost of accuracy) and improved accuracy (at the cost of explainability)”); Chelsea Chandler, Peter W Foltz, and Brita Elvevåg, "Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness," *Schizophrenia Bulletin* 46, no. 1 (2019): 13, https://doi.org/10.1093/schbul/sbz105; Rich Caruana et al., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission" (Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, Association for Computing Machinery, 2015). (“In machine learning often a tradeoff must be made between accuracy and intelligibility. More accurate models such as boosted trees, random forests, and neural nets usually are not intelligible, but more intelligible models such as logistic regression, naive-Bayes, and single decision trees often have significantly worse accuracy.”); Laura Sikstrom et al., "Conceptualising Fairness: Three Pillars for Medical Algorithms and Health Equity," *BMJ Health & Care Informatics* 29, no. 1 (2022): 3, https://doi.org/10.1136/bmjhci-2021-100459.

[142] Andrew D Selbst, "An Institutional View of Algorithmic Impact Assessments," 35, no. 1 (2021): 181.

[143] Diaz-Asper et al. (2023).

well as patient autonomy.[144] Additionally, there is apprehension that the use of AI in healthcare may diminish the personal connection between healthcare providers and patients, potentially overlooking individual values and preferences that are not explicitly captured in data or algorithms.[145]

1.7 Clinical Decisions in Scope

1.7.1 Introduction

To elucidate the ramifications of different clinical decisions on the assessment of discrimination in an AI-CDS system, this thesis concentrates on a subset of clinical decision categories. The reason for selecting these categories is because they exemplify promising use cases of AI and are of great importance to patients. Moreover, the distinction between these decision categories brings forward important nuances concerning the relevant methodological elements of assessing discrimination in an AI-CDS system.

1.7.2 Diagnosis

Diagnosis in medicine means to determine the disease that causes a patient's symptoms.[146] A diagnosis is a classification task: what is the correct label to put on this patient? The purpose of diagnosis is not to predict a future development (prognosis) but to conclude on whether the patient has a certain medical condition at the time of the assessment. Typically, a clinician will diagnose a patient based on a consultation where the patient is examined and explains the experienced symptoms, as well as the patient's medical history, and laboratory tests.[147] The typical reasoning begins with a broad consideration of which diseases might be causing the patient's symptoms (differential diagnosis). The list of possible causes is then narrowed down

---

[144] Hannah van Kolfschooten, "Eu Regulation of Artificial Intelligence: Challenges for Patients' Rights," *Common Market Law Review* 59, no. 1 (2022): 92; Daria Onitiu, "The Limits of Explainability & Human Oversight in the Eu Commission's Proposal for the Regulation on AI- a Critical Approach Focusing on Medical Diagnostic Systems," *Information & Communications Technology Law* 32, no. 2 (2023): 173, https://doi.org/10.1080/13600834.2022.2116354.

[145] van Kolfschooten (2022) 93-94.

[146] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri, "Improving the Accuracy of Medical Diagnosis with Causal Machine Learning," *Nature Communications* 11, no. 3923 (2020): 1, https://doi.org/10.1038/s41467-020-17419-7.

[147] Ledley and Lusted (1959) 9.

by way of exclusion.[148] This reasoning inevitably involves a considerable element of human discretion. In exercising this discretion, clinicians rely on their professional judgment,[149] drawing from their education, experience, and knowledge of findings from scientific research.[150]

Ongoing research is currently being conducted to evaluate the potential of utilizing AI as a diagnostic tool for a diverse range of diseases. To name only a few examples, AI-CDS systems might help diagnosing diseases such as diabetic retinopathy,[151] several types of cancer,[152] COVID-19,[153] heart failure,[154] Parkinson's Disease[155] and autism spectrum disorder.[156]

Throughout the thesis, it is assumed that the output from a diagnostic AI-CDS system can be categorised into two types: binary classification indicating whether a patient is positive or

---

[148] Ibid.

[149] J. G. Mazoué, "Diagnosis without Doctors," *The Journal of Medicine and Philosophy* 15, no. 6 (1990), https://doi.org/10.1093/jmp/15.6.559. 560-561.

[150] David L Sackett et al., "Evidence Based Medicine: What It Is and What It Isn't," 312 (13 January 1996): 71-72.

[151] Feng Li et al., "Deep Learning-Based Automated Detection for Diabetic Retinopathy and Diabetic Macular Oedema in Retinal Fundus Photographs," *Eye* 36, no. 7 (2021), https://doi.org/10.1038/s41433-021-01552-8.

[152] E.g., Adam Yala et al., "A Deep Learning Mammography-Based Model for Improved Breast Cancer Risk Prediction," *Radiology* 292, no. 1 (2019), https://doi.org/10.1148/radiol.2019182716; Carmen C. Y. Poon et al., "AI-Doscopist: A Real-Time Deep-Learning-Based Algorithm for Localising Polyps in Colonoscopy Videos with Edge Computing Devices," *NPJ Digital Medicine* 3, no. 1 (2020), https://doi.org/10.1038/s41746-020-0281-z; Marc Combalia et al., "Validation of Artificial Intelligence Prediction Models for Skin Cancer Diagnosis Using Dermoscopy Images: The 2019 International Skin Imaging Collaboration Grand Challenge," *The Lancet Digital Health* 4, no. 5 (2022), https://doi.org/https://doi.org/10.1016/S2589-7500(22)00021-8.

[153] Mohamed Abd Elaziz et al., "New Machine Learning Method for Image-Based Diagnosis of Covid-19," *Plos one* 15, no. 6 (26 June 2020), https://doi.org/10.1371/journal.pone.0235187.

[154] Choi et al. (2020).

[155] Gupta et al. (2023).

[156] Jonathan T. Megerian et al., "Evaluation of an Artificial Intelligence-Based Medical Device for Diagnosis of Autism Spectrum Disorder," *NPJ Digital Medicine* 5, no. 57 (2022), https://doi.org/10.1038/s41746-022-00598-6.

negative in respect of a specific disease, or multi-class classification suggesting a diagnosis in a broader context. Binary classification models are typically applied in systems focused on testing for a particular disease. Multi-class models are used in systems intended to provide more diagnostic suggestions in a broader context.

1.7.3 Treatment Recommendation

The most important question following a diagnosis is what kind of treatment a patient should receive. For instance, choosing a treatment for lung cancer is a difficult medical decision which depends on several factors.[157] Various treatment options are available, but there may be considerable individual differences in how patients respond to different treatments.

Machine learning algorithms can be used to analyse information about prior patients and produce models for treatment recommendation for new patients.[158] AI-CDS systems for treatment recommendation are typically based on a combination of AI models that predict different target variables of relevance to the decision, including models for prognosis and models for assessing the risk of complications.[159] The treatment recommendation category encompasses decisions regarding which among several possible treatments a patient should receive. It also includes decisions on whether or not a patient should undergo a specific treatment. This aspect is particularly relevant in the context of surgical procedures, where the potential benefits of the procedure must be carefully weighed against the associated risks. The latter type of treatment recommendation is particularly considered in this thesis, as one of the case studies introduced in chapter 5 (the NORspine project) relates to the decision on whether to recommend spine surgery to individual patients.

1.7.4 Preventive Intervention

Due to their potential for predicting health-related events and developments before they occur, AI-CDS systems can facilitate preventive interventions that would not otherwise be

---

[157] "Treatment, Lung Cancer," NHS 29 July, 2021, https://www.nhs.uk/conditions/lung-cancer/treatment/.

[158] e.g., M. Berkan Sesen et al., "Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer," *AMIA Annual Symposium proceedings* 2012 (2012).

[159] Khoa A Tran et al., "Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection," *Genome Medicine* 13, no. 152 (2021), https://doi.org/10.1186/s13073-021-00968-x.

feasible.[160] For example, systems designed to predict adverse events can be used to monitor in-hospital patients continuously and call on medical personnel when a preventive intervention should be initiated.

Research on machine learning based on electronic health records suggests that AI-CDS can be used to predict adverse events such as sepsis,[161] acute kidney injury,[162] circulatory failure,[163] and post-operative delirium. In a study by Tomašev et al., a model developed through machine learning was able to predict 55,8 % of all inpatient episodes of acute kidney injury and 90,2 % of all acute kidney injuries that required administration of dialysis within a window of up to 48 hours before the event occurred.[164] In a similar vein, a study by Mikalsen et al. demonstrates how an ML algorithm can be applied to thousands of electronic health records in order to discover patterns associated with postoperative delirium, a complication often seen among elderly patients having gone through major surgery.[165] In theory, a model like the one developed by Mikalsen et al. could be used as clinical decision support for postoperative delirium.[166] Preventive intervention in relation to delirium, sepsis and other

---

[160] Ira S. Hofer et al., "Development and Validation of a Deep Neural Network Model to Predict Postoperative Mortality, Acute Kidney Injury, and Reintubation Using a Single Feature Set," *NPJ Digital Medicine* 3, no. 58 (2020), https://doi.org/10.1038/s41746-020-0248-0; World Health Organization (2021) 7.

[161] Shamim Nemati et al., "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the Icu," *Critical Care Medicine* 46, no. 4 (2018), https://doi.org/10.1097/CCM.0000000000002936.

[162] Nenad Tomašev et al., "A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury," *Nature* 572, no. 7767 (2019), https://doi.org/10.1038/s41586-019-1390-1.

[163] Hyland et al. (2020).

[164] Tomašev et al. (2019).

[165] Karl Øyvind Mikalsen et al., "Using Anchors from Free Text in Electronic Health Records to Diagnose Postoperative Delirium," *Computer Methods and Programs in Biomedicine* 152 (2017), https://doi.org/https://doi.org/10.1016/j.cmpb.2017.09.014.

[166] Mikalsen et al. conclude that “The proposed approach can be used as a framework for clinical decision support for postoperative delirium”: Mikalsen et al. (2017).

severe conditions could provide doctors with a significant head start in terms of prevention and treatment.[167]

1.7.5 Allocation of Scarce Resources

While many clinical decisions inherently involve some consideration of resource allocation, this thesis considers the allocation of scarce resources as a distinct category of clinical decisions. This category encompasses decisions where resource allocation is the central issue, rather than being just one of many considerations. In these situations, the question is not whether a patient would benefit from a given resource, such as a medical treatment. Rather, the question is whether the allocation of a resource to an individual patient should be prioritised, considering the interests of other patients and broader societal interests. One characteristic commonly associated with decisions falling into this category is the inclusion of threshold requirements or direct competition among patients. For example, an AI-CDS system intended to support the allocation of scarce resources could be one that provides a score indicating the match between donors and recipients, or one that produces a score predicting the severity of a patient's health needs in the near future. Moreover, the allocation of ventilators to patients with severe cases of COVID-19, as well as the decision on when to discontinue the usage of ventilators for individual patients, are example of a type of decision concerning the allocation of scarce resources for which the use of AI-CDS systems has been suggested.[168]

A fictional case study (the Simon Tesfay case) introduced in section 5.1 further illustrates how an AI-CDS system can be involved in the allocation of scarce resources.

1.8 The importance of EU law

The objective of this thesis is motivated partially by the proliferation of AI-CDS systems, partially by the concern about biases potentially causing discrimination, and partially by the EU legislature's response to AI. Consequently, the thesis concentrates on EU law. However, given the practical importance of non-discrimination laws at the national and constitutional

---

[167] Mustafa Suleyman and Dominic King, "Using AI to Give Doctors a 48-Hour Head Start on Life-Threatening Illness," *deepmind.com*, *DeepMind*, 31 July, 2019, https://deepmind.com/blog/article/predicting-patient-deterioration.

[168] World Health Organization (2021) 10.

levels in EU Member States, it is worth cementing the importance of EU law to dealing with the issue of bias and discrimination in AI-CDS systems.

This thesis particularly emphasizes the importance of EU law due to the EU's role in regulating AI technologies and its strong emphasis on enforcing fundamental rights in the context of AI systems. The AI Act will indeed be the world's first comprehensive AI regulation. When proposing the AI Act, the European Commission declared its ambitions "to preserve the EU's technological leadership and to ensure that Europeans can benefit from new technologies developed and functioning according to Union values, fundamental rights and principles."[169] This reference encompasses the non-discrimination principle, given its status as a fundamental principle of EU law. Thus, the EU has made it its ambition to ensure non-discrimination in AI systems. As mentioned in section 1.1, as part of the toolbox to ensure non-discrimination, the AI Act is likely to include one or more pre-deployment discrimination assessment requirements.

EU non-discrimination law is important not only because of its central role in regulating AI-CDS systems but also because it has a broader impact on non-discrimination laws in EU Member States and beyond.[170] For instance, in Norway, despite not being an EU member state, the legislature has stated in preparatory works (an important source of law in the Norwegian legal tradition) that the level of protection offered by Norwegian non-discrimination law shall be as least as good as that provided by the Equality Directives.[171] The Norwegian Supreme Court has followed this up by emphasizing the importance of case law from the CJEU when interpreting Norwegian non-discrimination law.[172]

---

[169] Explanatory Memorandum to the European Commission's AI Act Proposal, 21 April 2021: Section 1.1.

[170] Hellborg frames EU law as the engine driving the development of Swedish non-discrimination law: Sabina Hellborg, *Diskrimineringsansvar : En Civilrättslig Undersökning Av Förutsättningarna För Ansvar Och Ersättning Vid Diskriminering*, vol. 136, Skrifter Från Juridiska Fakulteten I Uppsala, (Uppsala: Iustus förlag, 2018), 37-38.

[171] Law proposal 5 April 2017 from the Norwegian Ministry of Children and Equality: Prop. 81 L (2016–2017) Act on Equality and Non-Discrimination, 45.

[172] Judgment of the Norwegian Supreme Court 29 June 2011 (Rt. 2011 s. 964).

1.9 Ethnicity and Sex as Protected Characteristics

1.9.1 Introduction

This thesis concentrates its analysis on provisions of EU law that prohibit discrimination on grounds of ethnicity and sex. Thus, it primarily develops methodological elements of assessing sex discrimination and ethnic discrimination. However, these elements, and the discussions in the thesis, are largely transferrable to other discrimination grounds. The following first illuminates the rationale behind the focus on sex and ethnicity and then establishes what is meant by ethnicity and sex within this thesis. These terms are used in accordance with how they are understood and applied in the context of EU non-discrimination law. The role (and controversies) of these characteristics in clinical decision-making is further explained in chapter 4, concerning the sources of biases in AI-CDS systems.[173]

1.9.2 Purpose of the limitation

EU non-discrimination law prohibits discrimination based on several different characteristics ('protected characteristics').[174] This thesis concentrates its analysis on provisions of EU law that prohibit discrimination on grounds of ethnicity and sex. This means that the thesis primarily illustrates how algorithmic discrimination may arise by reference to these specific discrimination grounds. However, it is important to note that this does not diminish the significance of addressing algorithmic discrimination based on other protected characteristics. Additionally, this focus does not formally restrict the applicability of the thesis's findings to the assessment of discrimination in an AI-CDS system solely with regard to sex and ethnicity.

However, when discussing examples of discrimination, it is important to consider the specific nuances associated with each protected characteristic. The protected characteristics in non-discrimination law often have unique complexities. For instance, unlike sex and ethnicity,

---

[173] Section 4.4.3.3.

[174] The widest range of characteristics is provided by Article 21 of the EU Charter, which mentions "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." However, not all of these characteristics are protected by the Equality Directives. The words "such as" in Article 21 of the Charter, which precede the list of prohibited characteristics, entail that Article 21 is not limited to the explicitly mentioned characteristics, but in practice the CJEU has refused to apply the Charter provision to other grounds: Judgment of 18 December, 2014, Kaltoft, C-354/13, ECLI:EU:C:2014:2463.

disability is a protected characteristic that not all individuals possess, and it encompasses a wide range of conditions which vary significantly within the protected group. There are also certain differences between sex and ethnicity as protected characteristics. EU non-discrimination law is often based on a binary distinction between biological men and biological women, whereas ethnicity is a multi-class characteristic.[175] It would not have been feasible within the scope of this thesis to treat several different protected characteristics with the level of attention that they each deserve. Each discrimination ground deserves in-depth analysis, but limitations of time and resources necessitate a focused approach. However, it is my aspiration that future research endeavours will expand upon the analysis of bias and discrimination in AI-CDS systems, perhaps refining the methodological elements developed in this thesis to accommodate the inclusion of other protected characteristics.

The selection of biological sex and ethnicity as the focal points of this thesis is motivated, in part, by the availability of examples of biases in AI systems (and comparable technologies) related to these characteristics. Both ethnicity and biological sex are central topics in current societal debates and controversies related to bias and discrimination in AI systems. Moreover, discrimination based on ethnicity and biological sex has deep historical roots in healthcare, making them particularly important categories to study and understand in the context of emerging technologies such as AI-CDS systems.

1.9.3 "Racial and ethnic origin"

The Racial Equality Directive (RED) prohibits discrimination on the basis of "racial or ethnic origin".[176] The term "racial or ethnic origin" is not defined in the Directive, and the CJEU has been very reluctant to reduce the scope of this term to more specific attributes. Moreover, the CJEU does not appear to distinguish between *race* and *ethnicity*. In the cases it has heard so far, it seems that the Court has considered discrimination based on ethnicity only. Academic literature suggests that the Court is perhaps more comfortable with the term 'ethnicity' as opposed to 'race.'[177] This would be understandable in light of how controversial the word 'race' is in many European countries. Importantly, as the school of thought known as 'critical

[175] See section 1.9.4 below.

[176] Articles 1 and 2 RED.

[177] Evelyn Ellis and Philippa Watson, *Eu Anti-Discrimination Law*, 2nd ed. ed. (Oxford: Oxford University Press, 2012), 32-33.

race theory' stresses, 'race' is a social and legal construct – not a biological fact.[178] This thesis primarily uses the word 'race' when referring to other literature that uses this term. Within this thesis, the reader should always assume that 'race' refers to social and/or legal constructs. This is also in line with the spirit of the RED. Recital 6 of the Directive's preamble explicitly rejects "theories which attempt to determine the existence of separate human races."

When it comes to the term 'ethnicity,' the CJEU has adapted the European Court of Human Rights' explanation of ethnicity as a concept which "has its origin in the idea of societal groups marked in particular by common nationality, religious faith, language, cultural and traditional origins and backgrounds."[179] This wording stems from cases decided under Article 14 of the European Convention on Human Rights, a provision which does not itself use the term 'ethnicity.'[180] Instead, it lists a number of more specific protected characteristics, including 'race' and several other grounds which may be connected with ethnicity, such as particularly colour, language, religion, national or social origin and "association with a national minority." It is likely that the CJEU would consider discrimination on these grounds (non-exhaustively) as ethnic discrimination.[181]

1.9.4 "Sex"

Historically, the prohibition of sex discrimination has been an important driving force behind the development of EU non-discrimination law. Sex discrimination has been prohibited since the establishment of the Common Market, initially for the economically oriented purpose of ensuring fair competition between Member States.[182] Since then, it has become recognised as

---

[178] Delgado, Stefancic, and Harris (2017).

[179] Judgment (GC) of 15 July, 2015, CHEZ, C-83/14, ECLI:EU:C:2015:480, para. 46; Judgment of 6 April, 2017, Jyske Finans, C-668/15, ECLI:EU:C:2017:278, para. 17.

[180] E.g. Judgment of the European Court of Human Rights of 13 December, 2005, Case of Timishev V. Russia (Applications 55762/00 and 55974/00), para. 55.

[181] Roderick Liddell and Michael O'Flaherty, *Handbook on European Non-Discrimination Law 2018 Edition* (European Union Agency for Fundamental Rights and Council of Europe, 2018), 197.

[182] Tamara K. Hervey, *Justifications for Sex Discrimination in Employment* (London: Butterworth, 1993), 37; Samantha Besson, "Gender Discrimination under Eu and Echr Law: Never Shall the Twain Meet?," *Human Rights Law Review* 8, no. 4 (2008): 658, https://doi.org/10.1093/hrlr/ngn023.

a fundamental right and a general principle of EU law.[183] In the TEU, "equality between women and men" is emphasised as one of the founding values of the EU.[184]

At the level of secondary EU law, sex discrimination is prohibited by several directives.[185] For instance, Directive 2006/54/EC, which applies in matters of employment and occupation, only concerns sex discrimination. This is also the case with the relevant directive regarding sex discrimination in the context of healthcare: Directive 2004/113/EC ('Goods and Services Equality Directive'/'GSED'), which prohibits discrimination between men and women in relation to access and supply of goods and services.

The abovementioned formulations in the TEU, as well as the provisions of the directives concerning sex discrimination, indicate that EU law relies on a binary definition of 'sex.' Such a binary definition may seem out of sync with the contemporary societal discourse on gender identity and gender plurality. The laws were created at a time when these issues were not discussed nearly as much as they are nowadays. The Directives do not explicitly clarify how they define 'sex,' or what the relationship is between 'sex' defined by basic biological properties (typically based on chromosomes and reproductive organs),[186] the broader notion of 'gender' (denoting categories that are socially constructed rather than biologically defined),[187] and an individual's self-defined gender identity. The terms 'sex' and 'gender' sometimes appear to be used interchangeably in EU law. For instance, the GSED mentions "gender equality" at certain points in the recitals,[188] while predominantly focussing on prohibiting "discrimination based on sex."[189] However, recital 12 of the GSED indicates that

---

[183] Judgment of 8 April, 1976, Defrenne, C-43/75, ECLI:EU:C:1976:56, para. 12; Besson (2008) 658.

[184] Articles 2-3 TEU.

[185] Directive 79/78/EC; Directive 2006/54/EC (Recast Directive); Directive 2010/41/EU (Self-Employment Equality Directive).

[186] "Gender and Health," World Health Organization (Web page) accessed 7 November, 2023, https://www.who.int/health-topics/gender#tab=tab_1.

[187] This is based on the World Health Organization's definition of 'gender': Ibid; On the distinction between sex and gender, see, e.g., Sharon Cowan, ""Gender Is No Substitute for Sex": A Comparative Human Rights Analysis of the Legal Regulation of Sexual Identity," *Feminist Legal Studies* 13, no. 1 (2005): 70, https://doi.org/10.1007/s10691-005-1457-2.

[188] Recitals 6, 7 and 16 GSED.

[189] e.g., Article 1 GSED.

the directive is probably oriented towards biological sex, as it refers to "physical differences between men and women."

Whether someone is discriminated against on account of traits that are typical of persons that are medically categorised as female or because they have an appearance that fits an offender's perception of a woman, is probably not decisive in relation to direct discrimination in EU law. In both cases, it would be arguable that discrimination occurs on the basis of sex. In relation to indirect discrimination, which is a group-oriented notion, the question arises whether one should count as 'men' those who are medically defined as men or those that appear in accordance with a social construction of what a man is. In practice, defining such groups would be challenging. As a general principle, however, measures that disadvantage persons according to their biological sex status and measures that disadvantage people according to their gender are probably both prohibited. Therefore, this thesis does not categorically distinguish between sex and gender. In practice, however, biological sex is important in the context of healthcare, and patients are routinely categorised according to their biological sex in this context. To reflect this situation, and to adhere to the terminology of EU law, this thesis primarily uses the term 'sex' in the following.

2 Research Method and Methodological Reflections

2.1 Doctrinal legal research – Law and Context

The objective of this thesis is to develop methodological elements of assessing discrimination in an AI-CDS system before its deployment. This chapter explains how these methodological elements are developed (the research method) and reflects on the methodology underpinning the research itself.

With a few reservations which will be explained in the following, this research is primarily anchored in a doctrinal method. Doctrinal legal research (or 'legal dogmatics') is commonly regarded as research that aims to determine what the law is. However, this does not really say much about the methods that doctrinal legal research may encompass. A more informative description of doctrinal legal research is "research which provides a systematic exposition of the rules governing a particular legal category, analyses the relationship between rules, explains areas of difficulty and, perhaps, predicts future developments."[190] In accordance with this understanding, doctrinal legal research encompasses a broad range of studies into the law which are predominantly occupied with finding out what the law is. Statements about what the law is, typically take the shape of 'legal doctrines.' Chynoweth thus defines doctrinal legal research as research which is "concerned with the discovery and development of legal doctrines."[191]

The notion of a legal 'doctrine' can be defined very generally as "rules and principles,"[192] and they can (and often are) studied as such, without the need for further knowledge of any

---

[190] This definition is cited by Hutchinson and Duncan, and stems from the Pearce Committee that reviewed research in Australian law schools in the 1980s: Dennis Pearce, Enid Campbell, and Don Harding, *Australian Law Schools: A Discipline Assessment for the Commonwealth Tertiary Education Commission* (Canberra: Australian Government Publishing Service, 1987); Terry Hutchinson and Nigel Duncan, "Defining and Describing What We Do: Doctrinal Legal Research," *Deakin L. Rev.* 17, no. 1 (October 2012): 101.

[191] Paul Chynoweth, "Legal Research," *Advanced research methods in the built environment* 1 (2008). 30.

[192] William Twining, *Law in Context: Enlarging a Discipline* (Oxford: Clarendon Press, 1997), 39.

particular context. However, the content of rules and principles can also be articulated in a manner that more specifically informs of how they are understood in a specific context. With this in mind, a legal 'doctrine' can more specifically be understood as a systematic formulation of what the law is in particular contexts, which implies that seeking knowledge of what the law is requires an application of the relevant legal rules to particular facts and situations.[193] The analyses in this thesis involve the application of EU non-discrimination law to the particular facts of AI-CDS systems. The formulation of how the law is understood in this context, is anchored in a doctrinal method.

As regards the sources that doctrinal legal research relies on, there is arguably a traditional, narrow understanding, and a broader understanding. A narrow understanding of doctrinal legal research might only cover research that primarily or entirely relies on sources that are internal to the legal system.[194] According to a broader understanding, one could say that doctrinal legal research is research that formulates legal doctrines,[195] even though non-legal sources are relied on to ensure that the doctrines are formulated with a high level of context-specificity. When this thesis is described as primarily anchored in a doctrinal method, this is based on the broader understanding of doctrinal legal research.

However, the purpose of the analyses in this thesis is not strictly doctrinal, as the methodological elements developed in this thesis amount to more than the articulation of legal doctrines. Legal doctrines do not, as such, prescribe the methodological elements of assessing discrimination in an AI-CDS system in a pre-deployment context. The method of developing these elements – i.e., considerations, principles, criteria, and methods of such assessment – is further described in section 2.2 below. Subsequently, section 2.3 discusses how such contextual developments align with 'law-in-context' approaches.

---

[193] Chynoweth (2008) 29-30; Pauline C Westerman, "Open or Autonomous? The Debate on Legal Methodology as a Reflection of the Debate on Law," in *Methodologies of Legal Research: Which Kind of Method for What Kind of Discipline?*, ed. Mark van Hoecke (Oxford: Hart, 2011), 90-91.

[194] Chynoweth (2008) 30; Hutchinson and Duncan (2012) 114-15.

[195] Chynoweth (2008) 20.

2.2 How the Thesis Develops Methodological Elements of Pre-Deployment Discrimination Assessments

The methodological elements that this thesis aims to develop have been framed in section 1.1 as consisting of a set of questions (considerations), including the principles and criteria that guide these considerations, and relevant methods of answering them. The development of these methodological elements is conducted through an analytical process involving two steps: interpretation of the non-discrimination principle in EU law and adaptation of its doctrines to the particular context of assessing discrimination in an AI-CDS system before its deployment. These two steps are not necessarily explicated or distinguished in the analysis. In practice, the line between these two steps may be blurred. Both steps involve an aspect of contextualisation.

The *interpretation* of law for the purpose of articulating legal doctrines of relevance to the assessment of discrimination in AI-CDS systems, involves interpreting the non-discrimination principle in the light of non-legal sources of knowledge about AI-CDS systems and the issue of biases in these systems. As Mantelero has highlighted, although fundamental rights principles apply to AI systems, they need *contextualisation* in order to be implemented in assessments of AI systems.[196]

The *adaptation* of the legal doctrines that constitute the non-discrimination principle in EU law, involves further contextualisation of those legal doctrines so that the considerations, principles, and criteria that should guide an assessment of discrimination in a pre-deployment setting may be articulated. The articulation of such elements of an assessment methodology arguably extend beyond what is traditionally understood as a contribution of doctrinal legal research. This contribution is a result of a method that is heavily informed by non-legal materials. Particularly, the list of questions that an assessor should raise and the proposals for methods of answering these questions, are not legal doctrines as such. Nevertheless, these methodological elements are closely related to the development of legal doctrines. The questions to raise during a pre-deployment discrimination assessment are articulated for the purpose of enabling a proper application of the non-discrimination principle in this context. They are therefore directly dependent on an interpretation of this principle. The same applies for methods that an assessor may use to answer the questions – those methods could not be

[196] Mantelero 2022 p 162

deemed relevant on any other basis than a contextually informed interpretation of the non-discrimination principle.

Moreover, the assertion that certain considerations *ought* to be conducted as part of a pre-deployment discrimination assessment involves an element of discretionary judgment. Non-discrimination law is not articulated with such an assessment in mind, and there may be different views on which considerations an assessment should involve. In this regard, it is important to note that the thesis aims to develop methodological elements that reflect an interpretation of EU non-discrimination law which is loyal to the legislative aims and the jurisprudence of the CJEU, rather than maximising practical feasibility, resource efficiency during testing and assessment of AI systems, or other objectives. The judgment exercised when articulating the methodological elements developed in this thesis is guided by this ambition of loyalty to EU law and CJEU jurisprudence.

Despite the ambition of loyalty to legislative aims and CJEU jurisprudence, it is worth noting that certain prioritisations are made when analysing legal sources of interpretation within this thesis. The thesis involves analyses of several components of the non-discrimination principle in EU law. If this project had a strictly doctrinal ambition, it might have considered the CJEU's case law in further depth, taking into account all cases that potentially illuminate some aspect of the non-discrimination principle. However, the objective of this thesis is not to cover every single CJEU ruling concerning the topic of discrimination. Rather, it is important to the objective of this thesis that methodological elements based on the non-discrimination principle cover the various components of this principle that an assessment methodology ought to reflect. Moreover, the examination of non-legal materials, particularly from computer science and medicine, is instrumental to the objective of the thesis. Such examination is therefore prioritised over an exhaustive study of CJEU case law. Regardless, CJEU case law from recent years is well-represented within the thesis, and literature on EU non-discrimination law has been studied for the purpose of identifying case law references worthy of further investigation.

One clear-cut choice that is made, however, is the choice not to specifically investigate the case law of the European Court of Human Rights (ECtHR). ECtHR case law, especially pertaining to Article 14 ECHR, may be relevant to the interpretation of EU non-discrimination law. However, it is primarily the case law of the CJEU that determines the proper interpretation of EU non-discrimination law. Besides, the inclusion of ECtHR case law

would expand the scope of interpretation sources to an extent that is disproportionate to the significance of ECtHR case law in relation to the objective of the thesis.

2.3 Law-in-context

Traditional understandings of doctrinal legal research are oriented towards deriving doctrines from legal sources: The law is "just there" within the available sources of legal interpretation, and the articulation of its doctrines does not require further knowledge of a particular context.[197] As noted above, this thesis relies heavily on knowledge of a particular context, and its objective is not limited to deriving doctrines from legal sources. When consideration of context is an important feature of the method one applies in research that is primarily legal, this is often referred to as 'law-in-context.'[198] The subject matter of bias and discrimination in AI-CDS systems is discussed across multiple academic disciplines. It generally benefits from a contextual approach.[199] In this thesis, the contextual knowledge is particularly important due to the objective of developing methodological elements of an assessment, rather than merely developing or deriving legal doctrines from legal sources.

Law-in-context is legal research which draws considerably on non-legal sources and may involve more salient interdisciplinary aspects than doctrinal legal research in a narrower sense. With reference to Twining, 'context' may include the practices of those most directly affected by the rules being analysed or the "light to be thrown on particular problems by the techniques and findings of other disciplines."[200] In this thesis, ML practices and techniques inform the development of methodological elements of pre-deployment discrimination assessments. Findings from ML research also influence this development. For example, it is argued in chapter 8 that deep neural networks may unintentionally rely on protected characteristics in a way that amounts to direct discrimination under EU non-discrimination law. This insight leads to the suggestion of specific considerations aimed at assessing whether

---

[197] Twining contrasts what he calls a contextual approach, cf. section 2.3 below, with treating law as "being 'just there' to be studied in isolation": Twining (1997) 44.

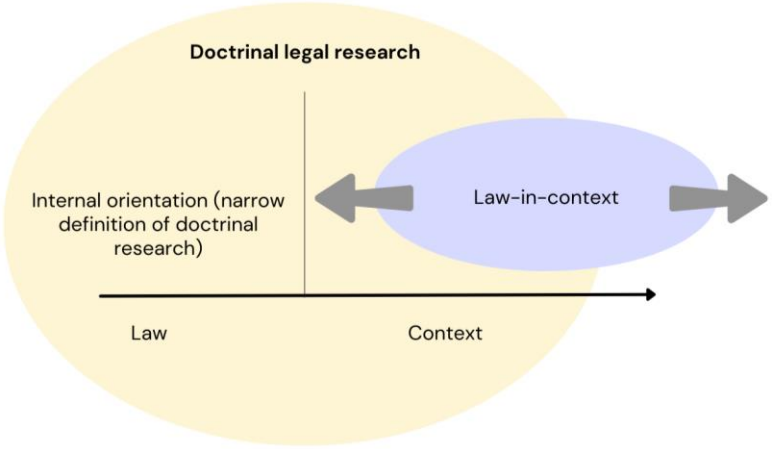
[198] Ibid; Chynoweth (2008). 31.

[199] Baldwin uses the label "transdisciplinary field" about "an area of study where different disciplines and research traditions talk to each other and where work is informed and influenced by these conversations". Robert Baldwin, Martin Cave, and Martin Lodge, *The Oxford Handbook of Regulation* (Oxford: Oxford University Press, 2010), 12.

[200] Twining (1997) 44.

an AI-CDS system might cause direct discrimination through such hidden inferences. These considerations could not have been developed based only on legal sources. Moreover, ML techniques heavily inform the suggestions of relevant methods to apply during a pre-deployment discrimination assessment.

Law-in-context approaches can either be seen as a sub-genre of doctrinal research or as a methodology of its own, depending on the exact law-in-context method a researcher applies and how one defines doctrinal legal research. In this thesis, doctrinal research is broadly defined, and it is submitted that the research is primarily anchored in a doctrinal method. The use of research material from other disciplines in this thesis can be described as 'auxiliary,'[201] meaning that the material serves a supportive and subsidiary function.[202] As Taekema and van Klink have put it: "The legal researcher defines a problem, which he cannot solve with legal methods only, so that there is a need for input from another discipline."[203] Auxiliary use of non-legal sources imply that the fundamental nature of the research does not change, even though an interdisciplinary aspect is introduced.



[201] Sanne Taekema and Prof van Klink, "On the Border: Limits and Possibilities of Interdisciplinary Research," *Bart van Klink and Sanne Taekema, Law and Method. Interdisciplinary research into Law (Series Politika, nr 4), Tübingen: Mohr Siebeck* 2011 (2011). 11.

[202] https://www.merriam-webster.com/dictionary/auxiliary

[203] Taekema and van Klink (2011). 11.

Figure 3. Law-in-context approaches can be defined within the frame of doctrinal legal research or outside, depending on how one defines doctrinal research and the role of context in a research project.

In this thesis, multidisciplinary literature is relied on to describe how ML is being used to create AI-CDS systems, how these systems may be used, and how bias and potential discrimination may occur in these systems.[204] Particularly, the thesis relies on literature from computer science and medical science. However, Part II includes a wider selection of multidisciplinary literature and can be said to contain more context than law. It relies primarily on material from non-legal disciplines. However, while these sources are distinctly non-legal, they serve a purpose within the overarching approach of the thesis, which is a doctrinally anchored legal method informed by multidisciplinary materials.[205] Hence, while the label of law-in-context may be appropriate for this thesis, the research method is fundamentally doctrinal.

2.4 EU law method and Particular Considerations Regarding the Forthcoming AI Act

The thesis relies on a doctrinally anchored legal research method which emphasises development of elements of an assessment methodology, and is informed by multidisciplinary literature. As described above, legal interpretation is a central part of pursuing the research objective of the thesis. Within a doctrinal research methodology, the exact norms of which sources to use and how they should be interpreted may vary between different legal jurisdictions and cultures. This thesis analyses EU law. Therefore, given the ambition of developing methodological elements in a manner that is loyal to the aims of the EU legislature and CJEU jurisprudence, it applies the interpretation method that is accepted when interpreting EU law. This method is defined by the approach taken by the Court of Justice of

---

[204] Insert reference to Synne Sæther Mæhle 2020, pointing out how materials from other sciences may be necessary to provide a proper context for legal research questions related to the use of big data in the health sector.

[205] Doctrinal legal research been described by Westerman as drawing on "the legal system as the main supplier of concepts, categories and criteria".Westerman p 94 in this book: Mark van Hoecke, *Methodologies of Legal Research : Which Kind of Method for What Kind of Discipline?*, 1st ed. ed. (Oxford,Portland, Oregon: Hart Publishing, 2011). While this is not descriptive of Part II of this thesis, the purpose of Part II is to prepare the ground and set the parameters for legal analysis.

the European Union. The CJEU's interpretation method is well known to legal practitioners and researchers in the EEA, and it is not described comprehensively in this thesis.[206] The method is essentially the same for interpretation of secondary[207] and primary EU law.[208] However, when it comes to interpretation of the Charter of Fundamental Rights, Article 52 of the Charter lays down certain specific interpretation rules.[209]

One aspect of the EU law method is worth highlighting before proceeding with this thesis: the role of preparatory works, which is relevant to this thesis's analysis of the forthcoming AI Act. As a general starting point, preparatory works have limited importance within the EU law method. The most important sources are the wording of statutes, their legislative aims, and CJEU jurisprudence. However, because the legislative process in the EU involves three different legislative institutions (the European Commission, the EU Council, and the European Parliament), the legislative aims may not always be clear or unified.[210] Therefore, the purposive considerations should primarily build on the legislative aims that are stated in the regulation itself or in the preamble, because the statements therein have been agreed upon by the three legislative institutions.

---

[206] This method of interpretation is characterised by consideration not only of the wording of a provision, "but also of its context, the objectives pursued by the rules of which it is part and, where appropriate, its origins": C-263/18 *Tom Cabinet*, para 38 and the references cited therein; C 53/81 judgment 23 March 1982 para 9. See also Samuli Miettinen and Merita Kettunen, "Travaux to the Eu Treaties: Preparatory Work as a Source of Eu Law," *Cambridge Yearbook of European Legal Studies* 17 (2015), https://doi.org/10.1017/cel.2015.6. 148.

[207] Secondary EU law are regulations, directives, decisions, recommendations and opinions: https://commission.europa.eu/law/law-making-process/types-eu-law_en

[208] Primary EU law includes the founding treaties, i.e., the Treaty on the Functioning of the European Union (TFEU) and the Treaty on European Union (TEU), as well as the EURATOM Treaty and the EU Charter of Fundamental Rights (the 'Charter') and the non-statutory general principles of EU law.

[209] For example, it specifies that the 'explanations' accompanying the Charter should be taken into account.Article 52(7) of the Charter; 'Explanations relating to the Charter of Fundamental Rights', OJ C 303, 14.12.2007, p. 17–35.

[210] Koen Lenaerts and Jose A. Gutierrez-Fons, "To Say What the Law of the Eu Is: Methods of Interpretation and the European Court of Justice," *Columbia Journal of European Law* 20, no. 3 (2014): 22.

As regards the role of preamble recitals, specifically, the CJEU has indeed confirmed their usefulness when interpreting secondary EU law, although they are not binding in themselves.[211] Their status as a source of legal interpretation is the subject of some debate.[212] Recitals are primarily helpful when they have an immediate connection to one or more operative provisions, in which case they might illuminate the rationale of those provisions. However, they may not be relied on to arrive at conclusions that would contradict the wording of a legislative act.[213] The use of a recital as a source of statutory interpretation is particularly questionable where the recital is not directly connected to a specific provision in the respective regulation or directive.[214]

Compared to preamble recitals, the preparatory works from the drafting history of a regulation or directive generally bear less weight as sources of interpretation. However, such documents are relied on by the CJEU in the interpretation of secondary EU law.[215] This thesis sometimes refers to two of the documents accompanying the Commission's AI Act proposal: the Impact Assessment (IA) and the Explanatory Memorandum (EM).[216] As preparatory works, these

---

[211] Their "usefulness as criteria for interpretation" have indeed been confirmed by the CJEU, see the cases referred to by Advocate General Ruiz-Jarabo Colomer: AG Opinion in Maruko, para 76.

[212] Tadas Klimas and Jurate Vaiciukaite, "The Law of Recitals in European Community Legislation," *ILSA J. Int'l & Comp. L.* 15 (2008); Maarten den Heijer, Teun van Os van den Abeelen, and Antanina Maslyka, "On the Use and Misuse of Recitals in European Union Law," *Amsterdam Law School Research Paper*, no. 2019-31 (2019).

[213] of 19 November 1998, C-162/97 Gunnar Nilsson,para 54; of Case C-308/97 Giuseppe Manfredi V Regione Puglia,para 30.

[214] Christa Tobler and Kees Waaldijk, "Case C-267/06 Tadao Maruko V Versorgungsanstalt Der Deutschen Bühnen: Judgement of the Grand Chamber of the Court of Justice of 1 April 2008, Not yet Reported," *Common Market Law Review* 46, no. 2 (2009). 731.

[215] As regards their status in the interpretation of primary EU law, see Miettinen and Kettunen (2015). 146; Lenaerts and Gutierrez-Fons (2014). 22-23. One might argue that there is a trend towards increased importance of preparatory works in the interpretation of EU law. A recent example is the Opinion from AG Campos Sánchez-Bordona in Case C-667/21, in which the AG considers the preparatory works to the GDPR in paragraphs 81, ,73, 106, and 117.

[216] Impact assessment: European Commission, *Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Brussels, 21 April 2021); the Explanatory Memorandum is integrated in the Commission's AIA proposal (Appendix 1).

documents are not part of the AI Act and statements therein are not binding on the CJEU or the Member States. They are evidence of the drafting history. As such, they may only be relevant to the extent that they illuminate the purpose or origin of a provision and confirms an interpretation that would be plausible based solely on the wording of the provision.

Compared to the Impact Assessment, one may argue that the Explanatory Memorandum (EM) accompanying the AI Act proposal is slightly more important to the interpretation of the AI Act. The EM is also a preparatory work that is not binding in itself. It is not a part of the AI Act. However, in contrast to the IA, the EM was presented as an integrated part of the European Commission's proposal. As such, the EM sets out the rationale for proposing the AI Act and the objectives of the proposed regulation. Moreover, the EM is easily accessible and it relates to a complex and technical regulation. In the literature on EU law interpretation methods, the ease of access to preparatory works and the technical nature of regulations have been proposed as explanations as to why the CJEU appears to rely increasingly on preparatory works in the interpretation of secondary EU law.[217] Thus, the availability of the EM and the technical nature of the AI Act suggests that the EM can be given some weight in the interpretation, if it is supported by a strictly textual interpretation of the relevant provisions.

In relation to the AIA, it should also be noted that once the enacted version of the AIA is officially translated into the languages of the EU Member States, the operational provisions of the different language versions should be taken into account in the interpretation. Such language versions do not exist at the time of writing this thesis.

[217] Lenaerts and Gutierrez-Fons (2014) 29., with reference to Soren Schonberg and Karin Frick, "Finishing, Refining, Polishing: On the Use of Travaux Preparatoires as an Aid to the Interpretation of Community Legislation," *European law review* 28, no. 2 (2003).

2.5 Way of Reference to the Forthcoming AI Act

At the time of writing this thesis, the final version of the AIA has not yet been agreed upon by the EU legislature. However, there is agreement among the legislative bodies involved in the ongoing negotiations on the core principles and structure of the regulation. There is also reason to expect that many provisions will be adopted as proposed by the European Commission, to the extent that the EU Council and the European Parliament have not suggested changes to these provisions, according to their respective negotiation positions. The negotiation positions relied on in this thesis are attached to the thesis as Appendixes 1, 2, and 3. Appendix 1 contains the Commission's proposal of 21 April 2021. Appendix 2 contains the Council's negotiation position (the 'General Approach'), whereas Appendix 3 contains the Parliament's position (the 'Compromise Text').

When referring to the AIA in footnotes throughout the thesis, the Commission's proposal and the Parliament's draft Compromise Text are primarily relied on. The reason for referring to the Parliament's position rather than the Council's position is that the Parliament proposes a stronger focus on fundamental rights, including the right to non-discrimination. Particularly, the Parliament's proposal clarifies that considerations related to discrimination must be included in a pre-deployment assessment of an AI system. Consequently, focussing on the Parliament's position as a supplement to the Commission's proposal provides a picture of the proposed provisions which is adequate in the light of the objective of the thesis. Alongside of those two documents, the Council's position does not add anything of considerable importance, given the objective of the thesis. The relevant aspects of the draft Compromise Text are examined in Part III of the thesis.

The reader should note that it is often necessary to refer to the Commission's proposal when referring to provisions in which the Parliament's draft Compromise Text (as per Appendix 3) does not propose any changes, because Appendix 3 does not repeat all provisions in the Commission's proposal. The Commission's proposal is referred to in footnotes as 'AIA (EC),'whereas the Parliament's position is referred to as AIA '(EP).' Differences between the draft Compromise Text and the Commission's Proposal are highlighted where it is essential to the analysis presented.

As chapter 7 further elaborates, an important aspect of the Parliament's draft Compromise Text is that it explicitly requires that certain aspects of discrimination must be included in a pre-deployment discrimination assessment, compared to the Commission's proposal. Thus, it is currently not certain to what extent the AIA will require pre-deployment discrimination assessments. Chapter 7 primarily explores the extent to which such requirements will exist if the relevant parts of the Parliament's draft compromise text are adopted. However, as chapter 7 also considers, there are provisions in the Commission's proposal which may be interpreted as requiring a pre-deployment discrimination assessment. In the unlikely event that the AIA ends up not requiring any form of pre-deployment discrimination assessment, the contributions of this thesis will be less relevant in terms of complying with current law (after adoption of the AIA). However, chapter 7 emphasises the utility of this thesis's contributions also in relation to voluntary efforts to prevent discrimination in AI systems.

PART II: BIAS

3 Bias

3.1 Introduction

3.1.1 Purpose of the Chapter

This chapter serves two purposes. First, it aims to establish an understanding of the concept of 'bias' based on a multidisciplinary selection of literature.[218] The understanding of how 'bias' is conceptualised and defined in different contexts forms part of the knowledge foundation upon which the thesis proceeds. As such, this chapter provides a frame of reference that informs the analyses conducted later in the thesis. Second, this chapter searches for a working definition of 'bias' that is relied on throughout the remainder of the thesis.

The purpose of the working definition of 'bias' is to articulate the factual phenomenon the thesis seeks to explore at its most foundational level. While such an initial definition should clarify what this thesis means when it refers to 'bias.' it is important that the definition avoids prematurely narrowing the scope and pre-empting the subsequent analyses. Instead, the ambition in this chapter is to find a starting point that encompasses the many facets of bias and remains open-ended enough to adapt to the evolving discussions in the remainder of the thesis. For the subsequent analysis of bias and discrimination in AI-CDS systems to unfold coherently, it is important that the working definition of 'bias' is sufficiently broad and adaptable. Consequently, the working definition should not be articulated in a way that is easily confused with 'discrimination' or other legal concepts.

3.1.2 Introduction to a Discourse with Multiple Dimensions

The term 'bias' is frequently encountered in the multidisciplinary literature on big data, ML and AI. In technical machine learning literature, 'bias' is often discussed as a technical challenge that must be dealt with in the process of developing AI systems because bias can hamper the systems' performance.[219] Other works are more concerned with the social and

---

[218] The approach to non-legal material in this thesis is described in chapter 2.

[219] e.g., Christopher M Bishop and Nasser M Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4 (Springer, 2006); Hastie, Tibshirani, and Friedman (2017) 223, et seq.

societal implications of ‘biased’ AI, highlighting the social harms that biased AI may cause.[220] This type of bias literature often refers to social values and ethical principles, suggesting that biased AI is problematic because of the potential conflict with such values and principles.[221] A smaller portion of the bias-related literature directly discusses legal rules that may be violated if AI systems exhibit bias.[222]

Works that discuss the issue of bias (in AI systems and beyond) often do so without defining what is meant by ‘bias.’ The literature reviewed for this chapter shows that conceptual discrepancies and nuances emerge when ‘bias’ literature from different disciplines is compared. It also shows that ‘bias’ can have more than one meaning within the same field. Many works that address ‘bias’ seem to circle around what appears to be a common, often domain-specific, core understanding of the concept, without explicating what that core understanding is.[223] In certain contexts it may be the case that the meaning of the term ‘bias’ is self-evident to the readers. However, the term ‘bias’ may not be familiar to many legal readers, and it has no formal legal definition within the EU legal order.[224] At the same time, the issue of ‘bias’ in AI systems has been highlighted in numerous policy documents and

---

[220] e.g., Barocas and Selbst (2016); Tal Zarsky, "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making," *Science, Technology, & Human Values* 41, no. 1 (2015), https://doi.org/10.1177/0162243915605575.

[221] e.g., generally, Cathy O'neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Crown, 2016); Ruha Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code* (Oxford: Polity, 2019); Joseph J. Avery and Joel Cooper, "Racial Bias in Post-Arrest and Pre-Trial Decision Making: The Problem and a Solution," *Cornell Journal of Law and Public Policy* 29 (2019); Supriya Kapur, "Reducing Racial Bias in AI Models for Clinical Use Requires a Top-Down Intervention," *Nature Machine Intelligence* 3 (2021), https://doi.org/10.1038/s42256-021-00362-7.

[222] e.g., Wachter, Mittelstadt, and Russell (2021 B); Bradley Henderson, Colleen M Flood, and Teresa Scassa, "Artificial Intelligence in Canadian Healthcare: Will the Law Protect Us from Algorithmic Bias Resulting in Discrimination?," *Canadian Journal of Law and Technology* 19, no. 2 (2022); Robert P Bartlett et al., "Algorithmic Discrimination and Input Accountability under the Civil Rights Acts," *Berkeley Technology Law Journal* 36, no. 2 (2021), https://doi.org/10.15779/Z38N7XN5B; Joe Atkinson, "Automated Management, Digital Discrimination, and the Equality Act 2010," *Green's Employment Law Bulletin*, no. 159 (2020).

[223] e.g., Sikstrom et al. (2022).

[224] Kiseleva and Quinn (2021) 157.

ethical guidelines in recent years.[225] The term has also begun to appear in legislative initiatives on the regulation of AI systems in the EU, including the AI Act.[226] Accordingly, there is a growing need for conceptual clarification, particularly in a legal context.

3.2 Bias as a Multidisciplinary and Contextual Notion

3.2.1 Bias, Prejudice and Stereotyping – Lessons from Social Psychology

In social psychology, 'bias' is often equated with the somewhat more specific concepts of 'prejudice' and 'stereotyping.'[227] While typically not addressing AI technologies, the social psychological discourse on bias, prejudice and stereotyping is relevant to the AI discourse, because it concerns mechanisms that may lead to harmful behaviour, erroneous decision-making and discrimination – some of the most salient concerns in contemporary AI discourse.[228] Those concerns build on the premise that harmful behaviour and erroneous decision-making in the past can be repeated and reinforced by AI systems that rely on historical data for training.[229] In addition, human engineers partake in the development of AI systems and make decisions through which they can infect the systems with their own 'prejudices' or 'biases', e.g., in connection with various design decisions.[230] Social

---

[225] e.g., Executive Office of the President and Podesta (2014): 59; Gerards and Xenidis (2021): 41; World Health Organization (2021) 1; (FRA) (2022).

[226] Section 3.3.1 describes a proposal to prohibit biased AI that was initiated by the European Parliament before the European Commission proposed the AI Act. Provisions addressing 'bias' in the AI Act are discussed in chapter 6.

[227] According to Krieger, prejudice has been studied in psychology at least since the 1920s: Linda Hamilton Krieger, "The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity," *Stanford Law Review* 47, no. 6 (July 1995): 1174.

[228] Works that discuss unfair decision-making in specific domains, such as the criminal justice system, often rely on explanations of 'bias' and 'prejudice' from social psychology: e.g., Avery and Cooper (2019).

[229] Karen Yeung, *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework*, Council of Europe (2019), 7.

[230] Fahse, Huber, and van Giffen (2021) 96. Bias occurring as a result of various decisions during the development process is further discussed in chapter 4.

psychology provides theories on this human aspect of 'bias,' including fundamental work conceptualising these phenomena.

However, the field of social psychology seems to theorize more often over the notion of 'prejudice' than 'bias,' and the two terms are sometimes used interchangeably.[231] It is, therefore, worth also highlighting the notion of 'prejudice.' Indeed, the contemporary discourse on biases in AI systems often emphasise the reproduction of past, human prejudices as a salient side-effect of AI technologies.[232]

'Prejudice' etymologically consists of the Latin 'prae', which means *before* or *in advance,* and 'judicium,' meaning *judgment.*[233] Thus, it can be seen as referring to an advance judgment, a judgment that is made before the facts have been established.[234] Although 'prejudice' and 'bias' are sometimes used interchangeably, some authors seem to distinguish between the two concepts. For example, there are works in which it appears that 'prejudice' primarily refers to attitudes[235] held by individuals, whereas 'bias' seems to be used more widely to include also structural, institutional and group-level dynamics that work to the detriment of a group of persons.[236] Gordon Allport, in his seminal conceptual work called "The Nature of Prejudice" (1954), defines prejudice as "an antipathy based upon a faulty and

---

[231] Social psychology can be seen as "that branch of the social sciences which attempts to explain how society influences the cognition, motivation, development, and behavior of individuals and, in turn, is influenced by them": Dorwin Cartwright, "Contemporary Social Psychology in Historical Perspective," *Social Psychology Quarterly* 42, no. 1 (1979), https://doi.org/10.2307/3033880.

[232] e.g., World Health Organization (2021) 1.

[233] Merriam-Webster Dictionary, s.v. "Prejudice," under the section for 'etymology,' accessed November 7, 2023, https://www.merriam-webster.com/dictionary/prejudice.

[234] In connection with the etymological origin of the term, Blum notes that "the contemporary notion of prejudice involves an affective component as well as a judgmental one": Lawrence Blum, "Prejudice," Oxford Handbooks (Oxford University Press, 2009), 2.

[235] An 'attitude' can be defined as "an evaluative disposition – that is, the tendency to like or dislike, or to act favorably or unfavorably toward, someone or something": Anthony G. Greenwald and Linda Hamilton Krieger, "Implicit Bias: Scientific Foundations," *California Law Review* 94, no. 4 (2006): 948, https://doi.org/10.2307/20439056.

[236] See, e.g., how these terms are used particularly on pages 10-11 in Dovidio et al: John F. Dovidio et al., *The Sage Handbook of Prejudice, Stereotyping and Discrimination* (London: SAGE Publications, 2010), 10-11; See also how the terms are generally used in Rupert Brown, *Prejudice: Its Social Psychology*, 2nd ed. (Chichester: Wiley-Blackwell, 2010).

inflexible generalization."[237] Subsequent scholarship in the field modifies Allport's emphasis on feelings of *antipathy* and suggests definitions that see prejudice as any attitude, emotion or behaviour towards a group or member of a group that creates or maintains hierarchical status relations between groups.[238] Moreover, contemporary conceptualisations of 'prejudice' in social psychology combine a sociological orientation with a cognitive orientation. This way, 'prejudice' is seen as something that can be displayed by individuals as well as groups.[239]

'Stereotyping' is another concept that describes a certain attitude which is sometimes associated with socially harmful behaviour, and which can be repeated and reinforced by AI-driven and data-driven decision-making. 'Stereotyping' refers to a generalisation where some assumption is made about a group, or members of a group, based on information about other members of the group.[240] For instance, if one assumes that Norwegians are good skiers, this is a case of stereotyping.[241] As this example illustrates, not all stereotypes are inherently negative or demeaning.[242]

However, stereotyping can lead to erroneous decision-making regardless of whether the stereotyping entails a negative attitude towards a group. For example, it is stereotyping if employers assume that people who graduate with good grades are good workers.[243] The statistical evidence for this assumption may vary depending on the content of the job in

---

[237] Gordon W. Allport, *The Nature of Prejudice* (Reading, Mass: Addison-Wesley, 1954), 9.

[238] Dovidio et al. (2010) 7. Brown uses a similar definition: Brown (2010) 7; elsewhere, 'prejudice' is simply defined as any "attitude toward people based on their membership in a group", cf. "Encyclopedia of Social Psychology," (Thousand Oaks, California: SAGE Publications, Inc., 2007), under "Prejudice." https://sk.sagepub.com/reference/socialpsychology.

[239] Dovidio et al. (2010) 6. ("Despite divergent views, both psychological and sociological approaches have converged to recognize the importance of how groups and collective identities affect intergroup relations. Recent definitions of prejudice bridge the individual-level emphasis of psychology and the group-level focus of sociology by concentrating on the dynamic nature of prejudice…").

[240] Frederick Schauer, *Profiles, Probabilities, and Stereotypes* (Harvard University Press, 2006), 3-4.

[241] Similar examples are used by Schauer: Schauer (2006) 6.

[242] Sophia R. Moreau, "The Wrongs of Unequal Treatment," *University of Toronto Law Journal* 54, no. 3 (2004): 298; Tarunabh Khaitan, *A Theory of Discrimination Law* (Oxford: Oxford University Press, 2015), 54.

[243] Schauer (2006) 6.

question. For some employers, the stereotyping may be statistically sound, whereas for others it may lead to faulty assumptions and sub-optimal decision-making. Finally, it is worth highlighting that stereotyping may also be seen as a normative construct in the sense that it reflects a judgment of how members of a certain group *should* behave.[244] For example, it involves stereotyping if one asserts that men should not wear high heels.

What, then, is the difference between 'prejudice' and 'stereotyping'? The answer depends on the definitions one operates with. The term 'prejudice' is sometimes reserved for faulty[245] or statistically spurious generalisations.[246] However, 'prejudice' is also used about generalisations that are inappropriate or morally unacceptable, yet statistically sound.[247] For instance, Schauer does not find any basis for a clear distinction between 'prejudice' and 'stereotyping' and notes that, rather, there are "serious and unresolved definitional issues that surround the topic of generalization."[248] On the other hand, authors who see the presence of an antipathy as a necessary component of 'prejudice' can distinguish more easily between the two concepts. These authors may distinguish between prejudice and stereotyping by defining 'stereotyping' as something that can occur without any negative feelings, whereas the word 'prejudice' may be defined as referring to negative feelings or attitudes only.[249] Hence, Blum notionally conceptualises 'prejudice' as a judgment that can be favourable or antipathetic toward a group, but for the purposes of his writings he *defines* 'prejudice' as the combination of a "negative affect and (unwarranted) negative evaluation/judgment."[250] In comparison, he defines 'stereotyping' as a judgment based on generalisation, but it does not have to be combined with any particular affective state.[251]

---

[244] Krieger (1995) 1173-74.

[245] Allport (1954) 9.

[246]Schauer (2006) 15.

[247] Schauer (2006) 17-18.

[248] Schauer (2006) 17-18.

[249] Lawrence Blum, "Racial and Other Asymmetries," in *Philosophical Foundations of Anti-Discrimination Law*, ed. Deborah Hellman and Sophia Moreau (Oxford: Oxford University Press, 2013), 189; Blum (2009) 1.

[250] Blum (2009) 2-3.

[251] Blum (2009) 2.

Prejudice and stereotyping are *intrapsychic* phenomena, meaning that they occur within an individual.[252] This makes prejudice and stereotyping challenging concepts to deal with from a regulatory perspective. The goal of regulation is primarily to influence behaviour, not attitudes or emotions.[253] And even if influencing people's attitudes may be a regulatory aim in some cases, legal rules in democratic societies tend not to prohibit attitudes or emotions as such. Actions or statements that exhibit undesirable attitudes may be prohibited, such as in the case of hate speech, but it is nonetheless the actions and not the attitudes that are prohibited. Moreover, the feasibility of influencing someone's attitude is questionable if that someone is not aware of the attitude. Stereotypes and prejudices may not reflect an explicit belief that a person is aware of – they may be "mental association[s] between a social group or category and a trait" held by someone who is not aware of the mental association, in which case there are *implicit* stereotypes[254] or *implicit* prejudices[255] (one or both are sometimes referred to collectively as '*implicit bias*').[256]

Regardless of whether one uses the terms 'stereotyping' and 'prejudice' interchangeably or if one reserves 'prejudice' for cases of *negative* attitudes or judgments, these terms denote generalisations which can have harmful impacts if they are repeated or reinforced by AI

---

[252] Dovidio et al. (2010) 10.

[253] Consider, for example, the definition proposed by Black of 'regulation' as a "sustained and focused attempt to alter the behaviour of others…": Julia Black, "Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a 'Post-Regulatory' World," *Current Legal Problems* 54, no. 1 (2001): 142, https://doi.org/10.1093/clp/54.1.103.

[254] Greenwald and Krieger (2006) 949.

[255] See, generally: Anthony G. Greenwald et al., "A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept," *Psychological Review* 109 (2002), https://doi.org/10.1037/0033-295X.109.1.3; Calvin K. Lai, Kelly M. Hoffman, and Brian A. Nosek, "Reducing Implicit Prejudice," *Social and Personality Psychology Compass* 7, no. 5 (2013): 315, https://doi.org/https://doi.org/10.1111/spc3.12023. (defining implicit prejudices as "social preferences that exist outside of conscious awareness or control.")

[256] See, generally: Greenwald and Krieger (2006); Alexander R Green et al., "Implicit Bias among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients," *Journal of general internal medicine* 22, no. 9 (2007); Caliskan, Bryson, and Narayanan (2017); Dipesh P Gopal et al., "Implicit Bias in Healthcare: Clinical Practice, Research and Decision Making," *Future Healthcare Journal* 8, no. 1 (2021), https://doi.org/10.7861/fhj.2020-0233.

systems.[257] One type of harmful impacts may be related to the loss of accuracy that may occur when there is lack of relevant information and a decision is based on generalisations from the factors that are available to a (human or machine) decision-maker.

Whilst stereotyping is arguably rational from the viewpoint of a decision-maker trying to maximize predictive accuracy in lack of perfect information, stereotyping may nonetheless be seen as harmful to the value of human dignity, which non-discrimination law aims to promote and protect.[258] Thus, it makes sense for regulatory and legislative efforts addressing the potential harms of AI systems to address the possibility that AI systems and the data used to train them could be impacted by prejudice and stereotyping.

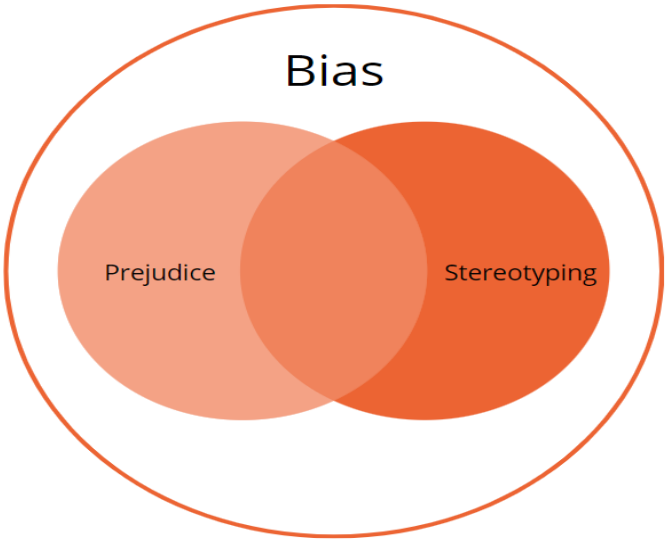


Figure 4: The figure illustrates that there is an area of overlap between the notions of prejudice and stereotyping, but also areas of meaning which are reserved for each of those terms. Bias covers both prejudice

---

[257] There is solid evidence of this happening in the criminal justice sector. For discussions of stereotyping in policing, see: David Rudovsky, "Law Enforcement by Stereotypes and Serendipity: Racial Profiling and Stops and Searches without Cause Symposium: Race Crime and the Constitution," *University of Pennsylvania Journal of Constitutional Law* 3, no. 1 (2001); On the implications of such historical stereotyping for the use of predictive algorithms in policing, see: Ferguson (2016); Kelly Blount, "Using Artificial Intelligence to Prevent Crime: Implications for Due Process and Criminal Justice," *AI & SOCIETY* (2022), https://doi.org/10.1007/s00146-022-01513-z.

[258] Section 4.2.

and stereotyping. However, bias may also have an undefined space of meaning outside of prejudice and stereotyping.

3.2.2 A Glimpse into Bias and Discrimination in Economics

Economists have extensively theorized and analysed issues of bias and discrimination in human decision-making. Within the domain of behavioural economics, a subfield that bridges psychology and economics, 'bias' often denotes a systematic pattern of flawed human reasoning.[259] This discipline particularly studies the rationality of human decision-making and the sources of irrational behaviours.[260] As an example from this field, behavioural economist Gigerenzer defines 'bias' as "a systematic discrepancy between the (average) judgment of a person or a group and a true value or norm."[261] Thus, from the vantage point of behavioural economics, decision-makers are 'biased' if their decisions deviate from the prediction of an economic model representing rationality and such deviation is attributed to systematic errors in reasoning. Here, 'systematic' implies that the error either manifests as a consistent pattern for an individual or is typical of human decision-making in general. For instance, the word 'bias' is sometimes used in economic literature to signify that a conclusion is *predetermined* by possibly flawed assumptions or by an implicit desire to arrive at a certain conclusion.[262] These preconceptions and implicit preferences exemplify common pitfalls in human reasoning.

Certain types of biases have been framed in economic literature as 'discrimination.' Writings on 'discrimination' in this field can therefore further illuminate how the field has conceptualised the 'bias' phenomenon. Economic models of 'discrimination' often distinguish between what Gary S. Becker coins "tastes for discrimination"[263] and models of

---

[259] e.g., Gerd Gigerenzer, "The Bias Bias in Behavioral Economics," *Review of Behavioral Economics* 5, no. 3-4 (2018): 306.

[260] e.g., Will Kenton, "What Is Behavioral Economics? Theories, Goals, and Applications," Toby Walters and Marcus Reeves eds. *Investopedia*, 16 January, 2023, https://www.investopedia.com/terms/b/behavioraleconomics.asp.

[261] For example, Gigerenzer (2018) 306.

[262] Edward S Herman, "The Institutionalization of Bias in Economics," *Media, Culture & Society* 4, no. 3 (1982): 278.

[263] Gary S. Becker, *The Economics of Discrimination*, 2nd ed., ed. Milton Friedman (Chicago: The University of Chicago Press, 1971), 16-17.

so-called “statistical discrimination” (introduced by Phelps, 1972 and Arrow, 1973).[264] A decision-maker is perceived as having a taste for discrimination when they are willing to make economical sacrifices in order to prioritize persons from some groups over others.[265] In comparison, statistical discrimination involves decision-making where genuine attempts are made at reaching an optimal decision. Due to absence of complete information, the decision-maker may rely on group membership when trying to reach an optimal decision.[266] For instance, a doctor in New Zealand might, based on statistical prevalence, suspect a Maori patient of having heart disease, given that here is a higher incidence of such diseases among the Maori compared to the broader population. Such an assessment, where the doctor leans on patient ethnicity, may be categorized as ‘statistical discrimination’ within the field of economics. While this term in economics does not imply a reference to non-discrimination laws, both ‘statistical discrimination’ and ‘taste-based discrimination’ are types of biases that can potentially constitute discrimination under EU non-discrimination law, depending on the circumstances.

In academic literature, it has been posited that statistical decision-making systems, including AI systems, do not “exhibit taste-based discrimination unless prejudice is explicitly programmed into them.”[267] This is rooted in the fact that ML algorithms, by their very nature, strive to identify and leverage the most predictive information available, thereby solving the task they have been assigned in an optimal manner. Should ML algorithms treat certain groups differently than others, such behaviour is prompted by the training data indicating that this is the best way of solving the given task. While there may be underlying issues with the

---

[264] Edmund S Phelps, "The Statistical Theory of Racism and Sexism," *The american economic review* 62, no. 4 (1972): 659; Kenneth J Arrow, "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. Orley Ashenfelter and Albert Rees (Princeton University Press, 1973). In subsequent literature in the field, the statistical discrimination model is usually credited to Phelps and arrow, see: Jonathan Guryan and Kerwin Kofi Charles, "Taste-Based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots," *The Economic Journal* 123, no. 572 (November 2013): 417, https://doi.org/10.1111/ecoj.12080; Stijn Baert and Ann-Sophie De Pauw, "Is Ethnic Discrimination Due to Distaste or Statistics?," *Economics Letters* 125 (2014): 270, https://doi.org/10.1016/j.econlet.2014.09.020.

[265] Becker (1971) 14.

[266] Guryan and Charles (2013) 418.

[267] Barocas, Hardt, and Narayanan (2019) 135.

data or the objective, which will be further explored in chapter 4, this differential treatment arises not from any inherent 'preference' by ML algorithms for certain groups. Rather, it is the result of optimising the path towards achieving a given objective. Conversely, human reasoning often does not lead to a course of action that strictly optimises a given objective. Behavioural economics highlight the myriad flaws of human reasoning, some of which stem from implicit or explicit preferences that extend beyond mere optimisation of the objective.[268]

3.2.3 Statistics, Computer Science, and Natural Language Processing

ML is a subset of computer science, which is the general term for sciences and practices in which information processing systems such as computers and software programs are studied.[269] In these scientific disciplines, 'bias' has been discussed from various perspectives. Most commonly, statistical sciences use the word 'bias' to refer to systematic errors (statistical estimates that miss the mark, i.e. that diverge from the value they are meant to predict).[270] The word 'bias' may also be used more generally to describe a statistical deviation from a given baseline, regardless of whether such deviation indicates an error. For instance, if half of the working population are men, but only 25 % of gynaecologists are men, this would be a statistical 'bias.'[271] In and of itself, this form of bias does not possess any inherent positive or negative connotations. It merely signifies that one statistical measurement diverges from another, which is established as the baseline.

In computer science literature oriented specifically towards AI technologies, 'bias' is often defined (more or less) in accordance with the traditional understanding of bias in statistics.[272]

---

[268] For comprehensive discussions of such flaws, see: Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011); Daniel Kahneman, Olivier Sibony, and Cass R Sunstein, *Noise: A Flaw in Human Judgment* (New York: Little, Brown Spark, 2021).

[269] "Computer Science," Geneva B. Belford (web page) accessed 28 August, 2022, https://www.britannica.com/science/computer-science.

[270] M Delgado-Rodríguez and J Llorca, "Bias," *Journal of Epidemiology and Community Health* 58, no. 8 (2004), https://doi.org/10.1136/jech.2003.008466; Cassie Kozyrkov, "What Is Bias?," *Towards Data Science*, 24 January, 2019, https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814.

[271] Danks and London (2017) 2.

[272] e.g., Shea Brown, Jovana Davidovic, and Ali Hasan, "The Algorithm Audit: Scoring the Algorithms That Score Us," *Big Data & Society* (January-June 2021): 4. (defining statistical bias as "the difference between an estimator's predicted value and the true value.")

This is not surprising, as AI typically relies on ML, which is fundamentally an advanced statistical approach. In line with typical statistical definitions, 'bias' in ML literature is sometimes defined as any discrepancy between the ideal or 'true' distribution of labels or outputs and the actual labels or outputs produced by an AI system.[273] When 'bias' is used like this, it is simply an attribute of statistical models, and it is not necessarily connected with social harm.

ML literature also uses the word 'bias' in other ways than to denote a systematic *error*.[274] Sometimes, 'bias' refers simply to the prior information or assumptions on which a learning algorithm is based.[275] As Hildebrandt explains, this type of 'bias' is inherent in ML and necessary for a predictive algorithm to produce an output at all.[276] These assumptions are necessary because the learning algorithm cannot receive complete information about the phenomenon it is trying to predict. It is therefore forced to generalise based on certain assumptions about that phenomenon. 'Bias' in this sense has been described as something as basic as "prior information, a necessary prerequisite for intelligent action."[277] ML literature typically calls this 'inductive bias.'[278] Hildebrandt proposes that this type of bias may also be called 'productive bias,' because of its necessity.[279]

When the existence of prior information or assumptions is called a 'bias,' this does not necessarily mean that the model is inaccurate in the sense that it performs worse than expected. This type of 'bias' rather refers to the fact that ML-based models are imperfect representations of reality and, in this sense, they are always 'inaccurate.' However, 'bias' in the sense of prior information or assumptions *can* lead to systematic errors and a poorly

---

[273] Hovy and Prabhumoye (2021) 2; Shah, Schwartz, and Hovy (2019) 2-3.

[274] Patrick Glauner, Petko Valtchev, and Radu State, "Impact of Biases in Big Data," *arXiv preprint arXiv:1803.00897* (2018): 1-2.

[275] Caliskan, Bryson, and Narayanan (2017).

[276] Hildebrandt (2021) 43.

[277] Caliskan, Bryson, and Narayanan (2017) 183.

[278] Mitchell (1997) 39-43; Eirini Ntoutsi et al., "Bias in Data-Driven Artificial Intelligence Systems—an Introductory Survey," *WIREs Data Mining and Knowledge Discovery* 10, no. e1356 (2020): 3, https://doi.org/https://doi.org/10.1002/widm.1356.

[279] Hildebrandt (2021) 43-44.

performing model.[280] This happens when the learning algorithm ends up relying too much on the prior assumptions and fails to learn the patterns in training data well enough.[281] In technical ML literature, this is called 'underfitting.' The term 'bias' is sometimes used synonymously with underfitting.[282]

3.2.4 Normative Definitions

Sometimes, the term 'bias' is used to refer to factual phenomena that are undesirable because they conflict with social values, legal rules, or ethical norms. Most significantly, bias is often discussed in relation to social values such as equality or fairness, and the legal and ethical norms which are associated with these values. When 'bias' is used in this way, the term 'bias' itself implies a negative evaluation: If something is 'biased,' it means that it does not adhere to the norms, standards or values it should adhere to. In other words, 'bias' is used as a *normative* notion with different value-underpinnings than the more neutral notions of bias typically referred to in technically oriented ML literature. For instance, 'bias' may defined as the amplification of existing health inequalities, in which case it is suggested by definition that bias is wrongful.[283]

In a relatively early discussion of 'bias' in computer systems (1996), computer scientist Batya Friedman and philosopher Helen Nissenbaum view computer systems as biased if they "systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others."[284] According to their definition, bias is characterised by being *systematic*

---

[280] Concerning "errors due to bias," see: Brenda Hali, "Understanding Bias-Variance Trade-Off in 3 Minutes," *Towards Data Science*, 2 December, 2019, https://towardsdatascience.com/understanding-bias-variance-trade-off-in-3-minutes-c516cb013513. Hali defines 'bias' as "simplifying assumptions made by the model to make the target function easier to approximate."

[281] Ibid.

[282] Glauner, Valtchev, and State (2018) 1-2; "The Complete Guide on Overfitting and Underfitting in Machine Learning," Avijeet Biswal (web page) accessed 7 November, 2023, https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting.

[283] Panch, Mattie, and Atun (2019) 1. ("… when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities in health systems.")

[284] Friedman and Nissenbaum (1996) 332.

and unfair.[285] This is an example of how computer scientists and/or ML professionals may define 'bias' differently from the traditional definitions used in statistical sciences, when they engage in discussions around the social and ethical harms of AI technologies.

In recent years, the subset of computer science literature that discusses ethical issues related to AI systems has often relied on various notions of 'fairness' as the benchmark for normative evaluation of bias in AI systems.[286] Particularly, so-called 'fairness metrics,' i.e. mathematical standards of 'fairness,' are meant to allow a binary classification of an AI system as 'fair' or not.[287] However, 'fairness' is itself an ambiguous term. As explained in chapter 1, this thesis does not delve into the different fairness metrics that have been highlighted in this vein of the literature. It suffices to note here that equating 'bias' with (the absence of) fairness does little to clarify the definition of the former.

The reliance on normative notions of 'bias' implies that the predictive accuracy of an AI system, per se, does not determine whether an AI system is 'biased.' An AI system may be 'biased' in a normative sense even if it is perfectly accurate, because the outcomes may be deemed as undesirable according to some norm, standard or value.[288] To illustrate this point, consider an example where accuracy in prediction does not eliminate 'bias' from a more normatively oriented position: An AI-CDS system is intended to optimize the organisation of psychiatric care. One of the target variables that the system can predict is the likelihood that a

---

[285] By "unfair discrimination" the authors mean that a system "denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate": Friedman and Nissenbaum (1996) 332.

[286] Selbst (2021) 129.

[287] Sam Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness," *Proceedings of KDD '17* (August 13-17 2017): 797, https://doi.org/10.1145/3097983.3098095; Selbst (2021) 129.

[288] Rachel KE Bellamy et al., "Think Your Artificial Intelligence Software Is Fair? Think Again," *IEEE Software* 36, no. 4 (2019): 78, https://doi.org/10.1109/MS.2019.2908514. ("Bias is such an issue because machine-learning software, by its very nature, is always a form of statistical discrimination. The discrimination becomes objectionable when it places certain groups or individuals at a systematic advantage and other groups or individuals at a systematic disadvantage.")

patient will be placed under mechanical restraint.[289] If ethnic minority patients are placed under mechanical restraint more often than other patients (e.g., due to more communication issues with personnel or due to prejudiced perceptions among personnel), the AI-CDS system is likely to predict that mechanical restraint will be needed more often for ethnic minority patients. This is a direct consequence of the situation in the real world where the training data are collected from. The AI system may therefore be said to have a high level of accuracy. Thus, there is little 'bias' in a strictly technical sense; the algorithm did a good job at learning from the data. However, this AI-CDS system reproduces an existing structural inequality that is stigmatising and may be contested, particularly from an 'equality' perspective, according to which stigma is harmful and undesirable regardless of predictive accuracy.[290]

On the other hand, consider an AI-CDS system that might have technical 'biases' without being particularly worrisome from an equality perspective: An AI-CDS system is intended to assist clinicians in the interpretation of chest X-rays for the purpose of diagnosing patients with pneumonia.[291] The system is trained on data from two different radiology sites. In the training data from site A, a larger proportion of the images are from patients who were diagnosed with pneumonia, compared to the training data from site B. At the same time, the AI-CDS system finds visual cues in the images which reveal which site each image is more likely to stem from. These visual cues may be extremely subtle and not detectable during human review of the training data. They may, for example, be caused by differences in the radiographic equipment used at the two radiological sites,[292] or by differences in how patients are positioned when the X-ray is taken. The machine learning algorithm may infer that patients from site A are more likely to have pneumonia. Consequently, the algorithm may infer that patients who had their X-ray taken at site A should be diagnosed with pneumonia

---

[289] The example is inspired by Andreas Danielsen et al., "Predicting Mechanical Restraint of Psychiatric Inpatients by Applying Machine Learning on Electronic Health Data," *Acta Psychiatrica Scandinavica* 140 (2019), https://doi.org/10.1111/acps.13061.

[290] Equality may encompass various aspects, see section 4.2.

[291] This example is based on Srishti Gautam et al., "Demonstrating the Risk of Imbalanced Datasets in Chest X-Ray Image-Based Diagnostics by Prototypical Relevance Propagation," *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (2022), https://doi.org/10.1109/ISBI52829.2022.9761651.

[292] Richard J Chen et al., "Synthetic Data in Machine Learning for Medicine and Healthcare," *Nature Biomedical Engineering* 5, no. 6 (2021), https://doi.org/10.1038/s41551-021-00751-8.

more often than patients who had their X-ray taken at site B. This constitutes a systematic inaccuracy and, thus, a 'bias' according to the traditional definition in statistical sciences. However, this bias is not relevant from an equality perspective, because the systematic difference in prediction occurs independently of each person's identity or group association.[293]

The fact that different normative viewpoints are embedded in the definitions that some authors or commentators rely on when they discuss bias in AI systems, is important to be aware of when approaching the discourse and the academic literature on this subject matter. While normative conceptualisations of bias may in theory rely on all kinds of values, chapter 4 introduces the specific normative lens of 'equality,' as the focus of the thesis sharpens. First, it is pertinent to establish a foundational working definition of 'bias' for the purposes of the thesis. To find further inspiration for such a definition, the following section turns to two different definitions of 'bias' that have been proposed in two documents that are particularly relevant to the legal analysis of bias and discrimination in AI systems.

3.3 Aspirational Definitions of 'Bias' from the European Parliament and the ISO

3.3.1 The 'Bias' Prohibition in the European Parliament's Resolution 20 October 2020

On 20 October 2020, the European Parliament presented a proposal for what it called a "framework on ethical aspects of artificial intelligence." This document, which formally constitutes a resolution of the Parliament, is hereinafter referred to as the 'AIER' (an abbreviation of the 'AI Ethics Regulation').[294] I will refer to this document in the past tense, to emphasise that it predates the Commission's AI Act proposal and that it is primarily of historical interest. The AIER was the European Parliament's own initiative for a legal framework governing the development and use of AI systems in the EU. However, it does not currently serve as a basis for the AI Act negotiations. As mentioned in section 1.1, the

[293] This assumes that the allocation of patients between site A and site B are not correlated with sociodemographic or other relevant factors.

[294] European Parliament, European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies, P9_TA-PROV(2020)0275 (2020).

Parliament's negotiation position in relation to the AI Act is found in a document referred to as the AIA Compromise Text.

At the time, the AIER proposal served both as a request from the Parliament to the European Commission for an AI regulation and as the Parliament's suggestion of what an AI regulation might contain. When the European Commission proposed the AI Act in April 2021, however, its content differs from the AIER proposal in several ways. For example, the AIER proposal explicitly prohibited discrimination and demanded that AI systems, including any software, algorithm or data used or produced by them, had to be "unbiased."[295] Thus, the Parliament proposed not only to include an explicit prohibition of discrimination but also a prohibition of 'bias' *per se*.

As a consequence of the proposed bias prohibition, the AIER proposal needed to define more precisely what would count as 'bias.' Because the AI Act does not define 'bias,' it is worth considering the definition in the AIER proposal as part of the search for a working definition of 'bias' that the remainder of this thesis can rely on, even though the AIER proposal is practically moot in terms of its legal importance. After all, it represents a significant effort by an influential legislative body to grapple with the complexities of defining 'bias' in relation to AI systems.

In the Parliament's AIER proposal, 'bias' was defined to mean "any prejudiced personal or social perception of a person or group of persons on the basis of their personal traits."[296] The requirement that an AI system, including its training data and outputs, had to be 'unbiased' thus meant that they could not contain any *prejudiced* perceptions of persons or groups.[297] The orientation towards the presence of prejudice would arguably have rendered this

---

[295] AIER proposal, Article 9.

[296] AIER proposal, Article 4(l). It should also be noted that Recital 23 to the AIER proposal suggested that AI systems should be considered 'biased' where they "display suboptimal results in relation to any person or group of persons, on the basis of a prejudiced personal or social perception…". As noted by Kiseleva and Quinn, this wording entails one element that lies closer to the typical statistical definition of bias, as it refers to a systematic inaccuracy: Kiseleva and Quinn (2021) 158. However, the bias definition in the AIER proposal still required that such an inaccuracy had to be based on a prejudiced perception.

[297] AIER proposal, Article 9(1).

definition – and prohibition – difficult to apply in practice.[298] Prejudice is an ambiguous concept that does not provide clarity in itself.[299] Moreover, as mentioned in section 3.2.1, prejudice is an intrapsychic phenomenon, which means that it would probably be challenging to determine whether an AI system is influenced by prejudice in practice.[300]

The 'bias' definition suggested by the Parliament in the AIER proposal illustrates the limitations of defining 'bias' by reference to certain *sources* of bias, particularly if those sources are explained in terms of intrapsychic phenomena. Such a definition would not function well as a working definition of 'bias' within this thesis, because the orientation towards certain sources of bias would pre-empt the discussion concerning the relevant sources of bias in AI-CDS systems in chapter 4. The importance of the different sources of bias in an assessment of discrimination, is something that the thesis considers as part of its objective to develop elements of an assessment methodology. Thus, a working definition of 'bias' should preferably pinpoint the factual phenomenon that it refers to without regard to the mechanisms causing that phenomenon to occur.

3.3.2 The ISO Standard on Bias in AI Systems

In 2021, the International Standardization Organization (ISO) released a technical standard on bias in AI systems.[301] The ISO bias standard defines several types of biases, all of which are based on one foundational definition of bias as a "systematic difference in treatment of certain objects, people or groups in comparison to others."[302] This wide definition is supplemented by more specific subcategories of bias defined in terms like 'human cognitive bias' ("bias that occurs when humans are processing and interpreting information"),[303] 'data bias' ("data properties that if unaddressed lead to AI systems that perform better or worse for different groups"),[304] and 'statistical bias' (a "type of consistent numerical offset in an estimate relative

---

[298] Kiseleva and Quinn (2021) 158.

[299] Section 3.2.1.

[300] However, it is not impossible that prejudice among persons in an AI development team can be observed as part of a discrimination assessment, cf. section 10.4.7.

[301] NEK ISO/IEC TR 24027:2021.

[302] NEK ISO/IEC TR 24027:2021, section 3.2.2.

[303] NEK ISO/IEC TR 24027:2021, section 3.2.4.

[304] NEK ISO/IEC TR 24027:2021, section 3.2.7.

to the true underlying value, inherent to most estimates").[305] These subcategories are all encompassed by the foundational definition of 'bias' that the ISO standard relies on. Due to its sweeping articulation and specific inclusion of definitions resembling the understandings of bias referred to in section 3.2, the ISO standard's foundational definition of 'bias' emerges as a promising candidate for the role as a working definition of 'bias' within this thesis.

Notably, because the foundational definition of 'bias' in the ISO standard focusses on a systematic difference in treatment, the word 'treatment' is central to the understanding of this definition. The standard therefore specifies that 'treatment' here means "any kind of action, including perception, observation, representation, prediction or decision."[306] Thus, the standard defines bias as a systematic difference in any kind of *action*, which includes intrapsychic phenomena such as a perceptions, as well as explicit actions such as predictions or decisions. This solution caters to an important conceptual challenge: When different disciplines discuss 'bias', they refer to different objects which may be 'biased.' For instance, psychologists refer to cognitive perceptions within human beings; statistical definitions refer to statistical estimates in general, including prediction models; behavioural economists refer to economic decision-making; moral philosophers and legal scholars tend to emphasize decision-making processes that impact individual persons or groups of persons. The utility of the ISO's definition of 'bias' is that it can be applied to all these objects:

- **statistical estimates and outputs** (treatment = predictions, and bias = systematic difference in predictions);
- **datasets** (treatment = representation, and bias = systematic difference in representation);
- **human reasoning and assumptions, including implicit mental associations** (treatment = perception, and bias = systematic difference in perception).

---

[305] NEK ISO/IEC TR 24027:2021, section 3.2.10. The ISO's notion of 'statistical bias' corresponds to the traditional statistical understanding of 'bias' that is observed in section 3.2.3. Compared to the ISO's specific definition of 'statistical bias', which reflects the definition of 'statistical discrimination' in economics, the general definition of 'bias' as a systematic difference in treatment is wider, because it does not require an inaccuracy or statistical offset.

[306] NEK ISO/IEC TR 24027:2021, section 3.2.2.

Moreover, the general definition of 'bias' in the ISO standard applies to biases that have positive, negative, or neutral impacts.[307] Consequently, under the ISO definition, an AI system that produces systematically different predictions for different patient groups is 'biased' according to this definition, even if the predictions are perfectly accurate. This definition is therefore wider than the technical definition often used in statistical sciences, in which bias is a systematic *error/inaccuracy*.[308]

While being wide enough to cover technical/statistical, psychological, and normative notions of 'bias', the ISO's general definition of 'bias' appears to capture the essence of how 'bias' has been conceptualised in the multidisciplinary literature referred to in section 3.2. Furthermore, the definition may work well in a text abound with references to 'discrimination,' because the definition of 'bias' does not in itself refer to 'discrimination' or other normative terms, like 'fairness,' etc.

The ISO's definition of 'bias' is relied on as a foundational working definition throughout the remainder of this thesis. Thus, 'bias' is defined in this thesis as a *systematic difference in treatment (including perception, representation, prediction, action and decision) of certain people or groups in comparison to others.*

3.4 The Relationship Between Bias and Discrimination

The many different definitions of 'bias' suggest that 'bias' means different things in different contexts. At the same time, many definitions refer to the same concept with a common origin in either social psychology, statistical sciences or economics. Understanding where this concept comes from and how it has been used and defined across different scientific disciplines is arguably key to an informed and coherent discussion of bias from a legal perspective.

Computer science literature sometimes uses 'bias' to describe the existence of prior information or assumptions or the *systematic inaccuracy of a prediction model*. In this statistical/technical sense, the only value at stake is that of predictive accuracy. However, when a predictive model is applied in different contexts, a systematic inaccuracy can impact other values such as health, safety and equality. Within this thesis, the problem of interest is

[307] NEK ISO/IEC TR 24027:2021, Section 5.2

[308] Section 3.2.3.

that biases in AI systems might disadvantage persons from certain groups in comparison with persons from other groups.[309] AI systems can have such effects even if they would be seen as highly accurate from a technical perspective.[310] This underscores the importance of clarifying whether 'bias' in a given context is meant as a reference to a strictly technical accuracy problem or whether it has normative connotations and, in the latter case, which normative connotations those are.[311]

As mentioned, this thesis continues with the ISO's definition of 'bias' as the definitional keystone. 'Bias' is understood herein as a *systematic difference in treatment (including perception, representation, prediction, action and decision) of certain people or groups in comparison to others*. This definition covers the most salient ways in which 'bias' is defined in the relevant multidisciplinary literature:

- Bias as prior information or assumptions that allow generalisations from imperfect information about the world, which is necessary to produce an output;
- Bias as an intrapsychic phenomenon where inferences are made based on a generalisation;
- Bias as a systematic error or inaccuracy without normative connotations;
- Bias as a normatively undesirable systematic error or other difference in treatment.

Starting with this wide, foundational definition, chapter 4 relies on the value of equality to clarify how certain biases that may arise in AI-CDS systems are particularly relevant to the objective of this thesis. This sharpens the focus of the thesis so that the development of

---

[309] Hale M. Thompson et al., "Bias and Fairness Assessment of a Natural Language Processing Opioid Misuse Classifier: Detection and Mitigation of Electronic Health Record Data Disadvantages across Racial Subgroups," *Journal of the American Medical Informatics Association* 28, no. 11 (2021): 2394, https://doi.org/10.1093/jamia/ocab148.

[310] Brown, Davidovic and Hasan call it "societal bias" when a systematic disadvantage occurs for one or more groups and the reason for this is societal: Brown, Davidovic, and Hasan (2021) 5.

[311] A survey of 146 papers analysing 'bias' in NLP systems illustrates the variations and calls out the lack of normative reasoning and conceptual clarity in this field. The authors call on researchers and practitioners to "articulate their conceptualizations of "bias" in order to enable conversations about what kinds of system behaviors are harmful, in what ways, to whom, and why": Su Lin Blodgett et al., "Language (Technology) Is Power: A Critical Survey of "Bias" in Nlp," *arXiv preprint arXiv: 2005:14050* (2020): 1.

methodological aspects of assessing discrimination in an AI-CDS system can concentrate on the most relevant types of bias.

Before proceeding, it is worth making some clarifications regarding the relationship between 'bias' and 'discrimination' based on the working definition of 'bias' that has been established in this chapter:

- **Bias is neutral whereas discrimination is negative.** 'Bias' denotes a systematic difference in treatment regardless of the consequences for persons or groups of persons. In comparison, discrimination presupposes that a person or group is disadvantaged or treated less favourably in comparison with others.
- **A systematic inaccuracy in a predictive model is always a bias, but it is not always discrimination.** For the systematic inaccuracy to constitute discrimination, it must lead to a particular disadvantage for a protected group (indirect discrimination) or it must cause a person to be treated less favourably than another on the basis of a protected characteristic (direct discrimination). The question of under which circumstances a systematic inaccuracy might constitute discrimination is considered particularly in Part IV of the thesis.
- **Bias can be a source of discrimination in AI systems, but discrimination by AI systems does not necessarily require that bias is found.** Bias in AI systems may cause discrimination when AI systems are used as decision-making support. The mechanisms through which this happens are explored in chapter 4. However, it should be noted that, according to the rule against direct discrimination, a single instance of less favourable treatment may constitute discrimination, whereas a single case of discrimination is not sufficient to find a 'bias', because a 'bias' is always systematic according to the definition applied in this thesis.
- **Compared to 'discrimination', the term 'bias' can be applied to a wider set of objects.** For instance, the term 'bias' lends itself to intrapsychic phenomena (e.g., prejudice), which implies that a human being may be 'biased.' In comparison, a human being cannot *be* discriminatory, because prejudiced attitudes or even racist beliefs do not constitute discrimination. Only certain actions, decisions or practices can amount to discrimination and *be* discriminatory. While the definition of 'bias' relied on herein covers perceptions and representations, mere perceptions or

representations (e.g., in a dataset) cannot constitute discrimination. A dataset may therefore be ‘biased,’ but not ‘discriminatory.’

4 Bias as an Equality Problem and Sources of Such Bias in AI-CDS Systems

4.1 Introduction

The previous chapter arrived at a working definition of 'bias' for the purposes of the thesis: Bias can be defined as a *systematic difference in treatment (including perception, representation, prediction, action and decision) of certain people or groups in comparison to others.*[312] This is a broad definition which encompasses the different facets of 'bias' highlighted in the multidisciplinary literature referred to in chapter 3.

This chapter further establishes the foundation for the analyses that are conducted in pursuit of the thesis' main objectives. It sets out with two aims in mind. First, the goal is to narrow the scope of the investigation from the broad range of biases possibly encompassed by the definition established in chapter 3 to those that are most relevant in the context of non-discrimination law. At this stage of the thesis, the narrowing of the scope is done without predetermining the outcomes of the subsequent, more detailed analysis of the non-discrimination principle in Part IV of the thesis. Therefore, this chapter relies on the concept of 'equality' to qualify the relevant biases as 'equality-related biases.' These are biases that interfere with the value of equality and which are therefore harmful from an equality perspective. They are the most important biases to consider in a pre-deployment discrimination assessment.

How the concept of equality is understood and applied within this chapter is clarified in section 4.2. below. Thereafter, section 4.3 applies the equality perspective as a lens through which certain biases in AI-CDS systems are qualified as 'equality-related' based on their interference with the value of 'equality.' Finally, section 4.4 examines the sources from which equality-related biases might arise in AI-CDS systems.

[312] The definition is based on the definition of 'bias' in NEK ISO/IEC TR 24027:2021.

4.2 Equality

4.2.1 Introduction

While bias in AI systems can lead to a range of undesirable consequences (e.g., general inaccuracy, distrust, and safety issues), the focus of this thesis is primarily on consequences related to the value of 'equality.' Although equality is a complex and contested concept,[313] it is commonly viewed as a foundational aim in the realm of non-discrimination law. While the precise objectives underlying non-discrimination law in the EU (and elsewhere) may be debated,[314] the importance of equality within this context is relatively uncontroversial. The following section elaborates on how this chapter utilises the concept of equality, drawing on some of its core aspects as widely recognised in legal discourse. Equality in the remainder of this chapter is understood as comprising five dimensions: 'formal equality,' as per section 4.2.2, and the four dimensions of 'substantive equality' proposed by Fredman, cf. section 4.2.3.

4.2.2 Formal Equality

Equality in the context of non-discrimination law can be seen as a two-faceted concept, encompassing a formal side – *formal equality* – and a substantive side – *substantive equality*. Formal equality promotes, as an ideal, formal equal treatment: The same rules and practices should be applied uniformly so that consistency of treatment is ensured.[315] Thus, like cases are treated alike,[316] regardless of individual characteristics deemed inappropriate or unjust for consideration. While there is debate around which characteristics deserve protection from an ethical or policy perspective, non-discrimination laws often specify the characteristics that are

---

[313] John Rawls, *A Theory of Justice*, revised (1999) ed. (Cambridge, Massachusetts: Harvard University Press, 1971), 86, etc; Ronald Dworkin, "What Is Equality? Part 1: Equality of Welfare," *Philosophy & Public Affairs* 10, no. 3 (Summer 1981): 185; Catherine Barnard and Bob Hepple, "Substantive Equality," *Cambridge L.aw Journal* 59, no. 3 (2000): 563-64; Sandra Fredman, "Substantive Equality Revisited," *International Journal of Constitutional Law* 14, no. 3 (2016): 720.

[314] Hervey (1993) 23.

[315] Sandra Fredman, *Discrimination Law*, 3rd ed., Clarendon Law Series, (Oxford: Oxford University Press, 2022), 251.

[316] Barnard and Hepple (2000) 562; Ellis and Watson (2012) 4; Fredman (2016 B); Anne Hellum and Vibeke Blaker Strand, *Likestillings- Og Diskrimineringsrett*, 1. utgave. ed. (Oslo: Gyldendal, 2022), 36.

protected by the respective laws.[317] Hence, in the context of non-discrimination law, formal equality means that individuals are treated equally regardless of the characteristics that the law protects. In EU non-discrimination law, formal equality is reflected by the prohibition of direct discrimination based on characteristics such as sex and ethnicity.

4.2.3 Substantive Equality

4.2.3.1 Various Dimensions of Substantive Equality

Non-discrimination laws tend to strive for equality beyond formal equality, recognising that formal equality sometimes leaves vulnerable groups at a disadvantage compared to others. For example, if only persons of a certain height are admitted to a police academy, this is a practice that does not differentiate based on sex and, thus, it amounts to formal equal treatment between men and women. However, in practice, women are more often disadvantaged by the height requirement because they are statistically less likely to meet this requirement. To address such disparities, which exist regardless of formally neutral practices, the ambitions of non-discrimination law tend to reach beyond formal equality. For example, prohibitions on indirect discrimination are intended to capture practices placing a protected group at a particular disadvantage compared to others, despite the formally neutral appearance of such practices.[318]

Various notions of equality beyond formal equality may be collected under the umbrella term of 'substantive equality.'[319] Further below, Fredman's specific framework of substantive

---

[317] Certain characteristics are widely recognised as worthy of protection in most Western societies. McColgan notes that there is "a general presumption that bare differences of "race" will almost never make people relevantly different for the purposes of the entitlement to be treated in like fashion": Aileen McColgan, *Discrimination, Equality and the Law* (Oxford: Hart Publishing, 2014), 106.

[318] e.g., Michael Connolly, *Discrimination Law*, 2nd ed. (London: Sweet & Maxwell, 2011), 10-12; Ellis and Watson (2012) 142-43.

[319] This use of the term 'substantive equality' as a reference to measures and ambitions that extend beyond formal equality or recognises the limitations of formal equality is line with the interpretation of substantive equality applied by the UN Committee on Economic, Social and Cultural Rights and the UN Committee on the Elimination of Discrimination Against Women: UN Committee on the Elimination of Discrimination Against Women (CEDAW), *General Recommendation No. 25, on Article 4, Paragraph 1, of the Convention on the Elimination of All Forms of Discrimination against Women, on Temporary Special Measures* (2004), para. 8,

equality is introduced and established as the lens that the remainder of this chapter applies. First, it is worth noting that different notions of substantive equality may place varying emphasis on certain objectives or values. Within the realm of non-discrimination law, substantive equality is often couched in terms of 'equal opportunities' and, to some extent, the promotion of 'equal outcomes.'[320]

'Equal opportunities' primarily refers to the objective of creating a 'level playing field,' ensuring that all individuals have similar chances without structural hindrances.[321] As a legislative objective, this is usually not controversial. Equal outcomes as an objective, on the other hand, often is. Equality of outcomes can be understood as aiming for a uniform distribution of the results of a decision-making process across protected groups. To illustrate, certain study programmes at Norwegian universities award additional points to male or female applicants in hopes of achieving gender balance in enrolment.[322] Thus, the objective of enrolling a certain number of persons from each sex is prioritized over the individual assessment of each applicant regardless of sex.

As an example of measures going even further in the direction of equal outcomes, publicly owned and publicly traded companies in Norway are obligated to maintain a certain gender

---

https://www.un.org/womenwatch/daw/cedaw/recommendations/General%20recommendation%2025%20(English).pdf. ("In the Committee's view, a purely formal legal or programmatic approach is not sufficient to achieve women's de facto equality with men, which the Committee interprets as substantive equality."); Social and Cultural Rights (CESCR) UN Committee on Economic, Social and Cultural Rights, *General Comment No. 20: Non-Discrimination in Economic, Social and Cultural Rights* (2009), para. 8, https://www.globalhealthrights.org/instrument/cescr-general-comment-no-20-non-discrimination-in-economic-social-and-cultural-rights/.

[320] Fredman (2016 B) 720; According to the CEDAW, equality of outcomes is "the logical corollary" of substantive equality: UN Committee on the Elimination of Discrimination Against Women (CEDAW) (2004): para. 9.

[321] Fredman (2022) 302.

[322] Regulation 6 January 2017 No. 137 on Admission to Higher Education; Anne Hellum, Ingunn Ikdahl, and Vibeke Blaker Strand, "Between Norms and Institutions: Unlocking the Transformative Potential of Norwegian Equality and Anti-Discrimination Law," in *Nordic Equality and Anti-Discrimination Laws in the Throes of Change: Legal Developments in Sweden, Finland, Norway, and Iceland*, ed. Anne Hellum et al. (New York: Routledge, 2024), 151.

balance in their boards.[323] Relatedly, at the EU level, Directive (EU) 2022/2381 obliges EU Member States to ensure (by 2026) that women hold certain percentages of board and director positions in listed companies.[324] However, provisions requiring a certain representation of a protected group may be controversial because they may involve formally different treatment of individuals.[325] To facilitate such differential treatment, non-discrimination laws may include specific rules on the extent to which so-called 'positive action' is permitted (i.e., specific measures to prevent or compensate for existing disadvantages).[326] The extent of permitted positive action may vary between EU Member States.[327]

In addition to an emphasis on equal opportunities and equal outcomes, substantive equality is often seen as encompassing the protection of fundamental values such as personal freedom, dignity,[328] and integrity.[329]

4.2.3.2 Substantive Equality as Allocational and Representational Harms

As a potential understanding of substantive equality in the specific context of AI technologies, the notions of allocational and representational harms could be considered. In the multidisciplinary, contemporary discourse on AI bias as a societal issue, the notions of

---

[323] Act 13 June No. 44 on Limited Liability Companies (*aksjeloven*), § 6-11 a; Act 13 June No. 44 on Public Limited Companies (*allmennaksjeloven*), § 20-6.

[324] Directive of the European Parliament and of the Council on Improving the Gender Balance among Directors of Listed Companies and Related Measures (Directive (Eu) 2022/2381).

[325] Hellum, Ikdahl, and Strand (2024) 149.

[326] Article 5 RED; Article 6 GSED; Colm O'Cinneide, "Positive Action and Eu Law," (2011): 6-7. https://www.era-comm.eu/oldoku/SNLLaw/04_Positive_action/2011-111DV20-O'Cinneide_EN.pdf.

[327] There has been some debate in Norway regarding whether Norway's extensive use of positive action in the education sector is compatible with EU law: Hellum, Ikdahl, and Strand (2024) 152.

[328] Fredman (2016 B) 713; Denis G. Réaume, "Dignity, Equality, and Comparison," in *Philosophical Foundations of Discrimination Law*, ed. Deborah Hellman and Sophia R. Moreau (Oxford: Oxford University Press 2013), 8.

[329] Hellum and Strand observe a development in non-discrimination law thinking in which equality is being more tightly interwoven with dignity: Hellum and Strand (2022) 39; Storvik frames the right to non-discrimination as a protection of mental integrity: Marius Storvik, "Rettslig Vern Av Pasienters Integritet I Psykisk Helsevern" (PhD UiT Norges arktiske universitet, 2017), 19.

allocational (or distributive[330]) and representational harms have gained traction as a framework of understanding why biases may be harmful.[331] Allocational harms can be understood as the most direct negative consequences of a decision or prediction; goods and opportunities are withheld for certain groups and thus allocated in a manner that contradicts equality.[332] In healthcare, people experience allocational harms for example when they are denied treatment because they are misdiagnosed or because they are not prioritised. Inequality in the allocation of goods, resources and opportunities is among the chief allocational concerns in contemporary AI bias discourse.[333]

Representational harms are the less tangent consequences of AI bias.[334] Crawford explains the label of representational harms as referring to diffuse effects that may not be immediately or concretely noticeable.[335] These harms may include the representation of certain groups in an unfavourable manner compared to other groups, or representation of certain groups in a manner that is demeaning or "fails to recognize their existence altogether."[336]

While influential on the AI bias discourse, the notions of allocational and representational harms do not add new dimensions to the existing understandings of substantive equality.

---

[330] Reuben Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," *Proceedings of Machine Learning Research* 81, no. 1 (2018): 8.

[331] Kate Crawford, "The Trouble with Bias - Nips 2017 Keynote," (2017). https://www.youtube.com/watch?v=fMym_BKWQzk; Blodgett et al. (2020).

[332] Crawford, "The Trouble with Bias - Nips 2017 Keynote."; Blodgett et al. (2020); Dobbe et al. argue that focussing only on models that "inherit pre-existing biases from training data" is too narrow because it only considers allocative harms while neglecting representational harms: Roel Dobbe et al., "A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics," *arXiv preprint arXiv:1807.00553* (2018): 1; Unfair allocation of goods/resources and opportunities is also discussed in by Barocas, Hardt and Narayanan: Barocas, Hardt, and Narayanan (2019) chapter 4.

[333] Brent Mittelstadt, Sandra Wachter, and Chris Russell, "The Unfairness of Fair Machine Learning: Levelling Down and Strict Egalitarianism by Default," *arXiv preprint arXiv:2302.02404* (2023): 19; generally: Lebret (2023).

[334] As Binns summarises, "the problem is not necessarily one of specific harms to specific members of a social group, but rather one of the way in which certain groups are represented in digital cultural artefacts": Binns (2018) 9.

[335] Crawford, "The Trouble with Bias - Nips 2017 Keynote."

[336] Blodgett et al. (2020).

Fredman's four dimensions of equality, which are introduced below, constitute a more complex framework that facilitates a more nuanced explanation of how bias in AI-CDS systems can be seen as an equality-related problem. This is therefore the primary analytical framework relied on in this chapter. However, the distinction between allocational and representational harms is referred to at several instances throughout the thesis, as it provides an effective way of distinguishing between two main categories of harms associated with biased AI systems.

4.2.3.3 Fredman's Four-Dimensional Framework

In the light of the elusive and multifaceted nature of the concept of substantive equality, Fredman has developed an analytical framework to "assist in determining whether actions, practices or institutions impede or further the right to equality."[337] While Fredman refers to "the right to equality" and discusses the content of such a right at the level of international human rights law, the framework is suitable also for the purpose of this chapter, which is to specifically qualify and examine equality-related biases in AI-CDS systems.

Fredman's analytical framework consists of four dimensions, stipulating the objectives encompassed by substantive equality:

- to redress disadvantage;
- to address stigma, stereotyping, prejudice, and violence;[338]
- to enhance voice and participation; and
- to accommodate difference and achieve structural change.[339]

These four dimensions of substantive equality are relied on in this chapter as an analytical framework for understanding how certain biases in AI-CDS systems could interfere with the value of equality. When such a 'lens' of equality is applied, this necessarily means that it is my interpretation of Fredman's framework that is applied. The following elaborates on my understanding of the respective dimensions of substantive equality.

---

[337] Fredman (2016 B) 713.

[338] The violence element is not particularly relevant in the context of AI-CDS systems and is not further considered here.

[339] Fredman (2016 B) 713 and 27.

The first dimension – to redress disadvantage – means to take into account the fact that different groups have different starting points. Some groups are disadvantaged in comparison to others, for reasons that they are not responsible for, such as hierarchical power asymmetries in society, socio-economic disadvantages, and under-representation.[340] This dimension of substantive equality resonates strongly with the aims underpinning EU non-discrimination law, as manifested particularly in the prohibition of indirect discrimination.

The second dimension of Fredman's framework revolves around addressing stigma,[341] stereotyping, prejudice, and violence. This calls for respect for the inherent worth of every individual and ensuring that they are recognised as individuals rather than being treated based on generalised group characteristics and assumptions. The right to be recognised as an individual and treated on the basis of one's own merits has been described by some scholars as the core of non-discrimination law.[342] Within the purview of US anti-discrimination law, formal inequality of treatment is often seen as inflicting a 'classificatory harm,' which suggests that the classification of a person according to some characteristic can be seen as harmful in itself (anti-classification theory).[343] The notion of classificatory harms dovetails the second dimension of Fredman's substantive equality framework. In EU non-discrimination law, the counteraction of stereotyping and prejudices is an important aim.[344] At the level of international human rights law, the issues of stereotyping and prejudices have particularly been discussed in the context of the UN Convention on the Elimination of Discrimination Against Women (CEDAW), which explicitly aims for the "elimination of

---

[340] Fredman (2016 B) 729;John E. Roemer and Alain Trannoy, "Chapter 4 - Equality of Opportunity," in *Handbook of Income Distribution*, ed. Anthony B. Atkinson and François Bourguignon (Elsevier, 2015), 217-18. ("Equality of opportunity exists when policies compensate individuals with disadvantageous circumstances so that outcomes experienced by a population depend only on factors for which persons can be considered to be responsible").

[341] Iyiola Solanke, *Discrimination as Stigma: A Theory of Anti-Discrimination Law* (Oxford, England,Portland, Oregon: Hart Publishing, 2017), 39.

[342] This aspect is sometimes conceptualised as a matter of 'dignity': Moreau (2004) 294-95; Fredman (2016 B) 724; Hellum and Strand (2022) 30.

[343] Barocas and Selbst thus argue that "considering membership in a protected class as a potential proxy is a legal classificatory harm in itself": Barocas and Selbst (2016) 695; Fredman (2016 B) 718-19;

[344] Hepple describes the counteraction of stereotyping as one of the main purposes of non-discrimination law: B. A. Hepple, *Equality: The New Legal Framework* (Oxford: Hart, 2011), 55-56.

prejudices [….] which are based on the idea of the inferiority or the superiority of either of the sexes or on stereotyped roles for men and women."[345]

Stereotyping and prejudice have been described in chapter 3 as intrapsychic phenomena. These intrapsychic phenomena may be harmful from an equality perspective even though they do not always constitute unlawful discrimination. Moreover, 'stigmatisation' occurs when a "generalised negative message about the group" is spread in society.[346] Stigma thus refers to a tendency to categorise and generalise that is more deeply rooted in societal structures, compared to stereotyping and prejudice.[347] Stigmatisation is recognised as a harmful effect of AI bias, because AI systems can produce information that contributes to the stigmatisation of a group.[348]

The third dimension – enhancing voice and participation – is centred on promoting social inclusion and political participation as facets of substantive equality.[349] This dimension actively counters the marginalisation of minority groups, a key focus in equality and non-discrimination law scholarship.[350] Marginalisation occurs when groups that lack influence, representation, or participation face further erosion in these regards. When voice and participation diminish, this means that there is a loss of agency – the interests and perceptions of marginalised individuals and groups will have minimal influence on decision-making processes and the shaping of society. In my understanding of Fredman's substantive equality framework, agency and influence are central values underpinning the 'participation' dimension of discrimination. In the context of AI technologies, agency and influence diminish when certain groups, which may be affected by the use of these technologies in practice, are

---

[345] Article 5(a) CEDAW; Simone Cusack, "The Cedaw as a Legal Framework for Transnational Discourses on Gender Stereotyping," in *Women's Human Rights: Cedaw in International, Regional and National Law*, ed. Anne Hellum and Henriette Sinding Aasen, Studies on Human Rights Conventions (Cambridge: Cambridge University Press, 2013).

[346] Solanke (2017) 21.

[347] Ibid.

[348] Harini Suresh and John V Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," *arXiv preprint arXiv:1901.10002* (2019); Schwartz et al. (2022) 2.

[349] Fredman (2016 B) 731-732.

[350] e.g., Ingunn Ikdahl, "Securing Women's Homes: The Dynamics of Women's Human Rights at the International Level and in Tanzania" (PhD thesis, University of Oslo, 2010), 67.

not involved in development processes or decisions on deployment, or in shaping the norms that guide development and deployment.

In addition to agency and influence, recognition can be understood as another important component of the third dimension of Fredman's equality framework. In non-discrimination law scholarship, 'recognition' generally refers to the "importance of inter-personal affirmation to our sense of who we are."[351] Recognition is arguably of particular interest in the context of AI technologies: In the discourse on the harms of biases in AI systems, Crawford highlights the potential for 'harms of recognition.' Such harms occur when a group is made invisible by an AI system, such as when facial recognition systems fail to recognise dark skin tones.[352] If the development of AI-CDS systems fail to recognise minority groups, this could entail marginalisation and impinge on voice and participation. Lack of recognition, such as in the case of facial recognition, results in loss of representation of a group in training data. In the context of machine learning, the lack of representation in training data is arguably tantamount to not participating in the shaping of a decision model. A machine learning algorithm is likely to produce a model that is better adapted to the groups that are well-represented in the training data.[353]

The fourth dimension – accommodating difference and effecting structural change – promotes the interests of those at risk of being overshadowed when societies prioritise the desires and lifestyles of the majority. This dimension of substantive equality thus alludes to what non-discrimination law scholars often label the 'transformative' aspect of non-discrimination law.[354] 'Transformative,' in this context, refers to non-discrimination law's call to restructure practices and societal norms, particularly those which cater predominantly to what has been

---

[351] Fredman (2016 B) 730-31.

[352] Crawford, "The Trouble with Bias - Nips 2017 Keynote."

[353] Further reasons why such underrepresentation might occur are discussed in section 4.4.2.

[354] Colm O'Cinneide, "Fumbling Towards Coherence: The Slow Evolution of Equality and Anti-Discrimination Law in Britain Special Issue on Human Rights and Equality," *Northern Ireland Legal Quarterly* 57, no. 1 (Spring 2006): 62; Bob Hepple, "The Equality Agenda," in *The Future Regulation of Work*, ed. Nicole Busby, Douglas Brodie, and Rebecca Zahn (London: Palgrave Macmillan, 2016), 51.

deemed ‘normal’ or ‘average.’ Such transformation requires that the underlying causes of inequality are addressed.[355]

The extent to which non-discrimination laws have and should have a transformative ambition in practice, is debated.[356] Laws prohibiting indirect discrimination may to some extent be perceived as accommodating difference and promoting structural transformation.[357] However, it should be noted that the transformative capacity of indirect discrimination laws is limited.[358] In contrast, laws subjecting certain entities to proactive ‘activity duties’ aimed at preventing discrimination, which section 7.5.2 returns to, have a stronger transformative aspect.

In the following examination of bias in AI-CDS systems as an equality problem, the focus is on equality-related impacts that interfere with the four dimensions in Fredman’s substantive equality framework. Due to the attention given to sex and ethnicity as discrimination grounds in this thesis, the following also centres on equality with regard to these characteristics. Before proceeding, however, it is worth sharpening the equality lens further by reflecting on how equality may be understood specifically within the context of healthcare.

4.2.4 Equality (and Equity) in Healthcare

First of all, it is important to note that the use of the term ‘equality’ in this thesis may differ from how the term is used in non-legal literature pertaining to ‘health equalities’ and ‘health inequities.’ Such literature sometimes distinguishes between ‘equality’ and ‘equity,’ typically defining ‘health inequalities’ as differences in the health of individuals or groups without

[355] UN Committee on the Elimination of Discrimination Against Women (CEDAW) (2004): para. 10.

[356] O'Cinneide (2006) 64. (“Serious disagreement also exists as to the appropriate limits to [equality law’s] transformative ambitions and scope of application.”); Hepple (2016) 51.

[357] Blaker Strand emphasises that protection from indirect discrimination allows people to be different: Vibeke Blaker Strand, "Diskrimineringsvernets Rekkevidde I Møte Med Religionsutøvelse" (PhD thesis, Universitetet i Oslo, 2011), 99; Ellis and Watson (2012) 142.

[358] However, Ellis and Watson emphasise that the concept of indirect discrimination is limited when it comes to achieving substantive equality through structural transformation: Ellis and Watson (2012) 468.

"any moral judgment on whether observed differences are fair or just."[359] In contrast, 'health inequities' are defined as health differences that could and should be prevented or mitigated because allowing them to exist is morally unacceptable.[360] This thesis does not distinguish between equality and equity. Importantly, equality is used here in a normative sense, as is common within legal research: Equality is a political and legislative aim and, thus, something that ought to be pursued. Its counterpart – inequality – is undesirable and worth mitigating through legislation, among other means. Non-discrimination law is based on the idea that the elimination of discrimination will contribute to the promotion of equality.

Given that this thesis refers to the value of 'equality' and not 'equity,' what might equality mean specifically in the context of healthcare? One potential answer is to provide the same treatment to persons with equal needs.[361] Another answer is equality in *access to healthcare*,[362] which is recognised by the UN Committee on Economic, Social and Cultural Rights (UNCESCR) as an aspect of the right to health.[363] Equality in access to healthcare can be defined as "provision of medical treatment that does not vary in scope or quality because of groups' characteristics or particular aspects of individuals composing these groups."[364] The ease or timeliness of access, or the quality or adequacy of services should not vary depending on who the patient is.[365] Equality in access to healthcare, in this sense, requires abidance by

---

[359] Mariana C. Arcaya, Alyssa L. Arcaya, and S. V. Subramanian, "Inequalities in Health: Definitions, Concepts, and Theories," *Global Health Action* 8, no. 27106 (2015): 2, https://doi.org/10.3402/gha.v8.27106.

[360] Ibid.

[361] A. J. Culyer and Adam Wagstaff, "Equity and Equality in Health and Health Care," *Journal of Health Economics* 12, no. 4 (1993): 433, https://doi.org/10.1016/0167-6296(93)90004-X.

[362] Culyer and Wagstaff (1993) 431.

[363] Social and Cultural Rights UN Committee on Economic, *Cescr General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12)* (11 August 2000), para. 19, https://www.refworld.org/pdfid/4538838d0.pdf.

[364] Marcin Orzechowski et al., "Social Diversity and Access to Healthcare in Europe: How Does European Union's Legislation Prevent from Discrimination in Healthcare?," *BMC Public Health* 20, no. 1399 (2020): 2, https://doi.org/10.1186/s12889-020-09494-8.

[365] Niall Crowley, *Equality, Diversity and Non-Discrimination in Healthcare: Learning from the Work of Equality Bodies*, European Network of Equality Bodies (EQUINET) (2021), 8, https://equineteurope.org/publications/equality-diversity-and-non-discrimination-in-healthcare-learning-from-the-work-of-equality-bodies/.

the more general values of formal equality of treatment irrespective of inappropriate characteristics as well as substantive equality, which encompasses equality of opportunities.

In contrast, equality of outcomes may be a more controversial aspect of equality in healthcare. In its most extreme sense, equality of health outcomes could mean that everyone should be equally healthy, which is an impossible proposal. Another extreme and unrealistic version of equal outcomes in a healthcare context would be that health resources such as medical treatments are equally distributed across groups. However, this is not feasible because diseases and other health related conditions are not distributed equally.

While the aforementioned, absolute interpretations of outcome equality in healthcare may be disregarded, equality of outcomes is not an irrelevant aspect of equality in healthcare. One example of its influence is where specific measures are implemented to address so-called 'social determinants of health.' According to the WHO's Commission on Social Determinants of Health (2005-2008), the social determinants of health are the "structural determinants and conditions of daily life," i.e., factors influencing people's health which are connected with the "circumstances in which people grow, live, work, and age, and the systems put in place to deal with illness."[366] If a decision or practice is aimed at addressing health problems that are related to these determinants, one might say that there is an aspect of outcome equality involved. Such measures do not focus on providing equal access to healthcare for patients with equal needs. Rather, they aim at providing healthcare specifically adapted to the social determinants of health, to increase the chances that patients affected by these determinants may have similar health outcomes to other patients.

4.3 AI Bias as an Equality Problem

4.3.1 Equality-Related Biases and Other Biases

In chapter 3, a broad, foundational working definition of 'bias' was established. Section 4.2 established an understanding of the value of equality, which consists of five dimensions: formal equality (section 4.2.2) and four dimensions of substantive equality (section 4.2.3).

---

[366] Commission on Social Determinants of Health, *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health (Final Report/Executive Summary)*, World Health Organization (Geneva, Switzerland: WHO Press), 1-2, https://iris.who.int/bitstream/handle/10665/69832/WHO_IER_CSDH_08.1_eng.pdf?sequence=1.

The present section now applies this equality perspective as a means of qualifying certain types of biases as 'equality-related' biases, thus identifying and categorising the main ways in which biases in AI-CDS systems could interfere with the value of equality.

4.3.2 Inappropriate Use of Personal Characteristics

One salient concern in the AI bias discourse is that AI systems might rely on personal characteristics in circumstances where society would not deem it appropriate to rely on those characteristics, regardless of whether they are predictive of the outcome the system is trying to predict.[367] Because inappropriate use of personal characteristics would lead to a systematic difference in treatment of groups that possess the characteristics relied on, this can be seen as a type of 'bias' according to the definition established in chapter 3.

An AI system that relies inappropriately on personal characteristics would contradict formal equality because there would be differential treatment based on such characteristics. It lies at the core of formal equality that people should be treated according to their individual qualifications, merits or needs, regardless of characteristics that society deems irrelevant or inappropriate in any particular context.[368] Furthermore, if patients are treated based on personal characteristics when such treatment is not medically necessary, this might contribute to stigma, stereotyping and prejudice, negatively affecting the second dimension of Fredman's substantive equality framework. In addition to these representational harms, allocational harms – impacting the disadvantage dimension of substantive equality – are evident if a person is denied a resource because an AI-CDS system relies on personal characteristics that should not be relied on.

4.3.3 Unequal Standards of Service (Disparate Performance)

The concern that is perhaps voiced most often in the societal discourse on bias in AI systems relates to the potential for unequal standards of service between different groups, due to

---

[367] House of Commons Science and Technology Committee, *Algorithms in Decision-Making* (23 May 2018), 21, https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/35104.htm; Crowley (2021): 8.

[368] Fredman (2016 B) 718-19; McColgan (2014) 111.

systematic performance variations in AI systems.[369] At a technical level, AI systems can display 'disparate accuracy' between groups. The performance, in terms of accuracy, typically drops for minority groups and groups that have been marginalised in the past. The harms of disparate accuracy manifest when AI models are used in the provision of important services, such as healthcare services, leading to unequal access to high-quality healthcare (to borrow an expression from Mittelstadt).[370]

From an equal opportunity perspective, patients who have coronary artery disease should have an equal chance of being diagnosed with the disease. It contradicts the disadvantage dimension of substantive equality if some groups have a lesser chance of receiving the correct diagnosis and the appropriate type of treatment. The third dimension of substantive equality – voice and participation – is affected because disparate accuracy to the detriment of marginalised groups could reinforce existing health disparities, further disempowering and marginalising these groups.[371] Furthermore, disparate accuracy is a reflection of structural inequalities which are to be transformed according to the fourth dimension of substantive equality (to accommodate difference and achieve structural change).

4.3.4 Repetition or Reinforcement of Stereotypes and Prejudice

The risk of reproducing and/or reinforcing historical stereotyping and prejudice is among the most salient equality-related concerns in the AI bias discourse. For example, algorithms used in the US for predicting the likelihood that a person charged with a criminal offense will commit a future crime are known to reproduce stereotyped attitudes of police officers in the past, causing black offenders to be assessed as more likely to commit future crimes by the

---

[369] Hardt (2014); Barocas and Selbst (2016) 720; Latrice G Landry and Heidi L Rehm, "Association of Racial/Ethnic Categories with the Ability of Genetic Tests to Detect a Cause of Cardiomyopathy," *JAMA cardiology* 3, no. 4 (2018), https://doi.org/10.1001/jamacardio.2017.5333; Paulus and Kent (2020) 1; Hovy and Prabhumoye (2021) 2.

[370] Brent Mittelstadt, *The Impact of Artificial Intelligence on the Doctor-Patient Relationship*, Council of Europe (December 2021), 44, https://rm.coe.int/inf-2022-5-report-impact-of-ai-on-doctor-patient-relations-e/1680a68859.

[371] Alvin Rajkomar et al., "Ensuring Fairness in Machine Learning to Advance Health Equity," *Annals of Internal Medicine* 169, no. 12 (2018), https://doi.org/10.7326/M18-1990; Sikstrom et al. (2022) 1.

algorithm.[372] The reproduction of negative stereotypes and prejudiced attitudes conflicts with the value of human dignity and may cause stigmatisation.[373] When persons or groups of are subjected to stereotypes or prejudice, this is considered to be harmful in itself (the stereotyping/prejudice dimension of substantive equality) regardless of materialised negative consequences for individual persons (the disadvantage dimension of substantive equality).[374] These harms of AI technologies interfere with the second dimension of substantive equality in Fredman's framework. They are representational harms that may arise in addition to allocational harms when stereotyping or prejudice is involved in decision-making.

4.4 Sources of Equality-Related Biases in AI-CDS Systems

4.4.1 Introduction

When assessing discrimination in an AI-CDS system it is helpful to have an understanding of how equality-related biases – potentially constituting discrimination – may arise in these systems. Consequently, such knowledge also underpins the methodological developments in this thesis. If certain biases (the equality-related ones) are potential causes of discrimination, an assessment of discrimination needs to build on knowledge of where these biases typically come from, so that an assessor may consider whether such sources of bias might have influenced the AI-CDS system being assessed. The further implications of finding that an AI-CDS system is influenced by different sources of bias, are explored in Part IV. At this point, it is worth noting that where a bias comes from is a relevant consideration at several instances of a pre-deployment discrimination assessment, according to the methodological elements developed in Part IV.[375] Moreover, understanding the underlying sources of bias is crucial to

---

[372] Julia Angwin et al., "Machine Bias," (23 May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[373] Moreau (2004) 297; Alexandra Timmer, "Toward an Anti-Stereotyping Approach for the European Court of Human Rights," *Human Rights Law Review* 11, no. 4 (2011): 709, https://doi.org/10.1093/hrlr/ngr036; Réaume (2013) 12; Fredman (2016 B) 726.

[374] Crawford, "The Trouble with Bias - Nips 2017 Keynote."; Anna Nilsson, "Same, Same but Different: Proportionality Assessments and Equality Norms," *Oslo Law Review* 7, no. 3 (2020): 139, https://doi.org/10.18261/ISSN.2387-3299-2020-03-01; Hovy and Prabhumoye (2021) 2.

[375] Examples from Part IV: Considering whether a bias is caused by hidden inferences matters to the distinction between direct and indirect discrimination, cf. section 8.6.5; If bias is caused by

an AI provider or deployer's implementation of targeted measures to mitigate or prevent biases in AI-CDS systems.

The following discusses and categorises equality-related biases in AI-CDS systems according to where the biases stem from. Four main categories of bias are identified: data bias, bias and processing and modelling choices, hidden inferences, and deployment bias. The two former categories are the largest, encompassing several subcategories of biases stemming from various sources. The purpose of this categorisation is to provide a frame of reference for the subsequent analyses carried out in this thesis. There is no claim made that these categories are exhaustive or that they represent the only appropriate way to categorise biases in AI-CDS systems. In fact, there may be partial overlaps and blurry lines between the different categorises proposed in the following. However, these categories represent the sources of equality-related bias that are most important to consider in a pre-deployment discrimination assessment, based on the multidisciplinary materials on which this chapter relies.

The five dimensions of equality established in section 4.2 are relied on to inform the identification of biases that are relevant from the equality perspective applied in this chapter. The following searches for sources of equality-related biases, rather than other biases.

4.4.2 Bias in Training Data (Data Bias)

4.4.2.1 Introduction

In the literature on algorithmic discrimination, 'biased' training data is often seen as the most prominent source of equality-related biases in algorithms and models.[376] In accordance with the definition of training data in the AI Act, 'training data' is used in the following as a

---

stereotyping or prejudice, this indicates direct causation, cf. section 10.4.7; Where bias comes is a relevant consideration when assessing causation and objective justification in relation to indirect discrimination, cf. section 10.5.2.

[376] Bart Custers, "Data Dilemmas in the Information Society: Introduction and Overview," in *Discrimination and Privacy in the Information Society* (Springer, 2013), 3-4; Barocas and Selbst (2016) 680-81; Hacker 2018 p 1146; Hildebrandt (2021) 2; Alessandro Castelnovo et al., "A Clarification of the Nuances in the Fairness Metrics Landscape," *Scientific Reports* 12, no. 4209 (2022): 3, https://doi.org/10.1038/s41598-022-07939-1; European Parliamentary Research Service: Study Panel for the Future of Science and Technology (2022): 20.

general term for training, validation and testing data.[377] However, data bias is a general term that requires further explanation and delimitation. Data (information) is a "product of many factors".[378] It is always a result of the context in which it is generated, collected and processed. What a learning algorithm learns depends on the examples to which it is exposed during training.[379] If labelled examples used in supervised learning are biased, the AI model will also be biased. Therefore, the mechanisms that may lead to bias in training data are important sources of bias in AI systems.

Machine learning typically uses authentic, historical data sets of relevance to the predictive task that the developer wants to make a decision model for.[380] Depending on the task that the system is going to perform, there are many different sources and types of data that can be used to train ML algorithms. The training data can include many different types of data, such as genetic information, lab values, clinical notes and time series (i.e., a sequence of time-ordered information from repeated or continuous measurements, such as medical examinations by electrocardiogram).[381] Some data types may be structured in neatly organised tables, as is typical for information in electronic health records about medical diagnoses, medications taken by a patient, laboratory test results, and sociodemographic characteristics.[382] Other data may be unstructured, such as free text clinical notes which may describe a patient's medical history, symptoms, life situation, etc.[383]

Machine learning may also rely on synthetic data, which raises further concerns about synthetic data bias, but those particular concerns are not addressed in this thesis due to the

---

[377] Section 1.5.3.

[378] Suresh and Guttag (2019) 1.

[379] Barocas and Selbst (2016) 680.

[380] Section 1.5.3.

[381] A. Anguera et al., "Applying Data Mining Techniques to Medical Time Series: An Empirical Case Study in Electroencephalography and Stabilometry," *Computational and Structural Biotechnology Journal* 14 (2016), https://doi.org/10.1016/j.csbj.2016.05.002; Kline et al. (2022) 2.

[382] Peter B Jensen, Lars J Jensen, and Søren Brunak, "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care," *Nature Reviews Genetics* 13, no. 6 (2012): 395, https://doi.org/10.1038/nrg3208; Jessica Irving et al., "Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk," *Schizophrenia Bulletin* 47, no. 2 (2020): 406, https://doi.org/10.1093/schbul/sbaa126.

[383] Irving et al. (2020) 406.

need to limit the scope of relevant material/literature and the insignificant role of synthetic training data in the development of AI-CDS systems at present.

The following describes various ways in which data bias with potential repercussions for the value of equality may arise.

4.4.2.2 Structural Inequalities in Healthcare (Historical Bias)

In literature on bias in AI technologies, historical inequalities are sometimes referred to as 'historical bias'[384] or 'systemic bias.'[385] Suresh and Guttag define historical bias as "a normative concern with the world as it is; it is a fundamental, structural issue with the first step of the data generation process and can exist even given perfect sampling and feature selection."[386] I will return to the issues of sampling and feature selection below. As the quote shows, the idea of historical bias is based on the notion that data that perfectly reflects the real world may be seen as biased if there is inequality in the reality that is reflected by the data. For example, a decision model based on historical health data might learn to predict worse outcomes for groups that have a disadvantageous starting point due to historical disparities.[387]

---

[384] Hacker (2018) 1148; Suresh and Guttag (2019) 2; Ninareh Mehrabi et al., "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 8, https://doi.org/10.1145/3457607; Friedman and Nissenbaum use the term 'preexisting bias': Friedman and Nissenbaum (1996) 332.

[385] Schwartz et al. (2022) 6. Hacker uses the term 'unequal ground truth': Hacker (2018) 1146.

[386] Suresh and Guttag (2019) 2.

[387] Matthew DeCamp and Charlotta Lindvall, "Latent Bias and the Implementation of Artificial Intelligence in Medicine," *Journal of the American Medical Informatics Association* 27, no. 12 (2020): 2021, https://doi.org/10.1093/jamia/ocaa094. ("For example, an algorithm to predict patient mortality or an individual patient's response to particular treatments could learn from existing racial, ethnic, and socioeconomic disparities in care and predict worse outcomes for those patients"); Charlotta Lindvall et al., "Ethical Considerations in the Use of AI Mortality Predictions in the Care of People with Serious Illness," *Health Affairs Blog*, *Health Affairs*, 2020, https://www.healthaffairs.org/do/10.1377/hblog20200911.401376/full/?MessageRunDetailID=3353581596&PostID=19618763&af=R&content=blog&mi=3egtxy&sortBy=Earliest&target=do-blog; Rahuldeb Sarkar et al., "Performance of Intensive Care Unit Severity Scoring Systems across Different Ethnicities in the USA: A Retrospective Observational Study," *The Lancet Digital Health* 3, no. 4 (2021), https://doi.org/10.1016/S2589-7500(21)00022-4.

If such a model is implemented in an AI-CDS system that supports the allocation of scarce resources based on a combination of each patient's needs and expected benefit, this could lead to an allocation of resources that perpetuates existing health disparities. This might be an AI-CDS system for allocating intensive care beds, ventilators or organs, to patients who are more likely than others to survive if they receive treatment.[388] There are many reasons why previous mortality rates may have been higher among ethnic minority patients (e.g., they may have been treated at lower quality institutions or they may previously have been discriminated against). These inequalities may become encoded in risk prediction models used to allocate scarce resources. These AI-CDS systems would reproduce disadvantages (colliding with the first dimension of Fredman's substantive equality framework) and might contribute to further marginalisation of disadvantaged groups (the third dimension). The factors causing these interferences with substantive equality may be unknown; they may be linked together in an inextricable chain of causal factors. They might be called 'structural' inequalities, because they are the result of underlying structures in society. There is not one entity that is responsible for structural inequalities; it is the "world as it is" that causes the model to produce undesirable outcomes in these cases.[389] However, in the context of healthcare there are specific reasons for concern about structural inequalities being reflected by AI-CDS systems. Access to healthcare is not equally distributed across groups that differ in terms of sex or ethnicity, for example.

Disparities between ethnic groups are perhaps studied most extensively in the US, due to the country's history with segregation, racism and oppression of African Americans.[390] Although there are particularities to the US health system that play a role in the explanation of how inequality across ethnic groups occurs, the existing research on health inequalities largely builds on generalisable insights that can also be applied more globally. Regardless of geographic and sociodemographic context, research on health disparities in the US is relevant as far it shows which types of inequalities there may be in the health sector and the

---

[388] Similar examples are given by Paulus and Kent (2020) 4-5.

[389] Suresh and Guttag (2019) 3-4.

[390] See, generally: Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care, *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care* (Washington DC: The National Academic Press, 2003), http://nap.edu/12875.

consequences of those inequalities. The following therefore makes several references to literature from the US.

Also in Europe there have been studies and reports suggesting that patient ethnicity may have impacted the access of health services and the quality of services.[391] A study from 2018 identifies ethnic minorities as a particularly vulnerable group in the context of access to health services.[392] It joins several reports both from mainland Europe and from the United Kingdom in indicating that discrimination and inequality is a problem in the health sector in many European countries.[393] These reports, which take into account empirical and self-reported data, suggest that ethnic minorities do not have access to health services of the same quality as the majority population and that services are to a lesser degree adapted to minority groups' needs.[394]

Ethnic disparities in healthcare arise from a myriad of complex factors.[395] Consider a study within a single health institution, which revealed that 'African American' women faced a higher likelihood of dying from breast cancer than 'White' women.[396] The authors of the study express uncertainty about the reasons for the disparity. In general, social and economic factors that correlate with ethnicity may often be involved, in addition to, sometimes, possible

---

[391] Rita Baeten et al., *Inequalities in Access to Healthcare - a Study of National Policies*, European Commission (Brussels, November 2018), 8, https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=8152&furtherPubs=yes; Crowley (2021): 14.

[392] Crowley (2021): 11.

[393] European Union Agency for Fundamental Rights (FRA), *Inequalities and Multiple Discrimination in Access to and Quality of Healthcare* (Luxembourg: Publications Office of the European Union, 2013); Jonas Frykman et al., *Discrimination - a Threat to Public Health Final Report - Health and Discrimination Project* (2007); Equalities Review Panel, *Fairness and Freedom: The Final Report of the Equalities Review* (Crown, 2007).

[394] Raj S Bhopal, "Racism in Health and Health Care in Europe: Reality or Mirage?," *European Journal of Public Health* 17, no. 3 (2007): 238, https://doi.org/10.1093/eurpub/ckm039.

[395] Alan Nelson, "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care," *Journal of the National Medical Association* 94, no. 8 (2002).

[396] Kevin Chu Foy et al., "Disparities in Breast Cancer Tumor Characteristics, Treatment, Time to Treatment, and Survival Probability among African American and White Women," *NPJ breast cancer* 4 (2018 2018), https://doi.org/10.1038/s41523-018-0059-5.

clinician prejudice or stereotyping. Communication issues and systematic differences in how and to what extent minority groups use health services may also contribute.[397] Groups that have experienced discrimination historically, might refrain from seeking medical treatment more often than other groups,[398] and from sharing relevant health information with clinicians and researchers due to mistrust.[399]

Disparities in access to or quality of healthcare services also occur between the sexes. Such disparities are equally undesirable regardless of whether it is men or women who experience the negative consequences. In practice, however, it is increasingly being recognised that women are often at the disadvantaged side of health-related disparities.[400] In 2021, the Norwegian government commissioned the Committee on Women's Health (*Kvinnehelseutvalget*) to investigate the challenges women encounter relating to health and healthcare accessibility.[401] In the report delivered in March 2023, the committee underscores that there are important differences between men and women that have not been sufficiently accounted for in the provision of healthcare services.[402] The committee stresses the importance of recognising biological sex as a determinant that invariably affects the trajectory and manifestation of diseases, responsiveness to medication, and overall utilisation of health services.[403] This is in accordance with findings in international health research.[404]

---

[397] The Norwegian Committee on Women's Health notes that national minorities have experienced systemic discrimination historically, which may affect how they use health services currently: Nou 2023: 5 Den Store Forskjellen. Om Kvinners Helse Og Betydningen Av Kjønn for Helse, 91 (NOU 2023: 5).

[398] European Union Agency for Fundamental Rights (FRA) (2013): 93; Annette J. Browne et al., "Can Ethnicity Data Collected at an Organizational Level Be Useful in Addressing Health and Healthcare Inequities?," *Ethnicity & Health* 19, no. 2 (2014): 249, https://doi.org/10.1080/13557858.2013.814766.

[399] Rajkomar et al. (2018) 868.

[400] James A. Marcum, "Clinical Decision-Making, Gender Bias, Virtue Epistemology, and Quality Healthcare," *Topoi* 36, no. 3 (2017), https://doi.org/10.1007/s11245-015-9343-2. p 502.

[401] NOU 2023: 5.

[402] NOU 2023: 5, 13.

[403] NOU 2023: 5, 49-50.

[404] e.g., Marianne J. Legato, Paula A. Johnson, and JoAnn E. Manson, "Consideration of Sex Differences in Medicine to Improve Health Care and Patient Outcomes," *JAMA* 316, no. 18 (2016),

Particularly, the Norwegian Committee on Women's Health notes that due to insufficient clinical understanding of differences between the sexes, which is partially explained by an historical lack of sex-specific research,[405] women are diagnosed at a later stage (compared to men) with conditions such as heart disease and bladder cancer.[406] This tendency suggests that clinicians might treat unlike cases alike when it comes to the assessment of these medical conditions, which is problematic from a substantive equality perspective. The Committee highlights that, due to insufficient knowledge, the sexes have been subject to formally equal treatment where sex-aware adjustments would have been preferrable. [407]

The exact reasons for sex-related differences in health outcomes are not always understood.[408] As illustration, certain studies have highlighted an intriguing dynamic: Female patients tend to have worse surgical outcomes when operated on by male surgeons.[409] The study by Wallis et al. found that sex discordance between surgeon and patient was generally associated with worse outcomes. However, the main driver behind this association was the outcomes for women who were treated by men. In the study, women treated by men had worse outcomes than men treated by women. If one factors in the prevalence of male surgeons, this

---

https://doi.org/10.1001/jama.2016.13995; Lorraine Greaves and Stacey A. Ritz, "Sex, Gender and Health: Mapping the Landscape of Research and Policy," *International Journal of Environmental Research and Public Health* 19, no. 5 (2022), https://doi.org/10.3390/ijerph19052563; Cara Tannenbaum, Colleen M. Norris, and M. Sean McMurtry, "Sex-Specific Considerations in Guidelines Generation and Application," *Canadian Journal of Cardiology* 35, no. 5 (2019), https://doi.org/10.1016/j.cjca.2018.11.011.

[405] NOU 2023: 5, 60.

[406] NOU 2023: 5, 49; Harun Fajkovic et al., "Impact of Gender on Bladder Cancer Incidence, Staging, and Prognosis," *World Journal of Urology* 29, no. 4 (2011), https://doi.org/10.1007/s00345-011-0709-9.

[407] NOU 2023: 5, 61.

[408] For instance, some studies suggest higher mortality rates in female patients in hospital after myocardinal infarction: Viola Vaccarino, Harlan M Krumholz, and Jorge Yarzebski, "Sex Differences in 2-Year Mortality after Hospital Discharge for Myocardial Infarction," *Annals of Internal Medicine* 134, no. 3 (2001), https://doi.org/10.7326/0003-4819-134-3-200102060-00007; Hani Jneid et al., "Sex Differences in Medical Care and Early Death after Acute Myocardial Infarction," *Circulation* 118, no. 25 (2008), https://doi.org/10.1161/CIRCULATIONAHA.108.789800.

[409] Christopher JD Wallis et al., "Association of Surgeon-Patient Sex Concordance with Postoperative Outcomes," *JAMA surgery* 157, no. 2 (2022), https://doi.org/10.1001/jamasurg.2021.6339.

observation implies a more pronounced surgical risk for women than men. Pinpointing the exact reasons behind this disparity remains challenging. Implications of the trend revealed by Wallis et al. for the development of an AI-CDS system for spine surgery are discussed in section 5.2.

4.4.2.3 Representation Bias, Type 1 (sample bias)

Representation bias is among the most commonly described sources of bias in AI systems.[410] The term 'representation bias' refers here to the quantitative representation of a group. For the purposes of this thesis, there are two relevant types of quantitative representation bias. The first type is when data is not representative of the relevant patient population, because it reflects a different demographic composition compared to the population where an AI-CDS system is intended to be used. This type of representation bias is often called 'sample bias.'[411] When sample bias is present in a dataset (a sample), the representation of a group within the dataset does not correspond to the actual presence of the same group in the relevant population. Instead, a group is either over- or underrepresented, compared to the relevant population.[412] For example, consider the situation where a model deployed at Hospital A has been trained on data from Hospital B, which serves a more homogeneous ethnic population compared to Hospital A. In this case, ethnic minorities will be even more marginalised within the training data than they are if one looks at the demographic of the population served by Hospital A. Situations like this are likely to occur in practice: in the US, a study found that the geographic distribution of patient cohorts included in training data used in development of clinical deep learning algorithms mostly came from only three states, which affected the demographic composition of patient groups represented in the training data.[413]

There are many reasons why underrepresentation can occur in a dataset. Some of the reasons are related to structural inequalities. For instance, some groups tend to leave a smaller 'digital

---

[410] Davide Cirillo et al., "Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare," *NPJ Digital Medicine* 3, no. 1 (2020): 2, https://doi.org/10.1038/s41746-020-0288-5.

[411] Vokinger, Feuerriegel, and Kesselheim (2021) 2.

[412] Barocas and Selbst (2016) 687.

[413] Amit Kaushal, Russ Altman, and Curt Langlotz, "Geographic Distribution of Us Cohorts Used to Train Deep Learning Algorithms," *JAMA* 324, no. 12 (2020), https://doi.org/10.1001/jama.2020.12067.

footprint' than other groups (the 'digital divide').[414] The digital divide can in some cases explain why some groups are underrepresented in training data. Elderly parts of the population may not be as active on the digital platforms from which data are collected as the younger parts of the population. In datasets collected from the health sector, underrepresentation of certain groups may occur as a result of those groups having had less access to healthcare, historically (for various reasons). Representation bias and structural inequalities may therefore be intertwined sources of bias.

Another source of sample bias in datasets is selection bias.[415] Selection bias occurs "when a rule other than simple random sampling is used to sample the underlying population that is the object of interest."[416] The issue is now with the process of data collection. A well-known example of selection bias is when the collection of data is based on consent: There may be systematic differences between the groups that consent to data collection and those that refuse. Moreover, selection bias may occur because of various decisions or barriers pertaining to the data collection process. For instance, a report from Data & Society on fairness in precision medicine describes how a set of guidelines for lung cancer screening required that candidates had to smoke 30 packs per year in order to be eligible for screening, not recognising that it is more common in the African American part of the population to smoke menthols, and people who smoke menthols smoke a smaller number of packs per year on average.[417]

---

[414] Nizan Geslevich Packin and Yafit Lev-Aretz. "Learning Algorithms and Discrimination." Chap. 4 In *Research Handbook on the Law of Artificial Intelligence*, eds. Woodrow Barfield and Ugo Pagallo, Edward Elgar Publishing Limited, 2018: 97. ("Classes of people with minimal or no digital footprint may discover that they are not included in opportunities that rely on predictive data-driven assessments.")

[415] Selection bias is a general problem in statistics. It is explained as such, for example, by Heckman: James J. Heckman, "Selection Bias and Self-Selection," in *Econometrics*, ed. John Eatwell, Murray Milgate, and Peter Newman (London: Palgrave Macmillan UK, 1990); Castelnovo et al. (2022).

[416] Heckman (1990) 201.

[417] Kadija Ferryman and Mikaela Pitcan, *Fairness in Precision Medicine*, Data & Society (February 2018), 24, https://datasociety.net/wp-content/uploads/2018/02/DataSociety_Fairness_In_Precision_Medicine_Feb2018.pdf.

4.4.2.4 Representation Bias, Type 2 (unequal representation)

The second type of representation bias is when data is representative of the population, but different groups are unequally represented in absolute terms, such as when there are minority groups in the population. This type of representation bias is referred to in the following as 'unequal representation'.

Minorities in the real world will usually be minorities in datasets, which means that, quantitatively speaking, there will be less data representing them, compared to the majority population. Health data registered in care facilities or in secondary health registries can generally be expected to contain less data on ethnic minority patients. The problem is that algorithms will be trained using quantitively fewer examples of the minority population, which may lead to less accurate predictions for the minority group (disparate accuracy).[418] In its 2021 Guidance on artificial intelligence in the health sector, the World Health Organization emphasises issues of unequal representation, noting that biases due to different group sizes will "emerge during modelling and subsequently diffuse through the resulting algorithm."[419] When there is unequal representation in training data, an AI-CDS system might

---

[418] Suresh and Guttag (2019) 4; Castelnovo et al. (2022) 3. In relation to prediction of COVID-19, see: Laure Wynants et al., "Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal," *BMJ* 369 (2020), https://doi.org/10.1136/bmj.m1328; Eliane Röösli, Brian Rice, and Tina Hernandez-Boussard, "Bias at Warp Speed: How AI May Contribute to the Disparities Gap in the Time of Covid-19," *Journal of the American Medical Informatics Association* 28, no. 1 (2020), https://doi.org/10.1093/jamia/ocaa210. In relation to AI that transcribes patient speech into clinical text, see: Allison Koenecke et al., "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences* 117, no. 14 (2020), https://doi.org/10.1073/pnas.1915768117. In relation to AI that diagnoses melanoma from skin lesions, see: Adewole S Adamson and Avery Smith, "Machine Learning and Health Care Disparities in Dermatology," *JAMA dermatology* 154, no. 11 (2018), https://doi.org/10.1001/jamadermatol.2018.2348. Note also what Xiaoxuan Liu at Birming University tells Wired magazine: "… algorithms might fail due to differences [between racial/ethnic groups] that are too subtle for doctors themselves to notice," and compared the situation to that of "having a Google Maps that doesn't go into certain post codes": Will Knight, "AI Can Help Diagnose Some Illnesses - If Your Country Is Rich," *Wired*, 18 November, 2020, https://www.wired.com/story/ai-diagnose-illnesses-country-rich/?

[419] To be precise, the WHO refers to this as "underrepresentation." However, underrepresentation is often used to denote cases where a group is underrepresented compared to its actual presence. World Health Organization (2021) 36 and 55.

learn to detect more accurately the symptoms that are typical for the majority group than the symptoms more often seen in a minority group.

4.4.2.5 Aggregation Bias

There may be a systematic difference in *how* the same factual situation is represented by data pertaining to different groups. In other words, the relationship between the input and the variable that a model is intended to predict may systematically differ between groups. This is often called 'aggregation bias,' because the bias occurs when data from different groups are aggregated and treated as if the target variable always has the same implications. For example, in Norway, immigrants from South-Eastern Asia have more type 2 diabetes than other segments of the population.[420] These persons also generally experience more complications at lower levels of obesity compared to other persons. They may receive less treatment than their obesity-related health condition warrants, because treatment guidelines are better adapted to the majority population.[421] A given score on the obesity scale actually means different things in terms of health implications for immigrants from South-Eastern Asia compared to other persons (according to the data observed in Norway).

Similarly, Cirillo et al. highlight the importance of accounting for the fact that symptoms of Parkinson's Disease may present differently in female patients compared to male patients.[422] The problem of aggregation bias has been demonstrated in research pertaining to decision support tools to diagnose and monitor diabetes.[423] Because there are known differences in diabetes complications between ethnic groups, research indicates that relevant "factors have different meanings and importances within different subpopulations" and that a single model is therefore "unlikely to be best-suited for any group in the population even if they are equally represented in the training data."[424]

---

[420] Kjersti Flugstad Eriksen, "Skal Forske På Hvorfor Innvandrere Fra Sør-Asia Oftere Får Diabetes," *Dagens medisin* 2023, https://www.dagensmedisin.no/diabetes-oslo-universitetssykehus-ous-overvekt/skal-forske-pa-hvorfor-innvandrere-fra-sor-asia-oftere-far-diabetes/565230.

[421] Ibid.

[422] Cirillo et al. (2020) 3.

[423] Suresh and Guttag (2019) 5.

[424] Suresh and Guttag (2019) 5.

Furthermore, it is possible that cultural and language-related differences may cause ethnic groups to present the same symptoms differently even if they are not actually experienced very differently.[425] Furthermore, there may be subtle patterns of variation in how clinicians communicate with different ethnic groups, which in turn may cause different information to be registered from two patients who are in similar situations.[426]

When aggregation bias occurs, this is likely to hamper an AI-CDS system's accuracy for all cases. If there is also representation bias in the dataset, the model might become better adapted to the majority group.[427] Aggregation bias is therefore a source of bias in AI-CDS systems that is closely connected to representation bias.

4.4.2.6 Historical Error Disparities, Including Stereotyping and Prejudice

Historical error disparities in decision-making can lead to disparate accuracy in AI-CDS systems, when the historical error disparities are reflected in training data.[428] The term 'historical error disparities' here refers to the fact that decision-makers may historically have made less accurate assessments for certain groups compared to others. Educational and health institutions sometimes do not provide the necessary training and other prerequisites for accurate collection and interpretation of data from minority patients by health personnel.[429] This can lead to a loss of accuracy in certain assessments, for example diagnostic assessments of skin lesions. Erroneous data on ethnic minority patients may also be explained by communication barriers between clinicians and patients, which are more dominant when the patient belongs to a minority group (as noted also in the previous section).

---

[425] Ana I. Balsa and Thomas G. McGuire, "Prejudice, Clinical Uncertainty and Stereotyping as Sources of Health Disparities," *Journal of Health Economics* 22, no. 1 (2003): 97, https://doi.org/10.1016/S0167-6296(02)00098-X., with further references.

[426] Ana I. Balsa, Thomas G. McGuire, and Lisa S. Meredith, "Testing for Statistical Discrimination in Health Care," *Health Services Research* 40, no. 1 (2005), https://doi.org/10.1111/j.1475-6773.2005.00351.x.

[427] Suresh and Guttag (2019) 5.

[428] The term 'error disparities' is used by Shah, Schwartz and Hovy to denote an unequal distribution of errors between groups: Shah, Schwartz, and Hovy (2019) 1.

[429] Bhopal (2007) 239. Notice however that Bhopal uses the phrase "institutional racism" about these situations.

One important source of historical error disparities in health data can be described in terms of measurement issues, i.e., differences in how, when and to what extent health data is recorded from different groups.[430] For example, certain pulse oximeters, smart watches or body sensors might register more incorrect data or fail to register data due to biological differences not accounted for in the development and testing of the sensor which takes care of the measurement.[431] Measurement issues may lead to error disparities as well as underrepresentation of disadvantaged groups in datasets used for training.[432]

Moreover, ethnic minority groups have historically been either over- or underdiagnosed with certain illnesses.[433] Disproportionate misdiagnosing of an ethnic minority group, historically, could lead to biased AI-CDS systems because diagnostic data are used as examples for ML algorithms to learn from.[434] Research on health disparities in the US has indicates that minorities receive lesser quality pain care compared with the non-Hispanic, White majority.[435] Related tendencies have been reported also in Norway: In a study of the provision

---

[430] Sendhil Mullainathan and Ziad Obermeyer, "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review* 107, no. 5 (2017): 479, https://doi.org/10.1257/aer.p20171084.

[431] Christine M. Cutillo et al., "Machine Intelligence in Healthcare—Perspectives on Trustworthiness, Explainability, Usability, and Transparency," *NPJ Digital Medicine* 3, no. 47 (2020): 3, https://doi.org/10.1038/s41746-020-0254-2; Michael W Sjoding et al., "Racial Bias in Pulse Oximetry Measurement," *New England Journal of Medicine* 383, no. 25 (2020), https://doi.org/10.1056/NEJMc2029240.

[432] James Zou and Londa Schiebinger, "Ensuring That Biomedical AI Benefits Diverse Populations," *EBioMedicine* 67, no. 103358 (2021): 2, https://doi.org/10.1016/j.ebiom.2021.103358.

[433] Steven J Trierweiler et al., "Clinician Attributions Associated with the Diagnosis of Schizophrenia in African American and Non-African American Patients," *Journal of consulting and clinical psychology* 68, no. 1 (2000), https://doi.org/10.1037/0022-006X.68.1.171; Michelle DeCoux Hampton, "The Role of Treatment Setting and High Acuity in the Overdiagnosis of Schizophrenia in African Americans," *Archives of psychiatric nursing* 21, no. 6 (2007), https://doi.org/10.1016/j.apnu.2007.04.006.

[434] Ziad Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (2019): 7, https://doi.org/10.1126/science.aax2342; Mullainathan and Obermeyer (2017); Rajkomar et al. (2018) 867.

[435] K. O. Anderson, C. R. Green, and R. Payne, "Racial and Ethnic Disparities in Pain: Causes and Consequences of Unequal Care," *Journal of Pain* 10, no. 12 (December 2009),

of epidural analgesia (pain relief) during labour, women from Sub-Saharan Africa were less likely than other groups to receive this type of pain relief medication.[436] If doctors systematically underestimate the pain reported by patients of a certain ethnicity, this group of patients will receive less treatment than other groups of patients who are actually experiencing the same level of pain.[437] Furthermore, minority groups may end up seeing the instances in the health system where relevant data is recorded at a lower rate, because they are less likely to be referred to further treatment.[438]

A similar scenario occurs in respect of sex discrimination, if men are more often (correctly) diagnosed with heart disease when presenting with the same symptoms as women. When there is a tendency for symptoms of a disease to go unnoticed in female patients, this may lead to delayed diagnosis and more false negatives for women.[439] For instance, the Norwegian Committee on Women's Health mentions that symptoms of stroke may differ between men and women and that this may increase the likelihood of missing an ongoing stroke in female patients.[440] Similarly, Vokinger and Gasser claim that an AI model that learns from EHR data may not recommend testing for cardiac ischemia for women due to the historical tendency of

---

https://doi.org/10.1016/j.jpain.2009.10.002; Adam D. DeVore and Adrian F. Hernandez, "49 - Quality and Outcomes in Heart Failure," in *Heart Failure: A Companion to Braunwald's Heart Disease (Fourth Edition)*, ed. G. Michael Felker and Douglas L. Mann (Philadelphia: Elsevier, 2020).

[436] Åsa Henning Waldum et al., "The Provision of Epidural Analgesia During Labor According to Maternal Birthplace: A Norwegian Register Study," *BMC Pregnancy and Childbirth* 20, no. 321 (2020), https://doi.org/10.1186/s12884-020-03021-8.

[437] Kelly M. Hoffman et al., "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences between Blacks and Whites," *Proceedings of the National Academy of Sciences* 113, no. 16 (2016), https://doi.org/10.1073/pnas.1516047113.

[438] Research from the US suggests that Black patients may be less likely to be referred for cardiac catheterization than White patients, after seeing a physician about chest pains: Kevin A. Schulman et al., "The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization," *New England Journal of Medicine* 340, no. 8 (1999), https://doi.org/10.1056/nejm199902253400806.

[439] Jenna Wiens, W Nicholson Price, and Michael W Sjoding, "Diagnosing Bias in Data-Driven Algorithms for Healthcare," *Nature Medicine* 26, no. 1 (2020): 25-26, https://doi.org/10.1038/s41591-019-0726-6.

[440] NOU 2023: 5, 51.

clinicians to misinterpret women's symptoms.[441] As a consequence, an AI-CDS system may cause women to receive potentially lifesaving treatment with more delay compared to men.[442] Whenever this type of systematic error disparity exists, an ML algorithm may infer that women are less likely to have a disease than men.

The reasons for historical error disparities are often challenging to disentangle.[443] Some error disparities may be a result of cognitive biases in the past. If past assessments are influenced by prejudiced or stereotyped attitudes, this will be reflected in training data. In healthcare, research indicates that ethnic stereotypes exist among clinicians in ways that could negatively affect the care given to ethnic minority patients.[444] To illustrate by reference to an anecdotal example, a Norwegian clinician gives her own account of outright prejudice from a colleague against a patient of Somali origin, in the Journal of the Norwegian Medical Assocation.[445] Allegedly, the colleague instructed that pain estimation for a Somali patient should be reduced to half of what the patient reported.[446] Prejudice and stereotyping in clinical reasoning may also occur in relation to the biological sex of the patient.[447] In addition to causing direct allocational harms, the reflection of prejudice and stereotyping in training data are likely to manifest in new instances of prejudice and stereotyping, contrary to the objective of substantive equality as described in section 4.2.3.

In addition to prejudice and stereotyping, another relevant cognitive bias that might cause equality-related biases in an AI-CDS system has to do with general tendencies in human reasoning that work to the detriment of minority groups. Psychologists Michael Billig and Henri Tajfel have established that there is a tendency to perceive persons who one define as

---

[441] Kerstin N. Vokinger and Urs Gasser, "Regulating AI in Medicine in the United States and Europe," *Nature Machine Intelligence* 3, no. 9 (2021): 738, https://doi.org/10.1038/s42256-021-00386-z.

[442] Vokinger and Gasser (2021) 738.

[443] Cirillo et al. (2020) 2.

[444] Vickie L. Shavers et al., "The State of Research on Racial/Ethnic Discrimination in the Receipt of Health Care," *American Journal of Public Health* 102, no. 5 (2012), https://doi.org/10.2105/ajph.2012.300773.

[445] Martine Rostadmo, "Svart Hud Er Tykkere Enn Hvit," *Tidsskrift for Den norske legeforening* (2021), https://doi.org/10.4045/tidsskr.21.0058.

[446] Ibid.

[447] Marcum (2017) 502.

outside of one's own group (outgroup) differently from those belonging to one's own group (ingroup).[448] This is sometimes called 'social identity theory.'[449] It is possible, in accordance with social identity theory, that clinicians in some instances treat ingroup patients more favourably compared to outgroup patients.

4.4.3 Bias in Data Processing and Modelling Choices

4.4.3.1 Introduction

Although the ML process is partially based on the autonomous learning of an algorithm, there is considerable room for human influence on the resulting model in important ways. Processing of data that has been collected for training purposes may necessitate various decisions from the AI development team, such as removing or merging features that exist in the data, filling in missing information, correcting errors, removing duplicated information or translating information into numerical values.[450]

For simplicity, all decisions made during the development process are referred to here as 'modelling choices.' Modelling choices not only involves the processing of training data, but also decisions such as what a model should be trained to predict (choice of target variable). The relationship between modelling choices and equality-related biases displayed by the model is complex and it may be difficult to foresee how modelling choices could lead to biases. The following outlines some potential ways in which modelling choices can contribute to equality-related biases. However, this is not intended as a comprehensive overview of modelling choices in machine learning projects.[451]

---

[448] Michael Billig and Henri Tajfel, "Social Categorization and Similarity in Intergroup Behaviour," *European Journal of Social Psychology* 3, no. 1 (1973), https://doi.org/10.1002/ejsp.2420030103; Henri Tajfel, *Differentiation between Social Groups : Studies in the Social Psychology of Intergroup Relations*, vol. 14, European Monographs in Social Psychology, (London: Academic Press, 1978).

[449] Saul Mcleod, "Social Identity Theory in Psychology (Tajfel & Turner, 1979)," Olivia Guy-Evans ed. *Simply Psychology*, 2 October, 2023, https://www.simplypsychology.org/social-identity-theory.html.

[450] Theobald (2020) 36.

[451] For instance, the choice of learning algorithm is not discussed here, although Kim points out that it reflects certain trade-offs: Kim (2022) 1552. Other decisions that are not discussed include,

4.4.3.2 Feature Selection

Choices made by developers may impact the training data and lead to data bias,[452] in which case one may categorise the bias as either data bias or modelling bias. For example, developers may in some cases choose which features the training data shall contain ('feature selection').[453] Here, I categorise feature selection bias as a type of bias that relates to modelling choices. This is done to emphasise the distinction between biases resulting from the most active choices that AI developers make and biases that are inherent in the data they collect. However, it is not asserted that this distinction is watertight. Influential decision are also made by AI developers at the data collection stage.

One might wonder why feature selection is sometimes done manually. Why not just feed all the available data into the learning algorithm? After all, the purpose of the algorithm is to find patterns in data that humans may not identify as easily. In practice, there are a few different reasons why AI developers select features. One example is cases where the available data contains explicit information on patient ethnicity and the developers decide that this information should not be relied on by an ML algorithm. Thus, they 'wash' the data to remove information on patient ethnicity. Another reason why certain features of the data may be removed is because developers assume that certain features are not relevant to the prediction task the algorithm is intended for. Irrelevant features can be removed from the training data to avoid confusing the algorithm and to limit the volume of data and, thus, the computing resources needed to process the data.[454] Reducing the features can also lead to a more interpretable model, which may be an objective for developers. For much the same reasons, developers may sometimes decide to *merge* features rather than removing features.[455] For example, they may decide that an algorithm should see 'heart valve surgery' as one

---

for example, the choice of cost function and optimizer, choice of data imputation and normalisation techniques, choice of hyperparameters for model tuning, and choice of model evaluation measures.

[452] Castelnovo et al. (2022) 3; Neeraj Tandon and Rajiv Tandon, "Will Machine Learning Enable Us to Finally Cut the Gordian Knot of Schizophrenia," *Schizophrenia Bulletin* 44, no. 5 (2018), https://doi.org/10.1093/schbul/sby101.

[453] Barocas and Selbst (2016) 688.

[454] Theobald (2020) 36-37.

[455] Theobald (2020) 38.

feature, instead of distinguishing between ‘valve repair surgery’ and ‘valve replacement surgery,’ which are two types of heart valve surgery.

Feature selection is particularly important when training data are collected from Electronic Health Records, because EHRs tend to include more information than what is deemed necessary for the development of a decision model for narrow tasks such as diagnosis-setting or treatment selection.[456] The removal of superfluous information before processing health data is also supported by the GDPR’s data minimisation principle, which requires that the processing of personal data is limited to what is necessary for the task at hand.[457]

Feature selection can lead to equality-related bias is if there are differences between groups, so that the relevant features for the decision task are not the same across groups.[458] If the selection of features to be included in training data prioritizes data that is more relevant to the majority population,[459] this could lead to disparate accuracy in the trained model, to the detriment of minority groups. For minority groups, the model may be considering an insufficiently rich set of data/factors.[460] This type of bias does not only affect minority groups. It can affect any group for which the selected features are less predictive compared to other groups. For example, consider the choice of whether or not to include smoking as a feature in a model intended to predict the risk of dying from COVID-19. In China, smoking is more common among men than women. At the same time, smoking is an important predictive

---

[456] Johnson et al. (2018) 2669.

[457] Mathias K Hauglid and Karl Øyvind Mikalsen, "Tilgang Til Helseopplysninger I Maskinlæringsprosjekter," *Lov og Rett*, no. 7 (2022), https://doi.org/10.18261/lor.61.7.3.

[458] Barocas and Selbst (2016) 688. (“Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve.”)

[459] Cognitive biases among developers is one reason why this happens: Schwartz et al. (2022) 25. (“… designers who must make decisions on what variables to include or exclude can impart their own cognitive biases into the model”.)

[460] On the potential for statistical discrimination as a consequence of incomplete data: Toon Calders and Indrė Žliobaitė, "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures," in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, ed. Bart Custers et al. (Berlin, Heidelberg: Springer Berlin Heidelberg, 2013), 52-53.

factor when predicting the risk of dying from COVID-19. If information about smoking habits is not included in the training data, an ML algorithm cannot discover the correlation between smoking and the higher mortality rate. Instead, the algorithm may assign more weight to the patient being male and thus overestimate the mortality risk also for men who do not smoke.[461] This specific type of bias has also been called 'omitted variable bias' in the ML literature.[462]

4.4.3.3 Ethnicity and Sex as Factors in Clinical Decision-Making

As mentioned in section 4.3.2, biased AI-CDS systems is an equality-problem if they involve an inappropriate reliance on personal characteristics. In this regard, the thesis concentrates on ethnicity and sex. A particular side of the 'feature selection' issue discussed in the previous section is whether or not the sex and ethnicity of patients should be included as features in the training data. If they are included, these factors will constitute feature variables relied on by the trained AI-CDS system and cause a systematic difference in how the system treats patients based on these characteristics. It is therefore worth considering the role of ethnicity and sex as factors in clinical decision-making, traditionally, as this might be a source of bias in AI-CDS systems.

The topic ethnicity or 'race' (which is often used in the US discourse) as a factor in clinical decision-making is controversial and subject to a longstanding debate.[463] There was a lively discourse around the topic in the early 2000s, catalysed by breakthroughs in gene sequencing technology and other approaches towards more 'personalised' medicine. On one end of the spectrum, advocates of 'race-based' medicine promote its potential benefits, while the opposition caution against the risk of reinforcing historical inequalities by including patient

---

[461] Alex Engler, "A Guide to Healthy Skepticism of Artificial Intelligence and Coronavirus," 2023 (2 April 2020). https://www.brookings.edu/articles/a-guide-to-healthy-skepticism-of-artificial-intelligence-and-coronavirus/.

[462] Mehrabi et al. (2021) 5. ("Omitted variable bias occurs when one or more important variables are left out of the model")

[463] Shedra Amy Snipes et al., "Is Race Medically Relevant? A Qualitative Study of Physicians' Attitudes About the Role of Race in Treatment Decision-Making," *BMC Health Services Research* 11, no. 183 (2011), https://doi.org/10.1186/1472-6963-11-183.

ethnicity as a factor in clinical decision-making.[464] There is no universal stance in the medical domain on whether, how or in which instances it is clinically meaningful and acceptable to use ethnicity as a part of the equation in clinical decision-making.[465] It is widely acknowledged that ethnicity is not a *causal* factor for medical conditions.[466] However, this does not mean that knowledge of patient ethnicity is always irrelevant in the sense that it cannot contribute to accurately diagnosing or making predictions about a patient's health.[467] Because socioeconomic, demographic, cultural, dietary and other factors sometimes correlate with ethnicity, ethnicity can be a factor of some predictive value when more precise information is not available.[468]

Historically, however, the use of race/ethnicity in medical decision making has not always been confined to circumstances where it contributes to predictive accuracy in the best interest of the patient. Kaufman and Cooper suggest that, to the extent that medically useful information can be conveyed by a patient's ethnic identity at all, that information is "much less than or much different than is commonly held, which puts many minority patients at a considerable disadvantage."[469] If the use of patients' racial or ethnic identity (whether self-reported or as assumed by health practitioners) has worked to the detriment of minority groups in the past, this means that the data used to train ML algorithms could be infected with

---

[464] The debate is described in retrospect by Vyas, Eisenstein, and Jones: Darshali A. Vyas, Leo G. Eisenstein, and David S. Jones, "Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms," *New England Journal of Medicine* 383, no. 9 (2020), https://doi.org/10.1056/NEJMms2004740.

[465] Jessica K Paulus and David M Kent, "Race and Ethnicity: A Part of the Equation for Personalized Clinical Decision Making?," *Circulation: Cardiovascular Quality and Outcomes* 10, no. 7 (July 2017), https://doi.org/10.1161/CIRCOUTCOMES.117.003823.

[466] Davenport and Kalakota therefore call an algorithm biased if it relies on ethnicity to make predictions about patients: Davenport and Kalakota (2019) 97.

[467] e.g.,Vyas, Eisenstein, and Jones (2020) 879.

[468] Kaufman and Cooper hypothesise that "justifiable uses of race in medical decision making will be rare," while also submitting that racial or ethnic identity can be highly informative: Jay S Kaufman and Richard S Cooper, "Use of Racial and Ethnic Identity in Medical Evaluations and Treatments," in *What's the Use of Race*, ed. Ian Whitmarsh and David S. Jones (Cambridge, Massachusetts: MIT Press, 2010), 200. On the use of race in algorithms in medicine, see also: Vyas, Eisenstein, and Jones (2020) 876-78.

[469] Kaufman and Cooper (2010) 200.

medically unjustified assumptions which may be reinforced if patient ethnicity is an available feature during training of an ML algorithm to be implemented in an AI-CDS system.[470]

When it comes to the biological sex of a patient, this is often an indisputably relevant factor in clinical decision-making.[471] The *reluctancy* to provide sex-specific assessment and care has arguably been more of an equality challenge than the actual use of sex as a factor in clinical decision-making.[472]As mentioned in section 4.4.2.2, the Norwegian Committee on Women's Health emphasises the importance of awareness around the inherent biological differences between men and women. While it is obvious that certain hormonal differences or differences in reproductive anatomy might directly influence the occurrence or progression of specific diseases or conditions, the committee finds that many differences are not properly considered in clinical decision-making.[473] By not considering the different needs and risks associated with each sex, medical practitioners might miss out on essential information that could guide patient care, potentially leading to suboptimal treatment. The implication is that the decision on whether to include patient sex as a feature that an AI model may rely on, has important consequences. Equality-related biases can arise if these features are included when they should not be, as well as if they are excluded when they ought to be considered. Just as with ethnicity, there is a potential pitfall in relying too heavily on biological sex in clinical decision-making, especially if there's no clear understanding of its importance in a particular

---

[470] Davenport and Kalakota (2019) 97.

[471] NOU 2023: 5, 50; Naomi Breslau et al., "Sex Differences in Posttraumatic Stress Disorder," *Archives of General Psychiatry* 54, no. 11 (1997), https://doi.org/10.1001/archpsyc.1997.01830230082012; Myrna M. Weissman and Gerald L. Klerman, "Sex Differences and the Epidemiology of Depression," *Archives of General Psychiatry* 34, no. 1 (1977), https://doi.org/10.1001/archpsyc.1977.01770130100011; Maria Teresa Ferretti et al., "Sex Differences in Alzheimer Disease — the Gateway to Precision Medicine," *Nature Reviews Neurology* 14, no. 8 (2018), https://doi.org/10.1038/s41582-018-0032-9; Berna C Özdemir and Anna Dorothea Wagner, "Consideration of Sex and Gender Aspects in Oncology: Rationale, Current Status, and Perspectives," *Italian Journal of Gender-Specific Medicine* 8, no. 1 (2022), https://doi.org/10.1723/3769.37567.

[472] Deborah Bartz et al., "Clinical Advances in Sex- and Gender-Informed Medicine to Improve the Health of All: A Review," *JAMA internal medicine* 180, no. 4 (2020), https://doi.org/10.1001/jamainternmed.2019.7194.

[473] Section 4.4.2.2.

context. Over-reliance or misuse of sex as a predictive factors without a solid medical foundation can lead to equality-related biases in AI-CDS systems.

4.4.3.4 Choice of Target Variable

The choice of target variable is a matter of defining the problem to be solved in a way that can be effectuated by an AI model. This decision is sometimes referred to as 'outcome choice.'[474] As explained in section 1.5.7, the target variable is the variable that the model is intended to predict – the output. Thus, defining the target variable is a question of what exactly an AI developer should instruct the algorithm to learn.[475] According to Obermeyer et al., it is a common problem in machine learning that biases occur because "algorithms are aimed at the wrong target to begin with."[476] The choice of target variable is arguably the most important modelling decision development teams make, because it determines what the model will predict.

The choice of target variable involves translating a clinical problem into a computation problem. The problem that one is trying to solve may be defined in a way that disadvantages certain groups.[477] Particularly, certain groups may be disadvantaged by an AI model if the target variable reflects structural inequalities.[478] This occurs, for example, if an model intended to predict where crime will occur actually predicts where an arrest is likely to occur. Arrests amount to a target variable that works to the detriment of groups living in areas that have been policed more intensely in the past. Due to the historical police behaviour, arrests more than actual crime, may be correlated with ethnicity. The use of target variables that are correlated with protected characteristics is often called 'proxy discrimination,' because the target variable serves as a proxy for the protected characteristic.

A real-world example concerning a model used to allocate healthcare services in the US illustrates how an inappropriate target variable can lead to equality-related bias in an AI-CDS

---

[474] Zou and Schiebinger (2021) 2.

[475] Ziad Obermeyer et al., "Algorithmic Bias Playbook," *Center for Applied AI at Chicago Booth* (2021): 2.

[476] Ibid.

[477] Barocas and Selbst (2016) 675 and 78; Calders and Žliobaitė (2013) 50-51.

[478] Schwartz et al. therefore argue that developers "should be able to demonstrate that the application is measuring the concept it intends to measure": Schwartz et al. (2022) 15.

system.[479] Obermeyer et al. show how the model predicted future healthcare costs incurred by patients as a proxy for their actual health needs.[480] In this case, the clinical problem was to predict health needs, and the computational solution chosen by the system developers was to predict expenditure. However, Black patients, as a group, generated less expenditure than White patients according to the training data.[481] Given this structural inequality, the target variable of expenditure meant different things for different patient groups.[482] In effect, Black patients needed to be more sick than White patients for the model to allocate health services to them. The study by Obermeyer et al. serve as the core inspiration for the fictional case study that this thesis offers in section 5.1.

4.4.3.5 Labelling (for Supervised Learning)

In supervised learning, pre-labelled data is typically used for training purposes. In relation to a classification problem, the label would be the category or class that an example in the dataset belongs to. If a trained model is intended to support the diagnosis of patients suspected of having schizophrenia, it would classify the patient as positive or negative in respect of this specific condition. If the model is developed using supervised learning, the training consists of repeated exposure to examples of patients whose data are labelled as either positive or negative, depending on whether they were diagnosed with schizophrenia or not. When the AI-CDS system has been deployed and is used to support the diagnosis of patients in the clinic, each patient is presented to the model as an unlabelled case. The model's task is to put the correct label on that case. Therefore, the process of labelling training data is closely connected with the choice of target variable.

Sometimes, the labels used for supervised learning are readily available in the collected data. However, other times, the data that is used to train models for AI-CDS systems need labelling because the target variable is not observable in the training data. In the terminology of Mikalsen et al., it is not an 'observed variable,' but a 'latent variable.'[483] In a text corpus, latent variables are words or phrases that are not explicitly mentioned, but which are

---

[479] Zou and Schiebinger (2021) 2.

[480] Obermeyer et al. (2019).

[481] Ibid; Wiens, Price, and Sjoding (2020) 25-26.

[482] Wiens, Price, and Sjoding (2020) 25.

[483] Mikalsen et al. (2017) 106.

nonetheless relevant to use as feature or target variables. Labelling is often necessary when the training data contains clinical notes from Electronic Health Records. For example, when training a model to predict postoperative delirium based on EHR data, Mikalsen et al. note that a "latent variable cannot be extracted directly from the EHR and could be the answer to a higher-level question such as *does the patient have postoperative delirium*?" By deciding that it was important to predict postoperative delirium, Mikalsen et al. chose a latent variable as the target variable. It was therefore necessary to label the EHR data so that it became observable to the algorithm whether each patient represented an example of postoperative delirium or not. Again, labelling and the choice of target variable are interconnected.

In the literature on equality-related biases in AI systems, it is often recognised that decisions on how to label the training data may contribute to an algorithm that is biased towards minority groups.[484] The labelling process may introduce cognitive biases of the development team.[485] In practice, the manual labelling of health data often requires medical expertise, which means that clinicians may be involved in the labelling. Thus, the labelling process is one point where error disparities or clinical stereotyping may contribute to a biased model. For example, heart conditions are more often missed in female patients compared to male patients; and women and ethnic minorities may be overdiagnosed with certain mental disorders. If clinicians repeat such tendencies when they are involved in the labelling of training data, the model will also reflect the same historical error disparities

Manual labelling is not always necessary and automated labelling procedures have become more widespread in recent years. When automated labelling is relied on, it is possible that bias is introduced to an AI model because of biases in the automated labelling procedure. Medical images usually need labelling because the images do not contain information on the diagnosis given to the patient. Labelling indicates to the learning algorithm whether each image represents an example of a positive or negative case. Medical images are sometimes

[484] Cutillo et al. (2020) 3.

[485] Shah, Schwartz, and Hovy (2019) 4.

labelled automatically using Natural Language Processing models, a technology that is prone to equality-related biases.[486]

4.4.4 Hidden Inferences

Reliance on inappropriate characteristics, which was highlighted in section 4.3.2 as an equality-related manifestation of bias, does not only occur because of feature selection choices made by AI developers. This type of bias can also occur where the presence of a characteristic, for example ethnicity, in training data is not observable to human developers, if the learning algorithm is able to infer ethnicity from the training data.[487] When this happens, it is possible that a model may end up using ethnicity as a feature variable without the knowledge of the AI-CDS system's developers. This is a potential problem that pertains particularly to non-interpretable neural networks based on deep learning techniques. For instance, consider a neural network that infers patient ethnicity from multi-modal data and covertly uses it as a feature variable. Furthermore, envisage that the algorithm associates patient ethnicity with a mental illness.[488] By relying on patient ethnicity and on this basis suggesting a diagnosis more often for patients from a certain group, an AI-CDS system could contribute to stigma about this group. Moreover, if the association that the AI-CDS system relies on is not medically sound, disparate accuracy could arise to the detriment of an ethnic group.

The possibility of hidden inferences is recognised in research concerning the use of neural networks for the interpretation of medical images in radiology. These models might be able to predict a patient's (self-reported) ethnicity with high accuracy, despite radiologists not being

---

[486] Laleh Seyyed-Kalantari et al., "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in under-Served Patient Populations," *Nature Medicine* 27, no. 12 (2021), https://doi.org/10.1038/s41591-021-01595-0; Hovy and Prabhumoye (2021).

[487] On the possibility that learning algorithms might infer various information, including protected characteristics, see, generally: Inga Strümke, Marija Slavkovik, and Clemens Stachl, "Against Algorithmic Exploitation of Human Vulnerabilities," *arXiv preprint arXiv:2301.04993* (2023). The issue is further discussed section 8.6.5.

[488] Historically, Black patients in the US have been diagnosed with schizophrenia more often than other ethnicities. Research suggests that differences in how clinicians respond to patients presenting similar symptoms is part of the explanation: Trierweiler et al. (2000); Hampton (2007).

able to identify ethnicity within these images.[489] Based on these findings, it has been suggested in the literature on speech-based AI-CDS systems for psychiatric diagnoses that such systems might also infer and rely on characteristics that are not observable to clinicians.[490] Understanding and mapping how algorithms combine input data to rely on features which, independently or combined, amount to inappropriate characteristics, is a serious concern in relation to the deployment of AI systems also more generally in society.[491]

4.4.5 Deployment Bias

Finally, biases can occur in AI systems at the stage of deployment. The performance of AI models often drops when they are deployed in a real-world setting, even if the model performs exquisitely during pre-deployment testing. One reason may be that the patient population or other circumstances in the real-world clinical setting differs from the patient population and circumstances reflected in the training data.[492] Variations of this problem are known by different names in the ML literature, including 'out of distribution,' 'out of population,' 'domain shift,' and 'context shift.'[493] Research suggests that AI-CDS systems may drop in accuracy by more than 10 % when tested at a different hospital than the testing site, partially due to different composition of ethnicity and sex in the patient groups combined with the system's different performance levels across demographic groups.[494] This problem is recognised in the WHO's 2021guidance for AI in the health sector.[495]

As an illustration of how the change of context from testing to deployment may lead to equality-related biases, consider a diagnostic model that is trained on data collected from a large, multinational database. This model's purpose is to assist in diagnosing a specific type of

---

[489] Imon Banerjee et al., "Reading Race: AI Recognises Patient's Racial Identity in Medical Images," *arXiv preprint arXiv:2107.10356* (2021).

[490] Hauglid (2022).

[491] See, generally: Frank Pasquale, *Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, Massachussetts: Harvard Universit Press, 2016).

[492] Vokinger, Feuerriegel, and Kesselheim (2021) 3.

[493] Schwartz et al. (2022) 26.

[494] Zou and Schiebinger (2021) 2; Ibid.

[495] World Health Organization (2021) 55. ("…an AI technology that is trained in one country and then used in a country with different characteristics may discriminate against, be ineffective or provide an incorrect diagnosis or prediction for a population of a different race, ethnicity or body type.")

cancer. While the database includes countries from around the world, it has less data from underdeveloped countries, including Eritrea. For the sake of illustration, let it be assumed that, in Eritrea, people are generally diagnosed with this type of cancer at an older age and at a more advanced stage of the disease, compared to patients in Norway.[496] The delayed diagnosis trend in Eritrea can be attributed to various factors distinct from those in Norway, such as lower levels of education, an underdeveloped healthcare system, and other barriers.

Now, envision deploying the model that was trained on multinational data in Norway. It will be utilized to support clinical decisions for all patients in Norway, including individuals like Simon Tesfay (whom the reader will become acquainted with in chapter 5 below) and other patients of Eritrean origin. If the model is applied to Simon Tesfay at a young age and an early stage of the disease, it may not perform as well as it would for ethnic Norwegian patients of similar age and disease stage. At this specific age and stage, the model might fail to recognise the disease in patients of Eritrean origin, leading to an increased number of false negatives for patients of Eritrean origin in comparison to those of Norwegian descent.

Having established that there are multiple, complex sources and causal mechanisms leading to equality-related biases in AI-CDS systems, the following case studies illustrate the implications of such biases from the perspectives of a potential victim of discrimination (section 5.1) and an AI development team (section 5.2).

[496] The example is inspired by experiences with breast cancer detection in Sub-Saharan Africa: Linda Nordling, "A Fairer Way Forward for AI in Health Care," 573 (26 September 2019), https://doi.org/10.1038/d41586-019-02872-2; Eleanor Black and Robyn Richmond, "Improving Early Detection of Breast Cancer in Sub-Saharan Africa: Why Mammography May Not Be the Way Forward," *Globalization and Health* 15, no. 1 (2019), https://doi.org/10.1186/s12992-018-0446-6.

5 Case Studies

5.1 Fictional Case Study: The Case of Simon Tesfay

5.1.1 Introduction

The following narrative is fictional. It is intended to illustrate how equality-related bias in an AI-CDS system might arise and impact the life of an individual. It also illuminates what it might take to become suspicious of potential discrimination and how challenging it may be to establish whether discrimination has indeed occurred.

This fictional case study is inspired by the study by Obermeyer et al. of ethnic bias in an algorithm used to allocate additional health services to patients in the US based on a prediction of future health expenditure as a proxy for health needs, which was introduced in section 4.4.3.4.[497] The study found that Black patients were less likely to receive additional health services compared to White patients with similar needs. The AI-CDS system described in the following case study serves a similar purpose to the algorithm studied by Obermeyer et al. However, the circumstances of the case are fictitious and somewhat adapted to a Norwegian context.

5.1.2 Simon Tesfay's Experience

Simon Tesfay (born 1971) has lived most of his life in Storevik, the administrative centre of the Storevik region. He came to Storevik as a child, when his family fled from the First Eritrean Civil War. For the last 10 years his general health has been deteriorating. He's been diagnosed with several medical conditions, for which he has frequently received treatment at the University Hospital of Storevik (UHS). During his many visits to the hospital, he has made some friends who are also ill. Simon bonds particularly well with Lars Holm, whom he has a lot of medical experiences in common with. They are the same age, and they have most of their diagnoses in common.

One day, Simon texted Lars about going out for coffee, to which Lars responded that he had to be at home, because he was waiting for a nurse to visit him. They postponed to the next day. When they met at the cafeteria in Storevik, Simon asked Lars if he was ill again. "No,

[497] Insert reference and see section 4.4.3.

no," said Lars. "The nurse's check-up was preventive. Apparently, I'm in such a bad shape that they are willing to do anything to keep me from occupying space at the hospital. So, they enrolled me in this preventive program thing."

A few weeks later, Simon felt ill and contacted his doctor. After a quick check-up, the doctor referred him to the department for internal medicine at UHS, where he was treated for kidney disease. After a week in the hospital, he was discharged.

Just as Simon exited the glass doors by the main entrance of UHS, a woman reached out to him. The woman was Marte Kirkerud, a journalist who said she was working a case concerning the hospital's use of AI systems. Simon told her that he had only been treated by humans, and never by robots. "The AI systems are really just complicated computer programs," Marte said. She explained to him that the hospital could be using AI systems as part of his treatment without his knowledge. She told him that the hospital relied on AI systems for many purposes, including for the purposes of predicting which patients would have the highest health needs during the next 12 months. Marte further told Simon that she found this type of prediction troublesome, because she had read that a similar system in the US was biased against black patients. She worried that an AI-CDS in use at UHS might display similar biases against ethnic minorities in Norway.

This led Simon to tell Marte the story about Lars Holm. "Now that you mention it," he said, "I can't really think of a good reason why Lars should receive preventive care, and not me. To me, it seems like our situations are pretty similar. You don't think UHS could be using an AI system to select patients for allocation of preventive care services, do you?" The slow nod from Marte let him know that, yes, that was exactly what she thought.

What Marte told Simon made him sad and angry, because he felt that he was being treated unfairly. He agreed to cooperate with Marte, who suggested that they should also involve the Norwegian Equality and Non-Discrimination Ombud. Together with the Ombud, they wrote a request for information from the hospital, requiring a detailed explanation of how the hospital selected patients for the Preventive Care Program as well as the reasons why Simon had not been offered to participate in the Preventive Care Program. The request specifically referred to the fact that Lars Holm received preventive services despite being in a comparable situation to Simon.

Furthermore, the hospital was asked to describe which data the AI-CDS system relied on to predict a patient's future needs. The hospital was also asked to provide statistics on the ethnic backgrounds of the patients who had been offered the Preventive Care Program, as well as statistics showing the overall ethnic composition of patients treated at the hospital since the program commenced. The Ombud suggested that this information might enable Simon to prove that the AI system was discriminatory.

5.1.3 University Hospital of Storevik (UHS)

UHS is a public hospital serving the population in the Storevik region in Norway. A clear majority of the population in Storevik consists of ethnic Norwegians (85 %). The largest ethnic minorities in Storevik are of Middle Eastern origin (7 %) and African origin (5 %), according to numbers provided by Statistics Norway.[498]

In its response to Simon Tesfay, UHS explained that deployment of AI systems for certain purposes began at the hospital in 2022. One of the first systems to be deployed was indeed used in the Preventive Care Program. In fact, the Preventive Care Program was an idea that the hospital management had been wanting to carry out for a long time, but they initially found that there was no way to efficiently select suitable patients for the program. Besides, the program would only work if the neediest patients could be identified through a targeted approach. However, when they read that a preventive care program had been established in the US based on an algorithm that automatically identified patients with special needs, UHS decided to start a procurement process. The hospital therefore announced a tender asking for an AI system designed for the purposes of identifying the patients who would have the largest health needs over a 12-month period.

A contract was eventually concluded with the Norwegian e-health company Norse Health ASA. They delivered an AI-CDS system which was safe and reliable because it was CE-marked in accordance with the Medical Device Regulation (initially, Norse Health ASA

[498] Storevik, in this case study, is not an actual place. However, Statistics Norway report on the representation of ethnic origins in different geographical areas within Norway. In the numbers from Statistics Norway, people are identified as originating from a foreign country if they or their parents are immigrants.

asserted that the AI-CDS system was not a medical device, but they eventually gave in to the Norwegian medical device authorities which argued that it was a medical device).

The AI-CDS system from Norse Health ASA continuously considers the future health needs of patients in the electronic health records held by UHS. If patients do not want the AI-CDS system to analyse their health records, UHS explained that they can notify the hospital thereof. When the system identifies a patient as qualified for the Preventive Care Program because the patient's predicted future use of health services exceeds a certain threshold, the system automatically notifies a clinician who makes the final decision on whether to offer the patient enrolment in the Preventive Care Program.

As regards the treatment offered to Lars Holm, the hospital wrote that they could not comment on the treatment of another patient, due to the duty of confidentiality. Moreover, they were not able to provide statistics showing the ethnicity of patients who had been offered the Preventive Care Program, or of all the patients treated at the hospital, because they did not keep track of their patients' ethnicity.

UHS further wrote in their statement that the hospital was extremely satisfied with the AI-CDS system and believed it had helped many patients to live better lives. They proudly proclaimed that the system's ability to identify the patients with the largest health needs were illustrated by the fact that the physicians agreed with the system's recommendation almost without exception.

5.1.4 The Workings of the AI-CDS System

In its response, the hospital enclosed technical documentation pertaining to the AI-CDS system. This documentation stated that the AI-CDS system was designed to predict which patients would require the most resources over the next 12 months, if they did not receive preventive treatment. More precisely, the algorithm's target variable was defined by a formula consisting of a patient's predicted number of visits to an outpatient clinic and the predicted number of days occupying a hospital bed. To predict these outcomes, the algorithm used all the information that was available about patients in their electronic health record and in the hospital's patient administrative system. This information includes the entire medical history with diagnoses, treatments, clinical notes, known diseases in a patient's near family, radiological images and physiological and biomedical measurements, etc. It also includes basic information such as a patient's age, occupation, address and place of birth.

Simon discussed the hospital's response with the Equality and Anti-Discrimination Ombud. They agreed that the response was not satisfactory and that it did not remove any suspicion about the AI-CDS system being discriminatory towards ethnic minorities. At the same time, it would be difficult to prove to a judge that Simon had been discriminated against. They needed more evidence. They contacted the hospital again and got it to agree to an external audit of the AI-CDS system.

5.1.5 Audit of the AI-CDS System in View of Discrimination

The hospital said it would not mind an external audit of the algorithm. It was agreed that researchers from the machine learning group at UiT the Arctic University of Norway should conduct the audit because of their outstanding expertise. The audit was funded partially through a crowdfunding project that had been established by Marte Kirkerud, and partially by UHS. The audit process took almost a year to complete, because it was necessary to collect data from patients who were not already registered at UHS, so that new data could be run through the AI-CDS system, in a simulation. The following is an excerpt from the report provided by the UiT group:

First of all, we find that the model is less accurate at predicting the health of ethnic minority patients, compared to the majority population. It seems that it tends to underestimate ethnic minority patients' future health needs. In our experiment with the algorithm, we gathered the data of 100 000 patients, making sure that half of them represented ethnic minorities, while half of them represented ethnic Norwegians. With such a patient cohort, one would expect the two groups to be assigned to the Preventive Care Program at similar rates.

Instead, when we ran the cases we had collected through the system, it identified ethnic Norwegian patients as eligible for the Preventive Care Program at a higher rate than ethnic minority patients. Out of all the cases we ran through the system (100 000 cases), the algorithm identified 5000 (5 %) of the cases as eligible for the Preventive Care Program. Out of the 5000 cases that were identified as qualified for the program, 3 350 (67 %) were ethnic Norwegians. If it had been the case that the Norwegian population was known to have worse health outcomes than the ethnic minorities in Storevik, this might have explained the difference. However, this does not seem to be the case. Quite on the contrary, there is some research suggesting that ethnic minorities in Norway have slightly worse health outcomes than the majority population. If anything, one would assume that minority patients should be assigned to the Preventive Care Program more often than other patients.

5.1.6 Implications

A long time has passed since Simon realised that he might have been discriminated against. However, although enormous resources have been spent on an audit of the AI-CDS system, he is still not sure that what happened amounted to discrimination under the law. At this point, he feels that the risk of losing in court is not worth it. Besides, a trial could take years. That is not how Simon intends to spend his retirement pension.

With reference to the clinical decision types mentioned in section 1.7, the AI-CDS system used by UHS in this case has elements of scarce resource allocation as well as preventive intervention. Throughout the remainder of the thesis, this case study serves as a reference point for discussing considerations of relevance to pre-deployment discrimination assessments of AI-CDS systems intended for these purposes.

The bias displayed by the system used by UHS is disadvantageous to ethnic minorities and might further increase the gap between these patients and the Norwegian majority in terms of utilisation of health services. Simon Tesfay's experience demonstrates the importance of subjecting AI-CDS systems to pre-deployment measures aimed at ensuring non-discrimination. It shows how challenging it can be for an individual patient to enforce the non-discrimination principle once it is suspected that discrimination has occurred. In Simon's case, it is notable that he even got more help investigating and building a case than most patients can hope for, because he happened to run into a journalist determined to unravel bias in AI-CDS systems.

5.2 Developing an AI-CDS System for the Prediction of Spine Surgery Outcomes (the 'NORspine project')

5.2.1 The Project

The second case study offered in this thesis is based on a project that aims for development and implementation of an AI-CDS system for the prediction of spine surgery outcomes. The project is a collaboration between the Norwegian registry for spine surgery (NORspine) hosted by the University Hospital of North Norway (UNN), UiT the Arctic University of Norway, the ICT unit of the Northern Norway regional health authority (Helse Nord IKT), Deepinsight (consultancy company) and DIPS (a Norwegian EHR system vendor). For the sake of disclosure, it is noted that the author of this thesis has been involved in one of the 'work packages' in the project pertaining to the consideration of legal solutions in an initial phase.

Although the project is not officially titled 'the NORspine project,' this title is used in this thesis, for ease of reference. It is important to note that the project is ongoing at the time of submitting this thesis. Several important decisions pertaining to the development process have not been finally settled and this case study does not necessarily give an accurate account of the project or the eventual results of the project. The purpose of the case study is to illustrate some of the bias-related challenges faced from a developer perspective, including the complexities of considering potential equality-related biases that might arise in an AI-CDS system.

The overarching goal is to improve outcomes from spine surgery. Clinicians involved in initiating the project note that a considerable proportion of patients have undesirable outcomes from spine surgery, according to self-reported data indicating that their conditions too often do not improve or become worse after surgery. If an AI-CDS system can help predict the outcome for individual patients, this would facilitate more accurate recommendations on who should have surgery and, thus, support the shared decision-making process between the patient and the surgeon. Hence, the clinical problem is to predict the outcome of surgery. The translation of this clinical problem into a computational problem raises several challenges such as what a good outcome is, how the perception of a good outcome may vary between patients, how reliable self-reported data are regarding surgery outcomes, and what exactly AI-CDS system should be instructed to predict.

While this thesis does not consider all aspects of the NORspine project, the following elaborates on the issues of data bias, feature selection, and the definition of the target variable. As explained in section 4.4.3, these decisions are potential sources of equality-related bias in AI-CDS systems.

5.2.2 Unequal Representation (Data Bias)

The ML algorithm in the project is trained on data concerning patients who have underwent spine surgery in the past. The data has been collected from NORspine.[499] Given that the Norwegian population is relatively homogenous in terms of ethnicity, it is to be expected that

[499] "Nasjonalt Kvalitetsregister for Ryggkirurgi," (web page) accessed 10 November, 2023, https://www.unn.no/fag-og-forskning/medisinske-kvalitetsregistre/nasjonalt-kvalitetsregister-for-ryggkirurgi.

the clear majority of patients represented in the data from this registry are ethnic Norwegians, and that there is considerably less data representing ethnic minorities. It is also worth noting that 'language barriers' is described as an exclusion criterion in the registry.[500] This exclusion criterion is more likely to exclude ethnic minorities from the registry, further marginalising these groups within the registry. Moreover, the registry partially relies on a digital public postal service (Digipost) to reach out to patients for data collection purposes.[501] This means that differences between groups in the use of the digital postal service might influence the demographic representation in the project's training data. If ethnic minorities use the Digipost service less than the majority population, this is yet another factor that potentially contributes to marginalisation and unequal representation.

The point of identifying these possible sources of bias is not to assert that these all apply to the NORspine project. For example, this thesis does not examine to what extent ethnic minorities really are underrepresented as users of the Digipost service. However, these are potential sources of bias that are relevant to consider in a project such as this. When data from the registry is used for training purposes, there is a risk that representation bias, due to any of the abovementioned reasons, could lead to disparate accuracy – an AI-CDS system that performs better for patients associated with the majority population, compared to ethnic minority patients.[502]

5.2.3 Feature Selection

When considering which feature variables to include in the training data, it is important to consider any known sources of equality-related bias that may influence the training data. As a starting point, one might observe that men and women appear to have different outcomes and the fact that there are anatomical differences between men and women. Consequently, the inclusion of the patient's sex as a feature variable may be considered because sex appears to be a medically relevant factor. However, inclusion of sex as a feature variable raises questions of exactly why men and women have different outcomes and whether this is in fact due to the anatomical differences between these two groups.

---

[500] E. Mikkelsen et al., "The Norwegian Registry for Spine Surgery (Norspine): Cohort Profile," *European Spine Journal* (Sep 17 2023): 3715, https://doi.org/10.1007/s00586-023-07929-5.

[501] Mikkelsen et al. (2023) 3716.

[502] Section 4.3.3, cf. section 4.4.2.4.

In the context of spine surgery data, the research suggesting that women have worse outcomes when operated on by men points out a pattern that could have implications for the feature selection.[503] If this is the case in the training data and there are more male surgeons than female surgeons in Norway, there is reason to believe that the trained model will suggest that surgery is recommendable for women less often than for men, if patients' sex is used as a feature variable. However, it is far from clear *why* women have worse outcomes when they are operated on by men.

Furthermore, the fact that women have worse outcomes when operated on by men indicates that the *surgeon's* sex is a predictive factor and, thus, a relevant feature variable which should be included in the training data. However, to do this in practice would have some intricate implications. It would only make sense to include the sex of the surgeon as a feature variable if it was also possible to use this information as an input when using the AI-CDS system to generate a prediction for the patient. In theory, the surgeon could be decided upon in advance, which means that the sex of the surgeon can be used as input for the model. The question is what good it would do. If it is the case in the relevant population that women actually have worse outcomes when the surgeon is a man, the inclusion of the surgeon's sex as an input feature would just mean that female patients with male surgeons would have less optimistic predictions. Arguably, the main benefit of including the surgeon's sex is interpretability; the inclusion of this feature enables detection of the correlation between unsuccessful outcomes for women and the sex of the surgeon. The correlation can now be used when explaining to the patient why the model predicts an unsuccessful outcome (the sex of the surgeon would only be one of many factors here) and help the patient and clinician make an informed choice.

5.2.4 Defining the Target Variables for Outcomes After Spine Surgery

The prediction of successful outcomes from spine surgery requires that a 'successful outcome' is somehow defined. 'Successful' is not one piece of information that can be observed within the training data. Consequently, what a 'successful outcome' means may be defined as one or more observable variables that an AI model can predict.

The definition of target variables in this project raises several questions that will be considered by an interdisciplinary development team, involving computer scientists and

[503] Wallis et al. (2022).

clinicians, as the project progresses. One challenge is how to accommodate the fact that patients have different preferences and, thus, different views on what a successful outcome is. To the extent that a successful outcome is defined by an AI development team, this implies that the development team chooses the preferences and values on behalf of all patients on whom the system will be used. This could in theory lead to an AI-CDS system that is better adapted to preferences commonly held by the majority population or by groups that are overrepresented in the development team. For example, men might be overrepresented in a development team. In the NORspine project, it is being considered how the preferences of individual patients can be included in the shared decision-making process.

One question with a potential for bias in this project is how far into the future, following surgery, the target variable(s) should be measured. One year after surgery? Five years? Ten years? If the target variable includes a point in time that is very far into the future, this could lead to representation bias,[504] in the form of underrepresentation of older patients compared with the actual population of patients that have spine surgery (which usually has many older patients).

As regards bias to the detriment of groups defined by ethnicity or sex, there may be target variables which are ingrained with structural inequalities, stereotyping or prejudice.[505] One such variable is the level of pain experienced by a patient. There are two aspects of this issue. Section 4.4 highlighted some indications suggesting that health personnel sometimes underestimate the pain experienced by certain ethnic minorities. This is one aspect, which relates to past stereotyping or prejudice. This implies that the recommendations that clinicians have given to certain groups in the past may have been based on an underestimation of the pain experienced by these patients. Consequently, the pain level, as a target variable, could mean different things for different groups. The other problematic aspect of using pain level as a target variable is the possibility that there might indeed be systematic differences in how different groups report self-experienced pain. For example, if women tend to report lower levels of experienced pain than men while AI developers assume that pain levels mean the

[504] See section 4.4.2.3.

[505] See sections 4.4.2.2. and 4.4.2.6.

same for men and women, this will imply that the target variable means different things for different groups.

To mitigate the risk of bias associated with pain level reports, pain levels as such are typically not relied on to define target variables related to spine surgery. A more common approach in this field is to use the Oswestry Disability Index (ODI) to measure outcomes for back pain.[506] This is the approach that is being considered in the NORspine project. In practice, patients fill out a questionnaire which leads to a score indicating their pain-related disability level. While it is assumed that this approach is less prone to group-level differences in the reported data, the possibility of such differences cannot be ruled out.

5.2.5 Implications

This case study illustrates some of the many decisions that a development team needs to make, particularly as regards how to translate a clinical problem into a computational problem that an AI system can solve. Moreover, it gives concrete examples of several of the potential sources of equality-related biases described in section 4.4. While the fictional case of Simon Tesfay in section 5.1 looked retrospectively at the experience of an individual, the NORspine case study takes the perspective of a development team in the early stages of developing an AI-CDS system. The case study demonstrates the complexity of the potential sources of bias and the importance of seeking knowledge about the sometimes subtle patterns that may be reflected in training data, feature variables and target variables. The case study is referred to at several instances throughout the remainder of the thesis.

[506] American Academy of Orthopaedic Surgeons, "Oswestry Low Back Disability Questionnaire." https://www.aaos.org/globalassets/quality-and-practice-resources/patient-reported-outcome-measures/spine/oswestry-2.pdf.

PART III: LEGAL FRAMEWORK AND PRE-DEPLOYMENT DISCRIMINATION ASSESSMENT REQUIREMENTS

6 Legal Framework

6.1 Introduction

This chapter introduces the relevant framework of EU law applicable to AI-CDS systems. Given the objective of the thesis, particular attention is given to EU non-discrimination law and the relationship between EU non-discrimination law and the other relevant acts of EU law applicable to AI-CDS systems.

The legal framework governing AI-CDS systems potentially consists of at least three different layers. It varies to what extent these layers are filled in with EU law versus national and constitutional laws. One layer consists of the laws governing the provision of healthcare services, including any laws that specifically target clinical decision-making. Section 6.2 notes that there is little EU law of relevance within this layer. Another layer consists of fundamental rights law, including laws that implement fundamental rights. Given the objective of this thesis, the right to non-discrimination and the manifestation of this right in the Equality Directives take a central role (section 6.3). However, given the data-driven nature of AI-CDS systems and the vast utilisation of personal data that occurs during training and use of these systems, the right to data protection and, particularly, the EU's General Data Protection Regulation (GDPR) are relevant (section 6.6).[507] Fundamental rights laws across the national and supranational levels typically overlap and intersect.

The third layer consists of product legislation – laws that govern various products and their manufacturers. In the EU, product laws for different products follow a well-established structure – the 'New Legislative Framework' (NLF).[508] NLF laws set out the requirements that products and manufacturers must comply with before a product can be placed on the EU

[507] 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.

[508] "New Legislative Framework," (web page) accessed 11 November, 2023, https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en;

market or put into service. It regulates products such as medical devices, lifts, toys, radio equipment, etc.[509] Traditionally, the NLF acts focus on ensuring the safety and performance of products.[510] This is the case, for example, with the Medical Device Regulation (MDR), which is discussed in section 6.5. The proposed AI Act is also modelled after the NLF and can be categorised as a product act. However, while addressing safety and performance (as is typical for NLF acts), the AIA is also strongly oriented towards fundamental rights protection (section 6.4). Product laws in the EU are instrumental to the functioning of a common European marketplace and therefore require a high level of harmonisation between Member States.

6.2 Legal Regulation of Clinical Decision-Making

In addition to EU law, the legal regulation of AI-CDS systems depends on the national laws of EU Member States and their international human rights obligations. However, there exist few laws specifically aimed at regulating clinical decision-making, resulting in limited regulatory supervision over these decisions. Clinical assessments predominantly fall under the purview of medical experts, who draw upon their expertise, derived from experience, education, and training.

Exactly how clinical decisions should be made, is typically not regulated in detail at a legislative level, although the specificity of health laws can differ among countries. Nevertheless, there are generally overarching legal principles that guide clinical decision-making. Paramount among these is the standard of care principle, which mandates that

---

Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys; Directive 2014/33/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to lifts and safety components for lifts (recast); Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC; Regulation 2017/745 (MDR).

[510] The New Legislative Framework was established through alignment of existing product safety legislation in accordance with Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a Common Framework for the Marketing of Products, and Repealing Council Decision 93/465/EC.

healthcare must be provided in accordance with certain minimum standards.[511] These standards are informed by fundamental principles of medical ethics. Moreover, clinicians are obligated to adhere to data protection obligations when they process patient information for clinical purposes. These obligations are anchored in the EU's General Data Protection Regulation (GDPR), which is typically augmented by more specific rules for data processing within the health sector at the national level.

As regards non-discrimination, it is possible that specific laws may exist within the legislation of Member States to combat discriminatory practices in clinical decision-making. Nevertheless, such specific rules are likely to be an infrequent occurrence. That said, the general principles of non-discrimination, found in Member State constitutions or in international law, remain applicable.

To promote equality and to protect citizens from discrimination, a right to non-discrimination is recognised in international human rights conventions at the level of the United Nations,[512] the Council of Europe – most notably the European Convention of Human Rights (ECHR),[513] in EU law,[514] and in the national laws and constitutions of many countries.[515] This right is implemented, and its enforcement is facilitated, via legislative acts at both national and supranational levels. Despite this, it is notable that regulatory oversight within the realm of clinical decision-making does not typically centre on ensuring compliance with non-discrimination laws.

---

[511] For example, this principle is incorporated in § 4 of the Norwegian Health Personnel Act (Act 2 July 1999 No. 64 on Health Personnel).

[512] International Convention on the Elimination of all forms of Racial Discrimination (ICERD) (adopted by the UN General Assembly 21 December 1965); International Covenant on Social, Economic and Cultural Rights (CSECR) (adopted by the UN General Assembly 16 December 1966), Article 2; International Covenant on Civil and Political Rights (CCPR) (adopted by the UN General Assembly 19 December 1966), Article 2; Convention on the Elimination of all forms of Discrimination Against Women (CEDAW) (adopted by the UN General Assembly 18 December 1979); Convention on the Rights of Persons with Disabilities (CRPD) (adopted by the UN General Assembly 24 January 2007).

[513] Article 14 ECHR.

[514] Section 6.3.

[515] In the Norwegian Constitution, § 98 enshrines a generally articulated principle of equality.

6.3 EU Non-Discrimination Law

6.3.1 Introduction to EU Non-Discrimination Law

EU non-discrimination law has evolved through the CJEU's case law based on the equal treatment/non-discrimination principle enshrined in Article 141 of the Treaty establishing the European Community (now Article 157 TFEU). This provision specifically obligates Member States to protect the principle of equal pay for male and female workers.[516] It was not until the advent of the 1997 Treaty of Amsterdam that the EU acquired legal authority to legislate against discrimination based on other characteristics than sex.[517] This Treaty broadened the EU's powers, authorising it to take action to combat discrimination on the basis of "sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation" (this wording is now found in Article 19 TFEU).[518]

This expanded mandate laid the foundation for the creation of the EU directives collectively referred to as the Equality Directives. Each of these directives prohibits discrimination based on certain characteristics and are applicable to specified sectors. Directive 2000/43 ('Racial Equality Directive'/'RED') prohibits discrimination based on ethnicity; Directive 2000/78/EC ('Employment Equality Directive'/'EED') prohibits discrimination on the basis of "religion or belief, disability, age or sexual orientation."[519] Directive 2004/113/EC ('Goods and Services Equality Directive'/'GSED') concerns equal treatment between men and women in the access to and supply of goods and services. The scope of application of these directives in relation to the provision of healthcare services is discussed in section 6.3.2 below.

---

[516] Article 8 TFEU states that the Union shall aim, in all its activities, "to eliminate inequalities, and to promote equality, between men and women." The non-discrimination principle is also expressed in relation to discrimination on grounds of nationality in Article 18 TFEU.

[517] Claire Kilpatrick, "Non-Discrimination," in *The Eu Charter of Fundamental Rights : A Commentary*, ed. Steve Peers et al. (London: Hart Publishing, 2014), 582-83.

[518] European Union, Treaty of Amsterdam Amending the Treaty Establishing the European Community, the Treaties Establishing the European Communities and Related Acts, 10 November 1997, Article 2(7) (which inserts into the TEC a provision corresponding to what is now Article 19 TFEU).

[519] As regards equal treatment of men and women in matters of employment and occupation, the Employment Equality Directive is supplemented by Directive 2006/54/EC (Recast Directive).

The Equality Directives obligate the Member States to implement the non-discrimination principle in a manner that provides, at minimum, the same level of protection as the rules set out in the Directives.[520] If a Member State fails to fulfil this obligation, the European Commission can initiate proceedings against it.[521] However, in practical terms, the preliminary ruling procedure under Article 267 TFEU serves as the primary canal through which the CJEU communicates its interpretation of the Directives to the Member States.[522]

Within their respective areas of application, the Equality Directives are specific expressions of the general non-discrimination principle in EU law.[523] This principle is reckoned as one of the fundamental principles of EU law. Given that the Equality Directives embody this same principle, they must be interpreted in conjunction with one another, and in alignment with the CJEU's jurisprudence on the general non-discrimination principle. At their core, these Directives rely on a unified definition of discrimination, despite their varied areas of application.

The Equality Directives prohibit four types of discrimination,[524] however, this thesis focuses on two: direct and indirect discrimination. By definition, direct discrimination occurs where one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of a protected characteristic.[525] On the other hand, indirect discrimination is where an apparently neutral provision, criterion or practice would put persons with a certain characteristic at a particular disadvantage, unless the disputed measure or practice is considered to be a proportionate means of achieving a legitimate aim (objective

---

[520] Article 6 RED.

[521] Article 258 TFEU.

[522] Nigel Foster, *Foster on Eu Law*, 7th ed. (Oxford: Oxford University Press, 2019), 187, etc.

[523] Ellis and Watson (2012) 130 and 42; CHEZ, C-83/14, para 42.

[524] In addition to direct and indirect discrimination, the Directives prohibit harassment (as defined, e.g., in Article 2(3) RED) and actions that amount to "an instruction to discriminate" (e.g., Article 2(4) RED).

[525] Article2(2)(a) RED; Article 2(2)(a) EED; Article 2(a) GSED; Directive 2006/54/EC of the European Parliament and the 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 (Recast), Article 2(1)(a).

justification).[526] The definitions of direct and indirect discrimination in EU law are analysed in Part IV of the thesis, in the pursuance of its main objective, to develop methodological elements of assessing discrimination in an AI-CDS system.

The non-discrimination principle is now enshrined in Article 21 of the EU Charter of Fundamental Rights. The Charter was introduced in 2001, but did not become binding on EU Member States until the Lisbon Treaty accorded it the status of primary EU law in 2009.[527] Article 21(1) of the Charter prohibits discrimination on "any ground," and enumerates a series of protected characteristics. These include, but are not limited to, "race, colour, ethnic or social origin." However, the Charter does not provide a specific definition of 'discrimination.' What discrimination means in Article 21(1) of the Charter must be ascertained through interpretation of the principle that this provision enshrines. This necessitates consideration of other related provisions in primary or secondary EU law that express the same principle, as well as the CJEU's interpretation of these provisions. Neither Article 21 of the Charter or the Equality Directives *establish* the non-discrimination principle in EU law.[528] Rather, it is a general principle of EU law, derived from international human rights law and the constitutional traditions of EU Member States.[529]

The applicability of the EU Charter is generally a subject of much academic discussion, and one that has received a lot of attention in CJEU case law in recent years. The Charter applies to EU institutions and to the Member States only "when they are implementing EU law," cf. Article 51(1) of the Charter. This has led to uncertainty about when a Member State is "implementing EU law." The CJEU has tried to clarify this issue in several cases, and it has

---

[526] Article 2(2)(b) RED; Article 2(2)(b) EED; Article 2(b) GSED; Article 2(1)(b) Directive 2006/54/EC.

[527] Article 6(1) TEU as amended by the Treaty of Lisbon in 2007: Treaty of Lisbon 13 December 2007 Amending the Treaty on European Union and the Treaty Establishing the European Community. Even before the Treaty of Lisbon provided the Charter with a formal legal status, the CJEU was referring to the Charter in its case law: e.g., Judgment (GC) of 13 March, 2007, Unibet, C-432/05, ECLI:EU:C:2007:163, para. 37; Ellis and Watson (2012) 119.

[528] Judgment (GC) of 17 April, 2018, Egenberger, C-414/16, ECLI:EU:C:2018:257, para. 78.

[529] Ursula O'Hare, "Enhancing European Equality Rights: A New Regional Framework," *Maastricht Journal of European and Comparative Law* 8, no. 2 (2001): 159; Judgment (GC) of 22 November, 2005, Mangold, C-144/04, ECLI:EU:C:2005:709, para. 74; Mangold, C-144/04; CHEZ, C-83/14, para. 42.

come to apply a rather broad interpretation of the phrase "implementing EU law" (broader than the wording itself would suggest).[530] In *Åkerberg Fransson*, the CJEU holds that the Charter applies where a Member State action falls within "the scope of" EU law.[531] However, the Court does not clearly define what it means by the "scope of EU law." In the *YS* ruling, the CJEU clarifies that the rights in the Charter are "applicable in all situations governed by EU law and they must be complied with inter alia where national legislation falls within the scope of EU law," provided that "EU law imposes specific obligations on Member States with regard to the situation at issue."[532] In the light of these directions, Article 21 of the Charter does not apply to clinical decision-making in general, because there is no EU law regulating clinical decision-making in general. There may be specific aspects of certain clinical decisions that are regulated by EU law, such as when EU law lays down safety requirements for blood transmissions, organ donations and transplantations.[533] Within the scope of such EU laws, Article 21 of the Charter is applicable.[534]

According to the same reasoning, Article 21 is also applicable to AI-CDS systems to the extent that these systems are governed by EU law. However, in cases where the Equality Directives are applicable, it is these Directives which must be applied, because they are more specific expressions of the non-discrimination principle within their areas of application. The following section discusses the applicability of the Equality Directives in the context of assessing discrimination in AI-CDS systems and, thus, whether an assessment methodology should be based on the Directives or Article 21(1) of the Charter. It also underscores why this distinction matters, even though the Directives and the Charter express the same fundamental principle of non-discrimination.

---

[530] Kilpatrick (2014) 581-82.

[531] Judgment (GC) of 26 February, 2013, Åkerberg Fransson, C-617/10, ECLI:EU:C:2013:105, para. 21.

[532] Judgment of 24 September, 2020, YS, C-223/19, ECLI:EU:C:2020:753, paras. 78-79.

[533] It should be noted that the Medical Device Regulation arguably already brings AI-CDS systems in under the scope of EU law, regardless of the AI Act. However, the MDR does not require any particular fundamental rights-oriented assessments before a medical device can be placed on the market.

[534] Judgment of 29 April, 2015, Léger, C-528/13, ECLI:EU:C:2015:288, paras. 46-48.

6.3.2 Applicability of the Equality Directives to AI-CDS Systems

In cases where a discrimination assessment based on EU law is required, this means that the non-discrimination principle in EU law is applicable at least as far as the mandatory discrimination assessment is concerned. To what extent pre-deployment assessments based on the non-discrimination principle are required for AI-CDS systems, is discussed in chapter 7. However, at this point, it is worth making a forward reference to the conclusions in chapter 7: There are, indeed, discrimination assessment requirements in the AI Act and they do refer to EU non-discrimination law. Therefore, EU non-discrimination law is applicable in these assessments. However, chapter 7 also finds that the discrimination assessment requirements in the AIA do not specify whether the assessments should be based on Article 21(1) of the Charter or the Equality Directives. Instead, it is concluded that the AIA most likely refers to the provisions of EU law that would apply to the underlying subject matter. If this is the Equality Directives, then a pre-deployment discrimination assessment should rely on the Equality Directives. If, on the contrary, the reference goes to Article 21(1) of the Charter, it follows that a discrimination assessment should be based on that provision.

In the light of this knowledge, the question arises whether there is any difference between basing a discrimination assessment methodology on the Equality Directives or Article 21(1) of the Charter. As noted in the previous section, these provisions express the same fundamental principle of EU law. However, there is one peculiar difference in the wording of Article 21(1) of the Charter, compared to the Equality Directives. Unlike the Equality Directives, Article 21(1) of the Charter does not explicitly distinguish between direct and indirect discrimination. At the same time, the possibility of objectively justifying potentially discriminatory practices according to Article 21(1) is available under Article 52(1) of the Charter. Hence, the question arises whether practices that constitute potential direct discrimination can be justified pursuant to Article 52(1) of the Charter. If they can, this means that there is, potentially, an important difference between conducting an assessment based on Article 21(1) of the Charter and basing it on the Equality Directives. Consequently, it seems that it might matter whether a pre-deployment discrimination assessment methodology is based on one or the other. Whether this is merely a difference in wording or an actual difference that survives further interpretation, is worth examining.

Even though the distinction between direct and indirect discrimination is not reflected in the wording of Article 21(1) of the Charter, such a distinction has been recognised in the CJEU's

case law under Article 21(1) of the Charter.[535] However, when applying Article 21(1) of the Charter, as the CJEU does in cases where the Equality Directives do not apply but where other acts of EU law are nonetheless applicable, the Court tends not to distinguish as clearly between direct and indirect discrimination as it does under the Equality Directives. For example, under Article 21(1), the Court tends to speak of "a difference in treatment" and assume that such a "difference in treatment" can be objectively justified under Article 21(1) of the Charter.[536] The notion of a "a difference in treatment" in this context refers to the non-discrimination principle and includes direct as well as indirect discrimination. This indicates that justification under Article 52(1) is available for direct and indirect discrimination alike. In academic literature, this view is supported by Ward, who finds (under some doubt) that, in cases where the Equality Directives do not apply, justification according to Article 52(1) of the Charter applies to direct as well as indirect discrimination.[537]

While the CJEU has never explicitly discussed the difference between justification of direct and indirect discrimination in cases decided on the basis of Article 21(1) of the Charter, the *Léger* ruling is of particular interest because the facts of the case arguably amount to direct discrimination.[538] The case arose when a medical doctor refused blood donation from the applicant (Mr. Léger) based on the fact that he had had sexual relations with another man. This decision was made in accordance with a French decree, and it was the legality of the decree that was disputed. Thus, the case stood between Mr Legér and the French authority responsible for the decree.[539] The question referred to the CJEU was presented to it as a

---

[535] Judgment of 29 October, 2020, Veselības Ministrija, C-243/19, ECLI:EU:C:2020:872, paras. 39-40.

[536] e.g., Veselības Ministrija, C-243/19.

[537] Ward argues that, at least in cases where the disputed measure is initiated by a Member State acting outside of the scope of the RED and the EED, "the distinction between the range of justifications available with respect to direct and indirect discrimination" is not relevant. Although she finds it "questionable", Ward further asserts that "[i]n these contexts, irrespective of whether the discrimination is direct or indirect, the justification test provided by Article 52(1) of the Charter applies": Angela Ward, "The Impact of the Eu Charter of Fundamental Rights on Anti-Discrimination Law: More a Whimper Than a Bang?," *Cambridge Yearbook of European Legal Studies* 20 (2018): 34, https://doi.org/10.1017/cel.2018.11.

[538] Léger, C-528/13.

[539] i.e., *Ministre des Affaires sociales, de la Santé et des Droits des femmes*.

matter of interpretation of Directive 2004/33/EC concerning technical requirements for blood and blood components. That directive laid down certain eligibility criteria for blood donors, one of which concerned sexual behaviour putting the donor at a high risk of acquiring severe infectious diseases that can be transmitted by blood.[540] The eligibility criteria for blood donors were intended to protect the health of blood recipients, especially against the risk of infectious diseases. The French decree implementing the directive entailed that if a man had had sexual relations with other men, this was to be deemed as a permanent contraindication to blood donation, due to the risk of HIV infections. The CJEU had to consider whether such a criterion for blood donation was lawful under EU law.

Although the CJEU does not mention it, the facts of *Léger* arguably amount to direct discrimination on the basis of sexual orientation. Despite this circumstance, the CJEU holds the door open to justification according to Article 52(1) of the Charter. It instructs the referring court to consider, as part of the objective justification test, whether there are ways to "identify more precisely the type of behaviour presenting a risk for the health of recipients, in order to impose a less onerous contraindication than a permanent contraindication for the entire group of men who have had sexual relations with a man."[541] It does not appear that the CJEU's reasoning around justification and proportionality is affected by the directly discriminatory nature of the French decree at issue.

Based on the *Léger* ruling and the wording of Article 21(1) of the Charter, it appears that justification of practices that amount to potential direct discrimination is possible in areas where Article 21, cf. Article 52(1) of the Charter, is applied. This conclusion, which is uncertain because the CJEU does not explicitly acknowledge that the *Léger* facts amount to direct discrimination, is relevant in circumstances where the Equality Directives do not apply. The ramification is that it does indeed matter whether a discrimination assessment methodology relies on Article 21(1) of the Charter or the Equality Directives. The difference is that an assessment methodology based on Article 21(1) of the Charter would lead to an assessment that is more lenient towards the deployment of an AI-CDS system potentially amounting to direct discrimination, as long as there is an objective justification for deploying the system. Where the Equality Directives apply, direct discrimination can only be justified in

---

[540] Léger, C-528/13, para. 18.

[541] Léger, C-528/13, para. 66.

case specific statutory exceptions are available, and there are no such exceptions of relevance to AI-CDS systems, currently.

As mentioned, the AIA does not explicitly clarify whether it refers to the Charter or the Equality Directives. Rather, the decisive question is whether the Equality Directives apply to the underlying subject matter, in which case the assessments mandated by the AIA must be based on these Directives. Hence, the question is whether the Equality Directives are applicable to clinical decision-making. This is not an uncontroversial question. The Equality Directives that potentially apply to clinical decision-making are the Goods and Services Directive, concerning sex discrimination, and the Racial Equality Directive, concerning ethnic discrimination. On their surface, according to their wording, both directives appear to apply to clinical decision-making. However, there are reasons why one might argue that the scope of applicability should be interpreted narrower than what the wording alone suggests, in the context of healthcare. The following first considers the wordings of the GSED and the RED, before turning to considerations that might lead to a narrower interpretation than what the wordings suggest.

The GSED applies (with a few exceptions)[542] "to all persons who provide goods and services, which are available to the public irrespective of the person concerned as regards both the public and private sectors, including public bodies, and which are offered outside the area of private and family life…"[543] The European Commission has clarified that public authorities providing services are governed by the GSED unless they are exercising public authority without any element of provision of a service.[544] Healthcare services generally have an element of provision of a service and there is no exception for healthcare services in the GSED.[545] Healthcare services are therefore covered by the GSED, regardless of whether they

---

[542] See Articles 3(3) and 3(4) GSED.

[543] Article 3(1) GSED.

[544] *Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee*, (5 May 2015), 3-4, https://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2015/0190/COM_COM(2015)0190_EN.pdf.

[545] Unlike the Services Directive, which excludes healthcare services according to Article 2(f): Directive 2006/13/EC of the European Parliament and of the Council of 12 December 2006 on Services in the Internal Market.

are provided by public, state-owned institutions or private actors, and regardless of how a Member State has organised and regulated the provision of healthcare services.[546] Recital 12 of the GSED's preamble also points in this direction. This recital clarifies that "differences between men and women in the provision of healthcare services, which result from the physical differences between men and women, do not relate to comparable situations and therefore, do not constitute discrimination." This recital would be superfluous if the GSED was not applicable to the provision of healthcare services.

Similarly, "healthcare" is covered by the wording of the RED.[547] Furthermore, the RED's preamble mentions healthcare as one of the areas where Member States should take specific action against discrimination.[548] Although the RED does not provide any indications as to what it means by "healthcare," a common understanding of this word encompasses the provision of healthcare services, including clinical decision-making.

Despite the wording, which for both directives imply that they are broadly applicable to the field of healthcare, the question of applicability proves more complex when one considers the generally hazy issue of EU competencies in health-related matters.[549] Article 168(7) TFEU (formerly Article 152 TEC) explicitly states that "Union action shall respect the responsibilities of the Member States for the definition of their health policy and for the organisation and delivery of health services and medical care." Against this background, Ellis and Watson find that, due to the EU's limited powers within healthcare, it is "unclear to what extent the Race Directive confers a right to equality of treatment" in this field.[550] Di Federico takes a more pessimistic view, as he argues that patients cannot invoke the RED unless they are in situations covered by a specific directive applicable in the field of healthcare, such as

---

[546] "Services" in the GSED means the same as "services" in Article 57 TEU (previously Article 50 TEC). The CJEU has clarified that the existence of a "service" does not presuppose that the service recipient pays for the service themselves: Judgment of 12 July, 2001, Geraets-Smits and Peerbooms, C-157/99, ECLI:EU:C:2001:404, para. 57; Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee (2015): 3.

[547] Article 3(e) RED.

[548] Recital 12 RED.

[549] Ellis and Watson (2012) 363-64.

[550] Ellis and Watson (2012) 438.

Regulation 883/2004/EC (Social Security Regulation), Directive 2011/24/EU (Patients Directive on Cross-Border Health), or Article 56 TFEU (freedom to provide services).[551]

One argument that supports a wide interpretation of the scope of application of the RED and GSED, is the status that the prohibition of discrimination based on ethnicity and sex holds in EU law and in the Member States of the EU.[552] With the 2007 Treaty of Lisbon, the EU Charter of Fundamental Rights was given the same status as the TEU and TFEU. The importance of this development when interpreting the scope of secondary EU law is confirmed by the CJEU's ruling in *CHEZ,* where the Court argues for a broad interpretation of the personal scope of the RED, considering the "nature of the right which it seeks to safeguard" and its status as a general principle of EU law.[553] Moreover, the CJEU's ruling in *Maniero,* concerning the RED's application in the field of education, further affirms that the RED cannot be interpreted restrictively, as it would undermine the fundamental goal of ensuring effective protection against discrimination.[554] Due to the status of sex discrimination as a general principle of EU law, the same arguments would apply to the application of the GSED.

The question of whether the RED and the GSED apply to healthcare remains unsettled. The abovementioned case law does not concern healthcare, directly. However, given the EU's commitment to combatting racial and sex-based discrimination, along with the CJEU's broad interpretation of the Equality Directives' scope, it is conceivable that the CJEU might extend the application of the RED and the GSED to clinical decision-making. Moreover, it is worth emphasizing that applying the principle of non-discrimination based on sex or ethnicity would

---

[551] Giacomo Di Federico, "Access to Healthcare in the European Union: Are Eu Patients (Effectively) Protected against Discriminatory Practices?," in *The Principle of Equality in Eu Law*, ed. Lucia Serena Rossi and Federico Casolari (Cham: Springer International Publishing, 2017), 235-36 and 51.

[552] Based on similar reasoning, Ellis and Watson note that the CJEU could give a "more generous interpretation to 'social advantages' in the field of race than it has for free movement": Ellis and Watson (2012) 364.

[553] CHEZ, C-83/14, para. 42.

[554] Judgment of 15 November, 2018, Maniero, C-457/17, ECLI:EU:C:2018:912, para 36; Rossen Grozev, "A Landmark Judgment of the Court of Justice of the Eu - New Conceptual Contributions to the Legal Combat against Ethnic Discrimination," *The Equal Rights Review* Fifteen (2015): 172.

not be tantamount to harmonizing healthcare provision within the EU, as that falls outside the EU's competencies.

In conclusion, it is submitted that there is a predominance of arguments in favour of considering both the RED and the GSED applicable in the context of clinical decision-making as a matter of *lex lata*. Consequently, it is asserted that these Equality Directives are applicable to clinical decision-making in the Member States. The implication is that the methodological elements that this thesis aims to develop, should primarily be developed on the basis of the Equality Directives. This leads to methodological elements based on the general rule that direct discrimination is not justifiable. Hence, it becomes important in a pre-deployment assessment to consider whether an AI-CDS system could lead to direct or indirect discrimination.[555]

6.4 The AI Act (AIA)

6.4.1 Overview

Recognising the possibility that AI systems can interfere with the safety and fundamental rights of EU citizens, the EU legislature has taken the initiative to create a common European regulatory framework for artificial intelligence – the Artificial Intelligence Act ('AI Act'/'AIA').[556] For an explanation of how this thesis refers to this proposed regulation and its different negotiation versions, see section 2.5.

The legislative aims of the AIA are to pursue a high level of protection of health, safety and fundamental rights, whilst promoting the implementation of AI systems and enhancing the internal market by ensuring free movement of AI-based goods and services across Member States.[557] The AIA is meant to lay down only "the minimum necessary requirements to address the risks and problems linked to AI."[558] This way, the European Commission intends

[555] See chapter 8.

[556] The AIA was proposed by the European Commission following its own 2020 White Paper on AI – A European Approach to Excellence and Trust: White Paper on Artificial Intelligence - a European Approach to Excellence and Trust (2020); as well as a request from the European Parliament to lay down a framework addressing ethical challenges associated with AI technologies: section 3.3.1.

[557] Recital 1 AIA (EP).

[558] AIA (EC), Explanatory Memorandum, 4.

to avoid disproportionate constraints on innovation and AI implementation.[559] Furthermore, of particular relevance to this thesis, the Explanatory Memorandum accompanying the Commission's proposal asserts the Commission's ambition to "minimise the risk of algorithmic discrimination."[560] This statement indicates that non-discrimination is among the pivotal objectives of the AIA. However, the AIA does not explicitly lay down a prohibition of discrimination, and it does not provide a definition of discrimination. This is an intentional choice, as the AIA is meant to supplement existing EU law where necessary to address AI-specific risks, rather than reiterating already established principles of EU law.[561]

To ensure safety and effective protection of fundamental rights, AI systems are regulated in the AI Act as products that must be evaluated based on safety requirements and fundamental rights, before they can be deployed within the EU. Thus, the AI Act is modelled after existing EU product safety legislation (the 'New Legislative Framework').[562] However, while traditional EU product safety laws are predominantly concerned with protecting the health and safety of EU citizens, the AI Act is equally concerned with fundamental rights, including the right to non-discrimination.[563]

The AIA is founded upon on a legislative risk assessment, through which the legislature has mapped the conceivable risks of AI systems. This risk assessment has led to the complete prohibition of certain types of AI systems by the AIA. However, most AI systems, including the majority of AI-CDS systems, are not subject to this prohibition. The AI systems that are not prohibited are divided into two main categories based on their anticipated level of risk to health, safety and fundamental rights: 'high-risk' AI systems and non-high-risk AI systems. According to the AI Act's risk classification scheme, AI-CDS systems will frequently be classified as 'high-risk' AI systems.[564]

The main duty subjects under the AIA are the 'provider' and the 'deployer' or 'user' (depending on the final negotiations), particularly those providing or deploying high-risk AI

---

[559] Ibid.

[560] Ibid.

[561] Hauglid and Mahler (2023) 2.

[562] "New Legislative Framework."

[563] Recitals 1, 2, 5, 10, 13, 15, and 27 AIA (EC); Hauglid and Mahler (2023) 6.

[564] Hauglid and Mahler (2023) 13.

systems.[565] This thesis uses the term 'deployer' rather than 'user,' in accordance with the European Parliament's Compromise Text. The majority of requirements are directed towards the 'provider,' a role defined as the entity "that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark."[566] A 'deployer' is defined as the entity "using an AI system under its authority." Consequently, AI developers are 'providers' under the AIA. This also applies for healthcare institutions where AI-CDS systems are developed in-house. In cases where healthcare institutions purchase AI-CDS systems from an external vendor, the vendor is the 'provider,' and the healthcare institution is the 'deployer.'

Title III AIA (Articles 9-15) governs 'high-risk' AI systems. These provisions stipulate that 'high-risk' AI systems and their providers and, in some cases, deployers, must satisfy certain requirements. These requirements relate to risk management measures, training data, technical documentation, record-keeping, transparency and information duties, human oversight, accuracy, robustness and cybersecurity. 'High-risk' AI systems are also subjected to certain requirements pertaining to measures aimed at ensuring compliance before the deployment of an AI system (preventive compliance measures), including a conformity assessment and certification scheme. The specifics of these preventive compliance measures are further examined in chapter 7, to ascertain the extent to which they require pre-deployment discrimination assessments.

The European Commission's AIA proposal does not provide for enforcement through individual remedies. Instead, the Commission assumes that ex post remedies are already available to individuals whose fundamental rights are infringed. The absence of individual ex post remedies is typical of the product safety approach that the AIA relies on. For example, should a patient suffer harm due to the use of a medical device, the remedies for the patient are not found in the Medical Device Regulation. Instead of directly addressing individual remedies, the MDR and other product safety acts refer these matters to other laws such as

---

[565] The term 'user' is relied on in the European Commission's proposal, whereas the 'deployer' term is introduced in the European Parliament's Compromise Text.

[566] Article 3(4) AIA (EP).

Product Liability Directive,[567] as well as national tort law and procedures. Specifically in relation to AI systems, the European Commission has proposed a new AI Liability Directive concerning the allocation of civil, non-contractual liability for losses and damages occurring in connection with the use of AI systems.[568] Additionally, certain changes to the Product Liability Directive have been proposed to facilitate its application to AI systems.[569] However, the lack of remedies in the Commission's AIA proposal received criticism.[570] In response, the European Parliament's Compromise Text contains provisions on certain individual and collective (for consumers) remedies.[571] To what extent these remedies end up being adopted, remains to be seen at the time of writing. It is worth noting that, because the AIA does not lay down an outright prohibition of discrimination, complaints and remedies in relation to algorithmic discrimination would probably have to rely on an AI provider or deployer's failure to fulfil their AIA duties. Potentially, a claim based on the AIA could assert that a provider or deployer has failed to conduct a proper pre-deployment discrimination assessment.

---

[567] Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products.

[568] European Commission, *Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)* (28 September 2022).

[569] These changes are part of a broader revision of the Product Liability Directive: European Commission, *Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products* (2022).

[570] European Data Protection Board and European Data Protection Supervisor., *Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (2021)* https://edpb.europa.eu/our-work-tools/ourdocuments/edpbedps-joint-opinion; Fanny Hidvegi, Daniel Leufer, and Estelle Massé, "The EU Should Regulate AI on the Basis of Rights, Not Risks," *Access Now*, 17 February, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach/; Vera Lúcia Raposo, "The European Draft Regulation on Artificial Intelligence: Houston, We Have a Problem," in *Progress in Artificial Intelligence*, ed. Goreti Marreiros et al. (Cham: Springer International Publishing, 2022), 70.

[571] Chapter 3a AIA (EP).

6.4.2 Bias and Discrimination in the AI Act

As noted, the risk that AI systems may cause discrimination is a salient concern in the preparatory works to the European Commission's AI Act proposal.[572] The preparatory works frame biased AI primarily as a 'fundamental rights risk,'[573] but do not clearly delineate which fundamental rights biased AI systems might interfere with. Rather, they broadly assert that the use of AI can have a negative effect on the fundamental rights protected by the EU Charter.[574]

The AIA's orientation towards non-discrimination is further emphasised in the preamble recitals. This is indeed evident in the Commission's proposal, yet the emphasis on non-discrimination is even more pronounced in the recitals as amended in the Parliament's Compromise Text. These recitals stress that data governance measures should aim to mitigate biases that might lead to discriminatory outcomes.[575] Furthermore, they highlight the possibility of discrimination in relation to certain AI applications, including healthcare services, in addition to applications intended for areas such as social scoring, education, employment, credit assessments, and migration.[576] The possibility of algorithmic discrimination in these AI applications indicate that these applications are 'high-risk' within the context of the AIA.

The AIA is the first EU law to specifically address 'bias' in AI systems. However, the AIA does not formally define the term 'bias.' The meaning of this term within the provisions where it is used, is therefore open for interpretation. This meaning may not necessarily align with the foundational definition of 'bias' that this thesis relies on, which builds on the ISO's definition, as presented in Part II. Nevertheless, in the absence of any authoritative sources that could shed light on the intended meaning of 'bias' within the AIA, the definition

---

[572] AIA (EC), Explanatory Memorandum, 4; European Commission, *Impact Assessment Annexes Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (Brussels, 2021), 37.

[573] AIA (EC), Explanatory Memorandum, 2 and 11.

[574] AIA (EC), Explanatory Memorandum, 11. In one regard, however, the issue of biased AI is specifically equated with the risk of discrimination, namely in relation to biometric identification systems: Recitals 33 and 44 AIA (EC).

[575] Recital 44 AIA (EP).

[576] Recitals 17, 35, 36, 37 and 39 AIA (EP).

provided in Part II of this thesis could serve as a useful point of reference for the interpretation. This definition, which sees bias as a systematic difference in treatment (including perception, prediction, etc.) of certain individuals or groups in comparison to others, arguably represents a natural, intuitive understanding of the term 'bias.' However, this definition of 'bias' is very broad and does not clarify which biases the AIA is concerned with. For instance, as noted in chapter 3, the foundational definition of bias encompasses 'bias' as a problem potentially related to equality and non-discrimination, among several other aspects of 'bias' – it could also refer to 'bias' as a purely technical issue tied to the functioning of an AI model, with potential health and safety impacts resulting from suboptimal performance.

One notable aspect of the European Commission's AI Act proposal is its frequent reference to the risk of 'discrimination' within the preamble while refusing to refer to 'discrimination' within the suggested operational provisions. In contrast, the operational provisions proposed by the Commission appear to focus on the concept of 'bias' and measures to detect and mitigate biases. The Commission's proposal does not define if or to what degree these 'bias' provisions are designed to prevent discrimination or to achieve other legislative objectives. However, the connection between bias and discrimination within the AIA is more explicit in the Parliament's Compromise Text, which refers to 'discrimination' in Article 10, concerning requirements for training data.[577] Recital 44 of the Compromise Text, which also relates to training data requirements, clearly situates bias as a potential source of discrimination. It refers to "discrimination that might result from the bias in AI systems".[578] Consequently, in the context of training data requirements in the AIA, bias is primarily framed as problematic due to the risk of algorithmic discrimination.

Given the significance of context in the interpretation of EU law, it is plausible that the term 'bias' could have varying definitions across different provisions within the AIA. Particularly, given the AIA's objectives, the issue of bias may be addressed due to its potential impacts on health and safety. For instance, in Article 14(4)(b) and Article 15(3) AIA, 'bias' is employed with two distinct connotations, primarily associated with issues other than algorithmic discrimination. Article 14(4)(b) AIA introduces the term 'automation bias,' defined as the

---

[577] Article 10(2)(f) and 10(3) AIA (EP). See also Article 4(1)(a)e AIA (EP), mentioning non-discrimination as one of the fundamental principles proposed for foundational AI models.

[578] Recital 44 AIA (EP).

"tendency of automatically relying or over-relying on the output produced by a high-risk AI system." This provision mandates human oversight measures to ensure deployers remain aware of this issue. It does not specifically address non-discrimination, instead focusing on the broader implications of flawed decision-making due to over-reliance on automation.

The term 'bias' also appears in Article 15(3), which applies to systems that "continue to learn after being placed on the market or put into service." For such systems, measures are required to mitigate "possibly biased outputs due to outputs used as an input for future operations ('feedback loops')." Article 15 is concerned with "accuracy, robustness and cybersecurity."[579] It addresses bias as a source of systematic errors, without reference to discrimination. In practice, however, feedback loops may cause or reinforce discrimination in AI-CDS systems, which implies that the mitigation of feedback loops could contribute to non-discrimination even if this objective is not highlighted in Article 15.

In addition to provisions and recitals that explicitly address issues of bias and/or discrimination, these issues are also tackled by provisions requiring consideration of fundamental rights more generally. There is a consistent focus on the protection of fundamental rights throughout the AIA, including various measures that AI providers and/or deployers are obliged to take to ensure compliance with fundamental rights before the deployment of an AI system. These requirements are further discussed in chapter 7.

6.5 The Medical Device Regulation (MDR)

6.5.1 Overview

AI-CDS systems are considered medical devices under the definition provided by the MDR.[580] Therefore, the MDR's requirements are applicable to these systems. While AI systems can be integrated into physical medical devices, an AI-CDS system is usually a standalone medical device, not incorporated into another medical device.

Medical devices are regulated at the level of EU law, with the important implication that medical devices not complying with EU law cannot lawfully be placed on the EU market or put into service in EU Member States. The MDR is part of the New Legislative

[579] Article 15 AIA (EC).

[580] Article 2(1) MDR; Hauglid and Mahler (2023) 9-10.

Framework.[581] As such, it is one of several acts of EU law stipulating detailed product safety requirements that product manufacturers and their products must comply with.

The MDR was proposed in 2012.[582] It is fair to say that the extent to which it anticipates AI medical devices is limited. The regulation explicitly applies to software systems for prediction, monitoring and diagnosis,[583] but specific characteristics of AI systems are not reflected in the safety and performance requirements prescribed for such software systems. For instance, the MDR is not concerned with the properties of the dataset used to train an ML algorithm.

The MDR imposes several general obligations on manufacturers, including the establishment and maintenance of a risk management system (including a post-market surveillance system), conducting clinical evaluations of their devices, preparing technical documentation, registering devices in a digital database, implementing a quality management system (if applicable), and establishing a system for recording and reporting incidents and corrective actions.[584] The role as 'manufacturer' under the MDR is held by "a natural or legal person who manufactures or fully refurbishes a device or has a device designed, manufactured or fully refurbished, and markets that device under its name or trademark."[585] In relation to AI systems, the role as manufacturer essentially corresponds to that of a 'provider' under the AIA.[586]

To ensure compliance with these requirements and the general safety and performance requirements set out in Annex I MDR, a conformity assessment scheme is established.[587] This scheme entails that certain measures and assessments must be conducted before a medical device can be placed on the market or deployed. The exact procedures depend on the risk classification of the device. There are four risk classes under the MDR (class I, class IIa, class

---

[581] Section 6.1.

[582] *Proposal for a Regulation of the European Parliament and of the Council on Medical Devices, and Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009*, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52012PC0542.

[583] Article 2(1) MDR.

[584] Article 10 MDR.

[585] Article 2(30) MDR.

[586] Section 6.4.1.

[587] Article 52 MDR.

IIb, and class III). As a main rule, conformity assessment of devices in classes higher than class I must involve a 'notified body,' i.e., a third-party designated to carry out conformity assessments in accordance with Chapter IV MDR.[588] AI-CDS systems generally belong in the higher classes, according to the MDR's risk classification rules.[589]

6.5.2 The MDR and Non-Discrimination

The MDR is an important part of the overall regulatory framework for AI-CDS systems. Its objectives are to ensure quality (performance) and a high level of protection from the health and safety hazards associated with medical devices, while also promoting innovation and trade.[590] 'Safety' in the MDR is closely connected with the performance of medical devices.[591] 'Performance' is defined as "the ability of a device to achieve its intended purpose as stated by the manufacturer."[592] In a medical setting, when a device fails to fulfil the purpose it is intended for, the failure can easily pose a health and/or safety risk to persons who rely on the device. While the MDR addresses these risks related to the performance of medical devices, including AI-CDS systems, it is not oriented towards equality-related biases or discrimination.

Once the AIA is enacted and enters into force, the MDR must be read in conjunction with the AIA to fully understand the applicable conformity assessment scheme for AI-CDS systems. This includes the pre-deployment discrimination assessment requirements applicable to these systems. However, given that the MDR does not refer to the non-discrimination principle, it also lacks any specific pre-deployment discrimination assessment requirement. Consequently, chapter 7 focuses primarily on the AIA requirements, rather than the MDR. The AIA

---

[588] Recital 42 MDR, cf. Chapter IV MDR.

[589] Rule 11, in Annex VIII MDR, was created to address risks posed by standalone software devices. More specifically, it addresses system posing a risk because they are intended to provide information that will be relied on in clinical decision-making: Hauglid and Mahler (2023) 16. This rule places software systems intended for decision-support or crucial patient monitoring tasks in class IIa or higher.

[590] Recital 2 MDR.

[591] Anastasiya Kiseleva, "AI as a Medical Device: Is It Enough to Ensure Performance Transparency and Accountability in Healthcare?," *European Pharmaceutical Law Review*, no. 1 (2020): 13, https://doi.org/DOI: 10.21 552/epIr/2020/1/4.

[592] Article 2(22) MDR.

requirements broaden the scope of mandatory pre-deployment assessments beyond mere health and safety considerations, as specified in the MDR, to encompass fundamental rights, including non-discrimination.

6.6 The General Data Protection Regulation (GDPR)

6.6.1 Overview

Development and deployment of AI-CDS systems involve the processing of health data, often in large quantities. This processing is governed by the General Data Protection Regulation (GDPR). The GDPR entered into force in May 2016 and was applicable on 25 May 2018, repealing the Data Protection Directive (Directive 95/46/EC).[593] Establishing fundamental principles and specific requirements for the processing of personal data, the GDPR applies only to the processing of 'personal data' and does not apply to the processing of anonymous data.

Personal data is defined as "any information relating to an identified or identifiable natural person."[594] The contentious issue of distinguishing between anonymous data and personal data under the GDPR is not discussed in this thesis.[595] Rather, this thesis assumes that AI-CDS systems are trained using personal data and that personal data is also utilised as input data when these systems are used in practice. Furthermore, it is acknowledged that the outputs generated by AI-CDS systems may constitute personal data. However, questions concerning the legal basis for processing of personal data are not specifically addressed, as these questions lie beyond the objective of the thesis.

The primary focus of obligations set forth by the GDPR is on the individuals or entities defined as 'controller' and 'processor' of personal data. The 'controller' is the entity (or entities in the case of joint controllership) that "determines the purposes and means of the processing of personal data."[596] On the other hand, the 'processor' is an entity that processes personal

---

[593] See the definitions of "personal data" and "processing" in Article 4 GDPR.

[594] Article 4(1) GDPR.

[595] Hauglid and Mahler (2023) 426-27; Nadezhda Purtova, "The Law of Everything. Broad Concept of Personal Data and Future of Eu Data Protection Law," *Law, Innovation and Technology* 10, no. 1 (2018), https://doi.org/10.1080/17579961.2018.1452176.

[596] Article 4(7) GDPR.

data on behalf of the controller.[597] During the development of AI systems, developers act as controllers of personal data used for training purposes. However, when an AI system is deployed by another entity, this entity becomes the controller of the processing activities required when operating the system in practice. These activities include the use of patient data as input data. They might also include any data used for testing and implementation purposes, depending on the exact responsibilities agreed between the developer or provider and the deployer. For instance, a healthcare institution deploying an AI-CDS system could assume the role of the controller for personal data throughout deployment and use of the system in practice.[598] Furthermore, if training data continues to be utilised after deployment of the system, the healthcare institution may also be regarded as the controller when such data is processed, for instance in connection with continuous learning activities or updates of a deployed model.

A controller can only process personal data in accordance with a lawful basis for processing pursuant to Article 6 GDPR. When it comes to health data,[599] Article 9 GDPR imposes additional requirements regarding the lawful processing grounds.[600] The necessity of a specific legal basis for processing personal data stems from the fundamental tenet that personal data belongs to the individual person to whom it pertains, even if controllers and processors are allowed to store and process the data.[601] The GDPR refers to these persons as 'data subjects' and grants them a set of specific rights concerning the processing of their personal data. These rights include the right to information, the right to rectification and erasure, the right to restriction of processing, the right to data portability, and the right to

---

[597] Article 4(8) GDPR.

[598] Depending on the data infrastructure relied on in each case, it is possible that the developer, or a third-party distributing the AI system, is a 'processor' during operation of the system.

[599] i.e., "personal data related to the physical or mental health of a natural person, including the provision of health care services, which reveal information about his or her health status": Article 4(15) GDPR.

[600] The development of an AI-CDS system based on health data is typically based on processing grounds related to scientific research purposes, which includes development: Article 9(j) GDPR, cf. Recital 159 and Article 89 GDPR. The use of an AI-CDS system in practice would rely on Article 9(h) GDPR, concerning the provision of healthcare services.

[601] However, the traditional legal concepts of ownership, title or property are not directly regulated for personal data in the GDPR.

object to certain processing activities.[602] In the context of AI-CDS systems, the relevant data subjects are primarily the patients whose data is used for development purposes and the patients whose data is used as input data when running the system in practice.[603]

While the GDPR does not specifically mention AI technologies, it does contain a provision that is applicable to decision-making where AI is involved. Article 22(1) GDPR states that individuals "shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." Regardless, decisions based solely on automated processing are lawful according to Article 22(2) GDPR, if authorised by Union or Member State law, or if they are based on the data subject's explicit consent. The interpretation of Article 22(1) and (2) has stirred up considerable academic debate. Particularly, there is debate over whether Article 22 should be interpreted as a right for data subjects to object to automated processing or as a prohibition of automated processing.[604] Currently, the prevailing

[602] Chapter 3 GDPR.

[603] In addition, depending on the functions of an AI-CDS system, data concerning other patients may be used to generate contextual information that increases the interpretability of outputs and enables a clinician to review the outputs.

[604] Isak Mendoza and Lee A Bygrave, "The Right Not to Be Subject to Automated Decisions Based on Profiling," in *Eu Internet Law: Regulation and Enforcement*, ed. Tatiana-Eleni Synodinou et al. (Cham, Switzerland: Springer, 2017); Damian Clifford and Jef Ausloos, "Data Protection and the Role of Fairness," *Yearbook of European Law* 37 (2018); Mariam Hawath, "Regulating Automated Decision-Making: An Analysis of Control over Processing and Additional Safeguards in Article 22 of the Gdpr," *European Data Protection Law Review* 7, no. 2 (2021): 164, https://doi.org/10.21552/edpl/2021/2/6; Luca Tosoni, "The Right to Object to Automated Individual Decisions: Resolving the Ambiguity of Article 22 (1) of the General Data Protection Regulation," *International Data Privacy Law* 11, no. 2 (2021), https://doi.org/10.1093/idpl/ipaa024; Peter Davis and Sebastian Felix Schwemer, "Rethinking Decisions under Article 22 of the Gdpr: Implications for Semi-Automated Legal Decision-Making" (paper presented at the Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023), held in conjunction with ICAIL, 2023).

view among legal scholars is that Article 22(1) implies a "qualified prohibition," which means that there is a prohibition which is lifted when the exceptions in Article 22(2) apply.[605]

This thesis does not delve into the interpretation of Article 22 GDPR. Even if Article 22 GDPR, depending on its interpretation, could be applicable and impose a qualified prohibition on certain AI-CDS systems, these systems can still be used with patient consent or in accordance with national laws that permit automated decision-making (now or in the future). Moreover, AI-CDS systems will often be used in a manner that does not constitute decision-making solely based on automated processing.[606]

In the context of this thesis's objective, the GDPR is particularly relevant due to its emphasis on preventive compliance mechanisms. The GDPR obliges controllers to undertake certain measures to ensure compliance before initiating the processing of personal data. The requirement of a Data Protection Impact Assessment (DPIA) in Article 35 is the most prominent example of such a preventive compliance mechanism. Because this requirement potentially entails a venue for pre-deployment discrimination assessment, it is further explored in chapter 7.

6.6.2 The GDPR and Non-Discrimination

Some argue that both data protection law and non-discrimination law aim to preserve personal autonomy.[607] However, while personal autonomy, particularly in terms of 'informational self-determination,'[608] is widely regarded as fundamental to data protection law, there is less

---

[605] Hawath (2021) 164; The Working Party on the Protection of Individuals With REgard to the Processing of Personal Data, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (Wp251rev.01)* (6 February 2018), 19, https://ec.europa.eu/newsroom/article29/items/612053. Tosoni, who asserts a different interpretation, also notes that the qualified prohibition argument represents the "majority view": Tosoni (2021) 161.

[606] The provision only applies to fully automated decision-making, defined as decision-making that relies "solely" on automated data processing. According to EDPB guidance, this means that there is no meaningful human involvement in the decision-making process: Data (2018): 20-21.

[607] Khaitan (2015) 10; Maximilian Von Grafenstein, "The Principle of Purpose Limitation in Data Protection Laws" (PhD thesis, Hamburg University), 393.

[608] This refers to the idea that individuals should have control over their personal data, how it is collected, used, and shared. Informational self-determination is a popular conception of privacy,

consensus on its role in non-discrimination law. Equality, rather than personal autonomy, is seen as the primary underlying value and objective of non-discrimination law.[609] Nevertheless, the concept of equality can be defined in various ways, and it can be argued that personal autonomy, understood as the freedom to make choices for oneself without unjustified limitations, is encompassed within the value of equality. Moreover, 'dignity' is often seen as foundational to non-discrimination law, and personal autonomy is sometimes recognised as one element of this broader value.[610] Additionally, it can be argued that data protection law and non-discrimination law are interconnected because non-discrimination is one of the several values upon which data protection law and, specifically, the GDPR, is based.[611]

One fundamental distinction between EU non-discrimination law and EU data protection law is the respective frameworks' attention to *procedural* sides of decision-making versus the *outcomes* of decision-making. This difference is reflected by the scope of application of the prohibition on discrimination in EU law, compared to the scope of application of the GDPR. The non-discrimination principle applies to any action, process, decision, practice, etc., however conducted, and regardless of the technology used, if the *outcome* is defined as 'discrimination.' In contrast, the GDPR only applies to the processing of personal data and

---

often attributed to Alan Westin: e.g., Daniel J. Solove, *Understanding Privacy* (Cambridge, Massachussetts: Harvard University Press, 2008), 24.

[609] Section 4.2.

[610] Solanke refers to Kant's equation of dignity with autonomy: Solanke (2017) 51.

[611] Goodman (2016) 495; Alessandro Mantelero, "Beyond Data," in *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI* (Springer, 2022), 21.

gives less attention to the outcomes of the processing.[612] Hence, the GDPR does not explicitly prohibit discrimination.[613]

However, it is arguable that the GDPR is more concerned with discrimination and outcomes of data processing activities than was its predecessor, the 1995 Data Protection Directive.[614] Particularly, the GDPR's Preamble connects the principle(s) of fair and transparent processing ('fairness principle') in Article 5 GDPR with the utilisation of technical and organisational measures capable of preventing discriminatory *effects*.[615] This has led certain data protection authorities to suggest that the fairness principle indeed encompasses a requirement of non-discriminatory outcomes of data processing.[616] In contrast, the prevailing view in academic

---

[612] Lee A Bygrave, "Minding the Machine V2. 0: The Eu General Data Protection Regulation and Automated Decision Making," in *Algorithmic Regulation*, ed. Karen Yeung and Martin Lodge (Oxford: Oxford University Press, 2019), 260; In relation to the 1995 DPD: Raphaël Gellert et al., "A Comparative Analysis of Anti-Discrimination and Data Protection Legislations," in *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, ed. Bart Custers et al. (Heidelberg: Springer, 2013), 71.

[613] In the words of Veale and Edwards, the GDPR is "quiet, although not silent, on bias and discrimination within algorithmic systems": Michael Veale and Lilian Edwards, "Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling," *Computer Law & Security Review* 34, no. 2 (2018): 403, https://doi.org/https://doi.org/10.1016/j.clsr.2017.12.002.

[614] 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 (Data Protection Directive/DPD).

[615] Recital 71 GDPR. Because Recital 71 refers to preventing discriminatory outcomes related to protected characteristics such as ethnic origin, etc., it clearly refers to the non-discrimination principle.

[616] Information Commissioner's Office, *Big Data, Artificial Intelligence, Machine Learning and Data Protection* (2017), 19; Moreover, building on Recital 71, there is an opinion from the Article 29 Working Party (now replaced by the EDPB) on profiling and automated decision-making, that recommends several measures aimed at preventing discrimination in relation to processing covered by Article 22 GDPR. The EDPB also notes that "fairness is an overarching principle which requires that personal data should not be processed in a way that is unjustifiably detrimental, unlawfully discriminatory, unexpected or misleading to the data subject" (without pointing specifically towards the *outcomes* of processing): European Data Protection Board (EDPB), *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default* (20 October 2019), 17-18,

literature is that the fairness principle in EU data protection law is primarily oriented towards ensuring procedural fairness, implying that there must be a certain balancing of interests between data subjects and controllers.[617] For example, Veale and Edwards assert that the fairness principle in EU data protection law "has never been substantially attached to non-discrimination in processing outcomes."[618] Whether the expansive interpretation applied by certain supervisory authorities is supported by the CJEU remains to be seen, and it is not clear what the practical consequences would be. One conceivable consequence of such an interpretation is that the assessment of the risk of algorithmic discrimination would be mandatory under Article 35 GDPR. While this is arguably not the case, currently, section 7.5.1 will note that the DPIA could nonetheless be a valuable venue for voluntary discrimination assessments.

In addition to the fairness principle, it is worth mentioning the restrictions on the processing of special categories of personal data in Article 9 GDPR. Article 9(1) GDPR lists a set of characteristics, the processing of which is prohibited as a starting point. The list is not entirely overlapping with protected characteristics in EU non-discrimination law.[619] Given the

---

https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf. Inspired by the EDPB guidance, the Norwegian Data Protection Authority (*Datatilsynet*) chose to focus on algorithmic discrimination in a regulatory sandbox rooted in the GDPR's fairness principle: Datatilsynet, *Ahus, Sluttrapport: Hjerterom for Etisk AI* (Februar 2023), https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/ahus-sluttrapport-ekg-ai/.

[617] Lee Andrew Bygrave, *Data Privacy Law: An International Perspective* (Oxford: Oxford University Press, 2014), 146; Raphaël Gellert, *The Risk-Based Approach to Data Protection* (Oxford University Press, 2020), 67; Clifford and Ausloos (2018) 179; Veale and Edwards (2018) 403; Gianclaudio Malgieri, "The Concept of Fairness in the Gdpr: A Linguistic and Contextual Interpretation" (paper presented at the Proceedings of the 2020 Conference on fairness, accountability, and transparency, 2020), 159; Kazirdaheh and Clifford also note that "it is clear that the fairness principle is primarily concerned with mitigating the negative impacts of the power and information asymmetries between the controller (and processor) and data subject": Atoosa Kasirzadeh and Damian Clifford, "Fairness and Data Protection Impact Assessments" (paper presented at the Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021), 148.

[618] Veale and Edwards (2018) 403.

[619] Some categories mentioned in Article 9(1) are not mentioned in Article 21(1) of the Charter (trade union membership, biometric data, and data concerning health except for disabilities). On the

particular attention that this thesis gives to discrimination based on sex and ethnicity, it is worth noting that the GDPR does not treat information revealing a person's sex as a special category of personal data.[620]

The categories of personal data that are mentioned in Article 9(1) GDPR can only be processed on the basis of the specific legal grounds mentioned in Article 9(2). A similar provision existed in the 1995 Data Protection Directive. Restrictive rules for certain data types in data protection law has historically been motivated by the assumption that those data are of a nature that increases the risk of discrimination when they are processed.[621] This idea also underpins Article 9 GDPR. Goodman describes the removal of special categories of personal data from datasets as the GDPR's "primary principle for addressing algorithmic discrimination."[622] However, the main thrust of Article 9 GDPR is that it sets out specific requirements pertaining to the legal basis for processing special category data. The provision governs the lawful collection and use of special category data, rather than the outcomes of processing activities where such data is used.

---

other hand, some grounds listed in Article 21(1) of the Charter are not mentioned in Article 9(1) GDPR: Marvin van Bekkum and Frederik Zuiderveen Borgesius, "Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the Gdpr Need a New Exception?," *Computer Law & Security Review* 48, no. 105770 (2023), https://doi.org/10.1016/j.clsr.2022.105770.

[620] The strict requirements applicable to special categories data would be challenging to apply to data that reveals a person's sex or gender. That would have turned basic information such as a person's first name into special category data.

[621] Committee of Ministers, *Resolution (73) 22 on the Protection of the Privacy of Individuals Vis-a-Vis Electronic Data Banks in the Private Sector*, Council of Europe (26 September 1973), Article 1, https://rm.coe.int/1680502830; Paul De Hert and Serge Gutwirth, "Privacy, Data Protection and Law Enforcement. Opacity of the Individual and Transparency of Power," in *Privacy and the Criminal Law*, ed. Erik Claes, Antony Duff, and Serge Gutwirth (Antwerpen: Intersentia, 2006), 78-80; van Bekkum and Borgesius (2023) 5.

[622] Goodman (2016) 495 and 502. Goodman finds, however, that such removal is not likely to prevent discrimination and that it could lead to loss of information needed to detect discrimination.

7 Pre-Deployment Discrimination Assessment Requirements

7.1 Introduction

The principal purpose of this chapter is to identify provisions within the legal framework applicable to AI-CDS systems that can be interpreted as necessitating a pre-deployment discrimination assessment. Hence, it clarifies the specific legal contexts in which the methodological elements developed in this thesis might be needed to fulfil the legal obligations of providers and deployers of AI-CDS systems. Relatedly, this chapter highlights the utility of these methodological elements also outside of the cases where a discrimination assessment is mandated by law.

The provisions identified in this chapter prescribe various types of assessments, which may be conducted based on various assessment methodologies, such as risk assessment methodologies and impact assessment methodologies. As noted in section 1.1, the methodological elements developed in this thesis may be useful regardless of the underlying assessment methodology that one applies. However, they may need some further development before they can be readily integrated into a broader assessment methodology and applied in practice. The aim of the thesis is to develop methodological elements based on the non-discrimination principle in EU law, which are relevant to the parts of a pre-deployment assessment that encompasses discrimination. Consequently, this chapter searches for any type of assessment of an AI-CDS system that is required before its deployment and which presupposes an application of the non-discrimination principle in EU law. This means, for example, that assessment requirements that refer to 'discrimination,' 'discriminatory effects,' 'fundamental rights,' or similar concepts, are relevant to consider in this chapter.

Furthermore, this chapter reflects on what it means to 'assess discrimination in an AI-CDS system.' To assess discrimination means different things, depending on the underlying assessment methodology that one applies. For each pre-deployment discrimination assessment requirement that this chapter identifies, it therefore reflects on what it means to assess discrimination based on the underlying assessment methodology that the respective pre-deployment discrimination requirement refers to. For example, do the relevant types of discrimination assessment requirements imply that one needs to assess *whether discrimination is present* in an AI-CDS system or not? Is it a matter of degree, so that one

must assess whether an AI-CDS systems is *more or less 'discriminatory'*? Or, is the assessment oriented towards some other benchmark?

The pre-deployment discrimination assessment requirements identified in this chapter are often found within provisions that prescribe a broader set of measures aimed at ensuring compliance with applicable requirements before the deployment of an AI-CDS system. Such measures are referred to in this thesis as 'preventive compliance measures.' Thus, a requirement of a pre-deployment discrimination assessment may be one component of a broader preventive compliance measure. For example, risk management is relied on as a preventive compliance measure in the AIA.[623] Risk *assessment*, which might include a discrimination assessment, is one component of this broader preventive compliance measure. This chapter highlights the importance of preventive compliance measures, including pre-deployment assessments, to ensure non-discrimination in clinical decision-making where AI systems are involved.

Following a similar model of regulation to the one known from EU product safety legislation,[624] the AIA relies on an overarching preventive compliance measure referred to as 'conformity assessment.' This is a procedure through which the provider of an AI system demonstrates that it has abided by the requirements set out in the AIA. Such a demonstration of conformity presupposes that the provider has satisfyingly carried out the specific preventive compliance measures prescribed by the AIA provisions applicable to providers of high-risk AI systems, including any pre-deployment discrimination assessments required therein. Consequently, the methodological elements developed in this thesis can be utilised by AI providers when demonstrating conformity, which is the essential step towards certification and access to the EU market. Given its importance as an overarching preventive compliance measure, the conformity assessment scheme for AI-CDS systems is outlined in section 7.3.4.

This chapter proceeds, first, by establishing the importance of preventive compliance measures aiming to ensure non-discrimination in the context of AI-CDS systems. This is done by describing the traditional enforcement regime facilitated by EU non-discrimination law and discussing its limitations, in section 7.2. Section 7.3 then situates the use of preventive

---

[623] Section 7.4.1.

[624] Section 6.4.1.

compliance measures within the context of different regulatory strategies, as described in the literature on regulation theory. A basic understanding of the strategies underpinning the AIA's preventive compliance measures is helpful to the understanding of the specific pre-deployment assessment requirements that are found within the AIA.

Section 7.4 explores the AIA's requirements that specifically require an assessment falling within the purview of what this thesis defines as a 'pre-deployment discrimination assessment.'[625] Section 7.5 considers other relevant pre-deployment assessment provisions which either encourage or require a discrimination assessment. Section 7.6 then discusses what the identified pre-deployment discrimination assessment requirements entail for those tasked with conducting these assessments. To clarify the legal basis on which the methodological elements for such an assessment should be developed, it first asks whether the relevant requirements refer to the Equality Directives or Article 21 of the EU Charter. Finally, implications of the identified discrimination assessment requirements in terms of what it means to assess discrimination in an AI-CDS system are discussed.

7.2 EU Non-Discrimination Law's Traditional *ex post* Enforcement Regime

To underscore the importance of preventive compliance measures aimed at ensuring non-discrimination, it is worth considering the existing enforcement regime facilitated by EU non-discrimination law. Although the exact content and structure of enforcement mechanisms can be shaped by the Member States, the Equality Directives lay down five mandatory components of an enforcement regime.[626]

First, member states must facilitate the engagement of associations, organisations, or other legal entities in the procedures that the member states provide for the enforcement of the non-

[625] i.e., any assessment that must be conducted before the deployment of an AI-CDS system, which encompasses aspects of 'discrimination' as defined in EU law.

[626] In addition to the five components mentioned in this section, the rules concerning reversed burden of proof can also be seen as one of EU non-discrimination law's measures to promote effective enforcement, although these rules relate specifically to the evidentiary aspects of litigation: Jan Niessen, "Making the Law Work-the Enforcement and Implementation of Anti-Dicrimination Legislation," *European Journal of Migration & Law* 5, no. 2 (2003): 250. Because the reversed burden of proof applies to ex post litigation, it is not further discussed within this thesis.

discrimination principle.[627] Second, member states are required to establish a sanctions regime, the content of which is not specified, but which must be "effective, proportionate and dissuasive."[628] Sanctions that are encouraged by the Directives include monetary penalties and/or payment of compensation to victims of discrimination.[629] For instance, the GSED requires "real and effective compensation or reparation" through measures which shall be "dissuasive and proportionate to the damage suffered."[630] Third, the Equality Directives impose a general information obligation on the member states, requiring them to bring the provisions of the directives to the attention of persons concerned.[631] Fourth, Member States must address the issue of 'victimization': Individuals should be protected from adverse reactions following an attempt to pursue a discrimination claim.[632]

Fifth, the Equality Directives mandate the establishment and funding of specialised bodies for the promotion and monitoring of discrimination in the Member States. Currently, EU law does not specify the functions that equality bodies may have or the mechanisms that they may use to ensure compliance. Their effectiveness in doing so has therefore been questioned.[633] To strengthen the EU-level requirements concerning equality bodies, two directives specifically addressing this issue were proposed by the European Commission in December 2022.[634] The

---

[627] Article 7(2) RED; Article 9(2) EED.

[628] Article 15 RED, cf. Recital 19 RED ("Persons who have been subject to discrimination based on racial and ethnic origin should have adequate means of legal protection"), cf. Recital 26 RED ("Member States should provide for effective, proportionate and dissuasive sanctions in case of breaches of the obligations under this Directive"); Article 17 EED; Article 25 Recast Directive; Article 14, cf. Recital 27 GSED.

[629] e.g., adequate compensation is emphasized in Recitals 33-35 of the Recast Directive.

[630] Article 8(2) GSED. See also Article 18 Recast Directive, requiring "real and effective compensation or reparation (…) in a way which is dissuasive and proportionate to the damage suffered."

[631] Article 10 RED; Article 12 EED; Article 15 GSED; Article 30 Recast Directive.

[632] Article 9 RED; Article 9 EED; Article 10 GSED; Article 24 Recast Directive.

[633] Sara Benedí Lahuerta, "Equality Bodies in the EU: Origins, Challenges and Future Prospects," in *Edward Elgar Research Handbook on European Anti-Discrimination Law (Forthcoming)* (Edward Elgar, 2021), 4-5.

[634] Proposal 7 December 2022 for a Council Directive on Standards for Equality Bodies in the Field of Equal Treatment Between Persons Irrespective of Their Racial or Ethnic Origin, Equal Treatment in the Field of Employment and Occupation Between Persons Irrespective of their Religion

proposals emphasise the need to equip equality bodies with the digital resources required to address the risk of algorithmic discrimination as well as using AI as a tool to identify patterns of discrimination.[635]

The existing enforcement requirements in EU non-discrimination law are strongly focused on ex post enforcement. The most preventively oriented component is the information obligation which rests on the Member States. However, as a means of ensuring compliance, the Member States' provision of information to regulated entities is a teethless measure. It does not require any preventive measures from AI developers or deployers. The fifth component – the establishment of specialised promotion and monitoring bodies – also has a preventive aspect, as these bodies are expected to engage in proactive efforts against discrimination. However, in countries where equality bodies engage in preventive measures, these measures tend to be limited to general research, development of good practices, and communication efforts.[636]

In summary, the enforcement regime facilitated by EU non-discrimination law relies heavily on ex-post enforcement.[637] The lack of preventively oriented measures in non-discrimination

---

or Belief, Disability, Age or Sexual Orientation, Equal Treatment Between Women and Men in Matters of Social Security and in the Access to and Supply of Goods and Services, and Deleting Article 13 of Directive 2000/43/EC and Article 12 of Directive 2004/113/EC (COM(2022) 689 final); Proposal 7 December 2022 for a Directive of the European Parliament and of the Council on Standards for Equality Bodies in the Field of Equal Treatment and Equal Opportunities Between Women and Men in the Matters of Employment and Occupation, and Deleting Article 20 of Directive 2006/54/EC and Article 11 of Directive 2010/41/EU (COM(2022) 688 final).

[635] COM(2022) 689, Recital 20.

[636] Niall Crowley, *Equality Bodies Making a Difference*, European network of legal experts in gender equality and non-discrimination (2018), 70-71.

[637] The provision in the current Equality Directives which goes the furthest in requiring preventive measures is Article 26 of the Recast Directive, which is entitled "prevention of discrimination." This provision says that member states shall encourage employers and those responsible for access to vocational training to "take effective measures to prevent all forms of discrimination on grounds of sex (…)." The provision thus appears to require something more than dissuasive ex-post enforcement mechanisms, but it does not set out any specific preventive regulatory mechanisms.

law (in the EU and elsewhere) has been criticised in academic literature.[638] Indeed, non-discrimination law would stand a better chance of realizing its objectives if it could prevent discrimination from occurring in the first place.[639]

The current regime has limited capabilities of ensuring non-discrimination in practice, for several reasons. One central issue with the ex post enforcement regime pertains to its reliance on individual initiative and litigation.[640] Victims of discrimination tend not to have access to the information they need to confirm their suspicions about discrimination, in which case it does matter whether they would be willing and able to pursue their claim in the first place.[641] In healthcare, specifically, patients are often in a dependent position, inclined to trust the clinical assessments being made. As the fictitious case of Simon Tesfay v UHS illustrates (see section 5.1), patients generally do not have access to information necessary to detect that discrimination has occurred. Pre-deployment discrimination assessment requirements for AI systems could arguably be a crucial addition to the existing enforcement regime.

---

[638] Catherine Barnard, "Gender Equality in the Eu: A Balance Sheet," in *The Eu and Human Rights*, ed. Philip Alston (Oxford: Oxford University Press, 1999), 258; O'Hare (2001) 154. ("None of these measures, however, encourage Member States to make provision for what Barnard terms pro-active remedies, such as a legally binding direction as to future conduct or a programme of affirmative action measures in the event of a finding of discrimination.")

[639] Niessen (2003) 255.

[640] Barnard (1999) 258; Mark Bell and Sara Kjellstrand, *Critical Review of Academic Literature Relating to the Eu Directives to Combat Discrimination*, European Commission Directorate-General for Employment and Social Affairs (2004), 28 and 30, https://www.antigone.gr/wp-content/uploads/library/documentation-of-EU-and-international-organizations/policy-documents/en/criticrevaclit.pdf; Niessen (2003) 254.

[641] Sandra Fredman, *Discrimination Law*, 2nd ed. (Oxford: Oxford University Press, 2011), 283. ("Direct discrimination is particularly difficult to prove, since most relevant evidence is in the hands of the respondent. Indirect discrimination and equal pay claims have their own difficulties, requiring complex compilation of statistics."); Henrard (2019) 95. ("Victims of discrimination tend to face tremendous difficulties in producing proof, since the required information is often only accessible to the perpetrator.")

7.3 Preventive Compliance Versus Ex Post Enforcement, 'Meta-Regulation'

7.3.1 Compliance and Enforcement

'Compliance' generally means to abide by the law, to fulfil one's duties, or to satisfy applicable requirements.[642] If discrimination occurs, this means that there is an incident of non-compliance with the non-discrimination principle. As stressed in section 6.4.2, a central aim of the AIA is to enhance compliance with the non-discrimination principle – in other words, ensuring non-discrimination. Pre-deployment discrimination assessments are one way of enhancing compliance, because they entail that the non-discrimination principle is considered when it is determined whether an AI system can be deployed.

Compliance is closely connected with the notion of 'enforcement.' The term 'enforcement' often refers to the act of invoking the consequences of non-compliance with a legal rule *after the fact* (i.e., after the rule has been violated).[643] Enforcement, in this sense, is referred to in this thesis as 'ex post enforcement.'[644] In a wider sense, enforcement can be understood as any measure aimed at ensuring compliance, regardless of whether these measures are applied preventively or after an allegation or suspicion of unlawful behaviour has arisen. Given this wide understanding of enforcement, preventive compliance measures within a legal framework can be seen as components of a larger system of 'enforcement.' Preventive compliance measures are often combined with ex post enforcement measures such as administrative sanctions and procedures for complaints and dispute resolution. Section 7.2 has already highlighted limitations associated with ex post enforcement in relation to non-discrimination law, specifically.

---

[642] Andrew Hopkins and Andrew Hale, "Issues in the Regulation of Safety: Setting the Scene," in *Changing Regulation: Controlling Risks in Society* (Pergamon-Elsevier Ltd, 2002), 7. ("A regulatory regime can choose to place its emphasis differently in both what it specifies as rules and what it looks at as evidence of compliance with the rules".)

[643] For example, Gellert defines enforcement narrowly: Gellert (2020) 51.

[644] In contrast to the definition of 'enforcement' relied on herein, Gellert considers enforcement to be "per definition *ex post*": Ibid.

7.3.2 Meta-Regulation

Preventive compliance measures, which may require different types of pre-deployment assessments, form parts of the regulatory strategy that the EU is relying on in its regulation of AI systems. To underscore the role and potential importance of these compliance measures, it is worth situating them in the light of existing theories on different regulatory strategies.

When determining how to regulate an area, i.e., how to best influence the behaviour of the relevant entities within this area,[645] various strategies may be taken.[646] Legal systems usually include ex post enforcement measures which enable a supervisory or judicial authority to force the fulfilment of an obligation, or to impose punishment or other sanctions on a rulebreaker (ex post enforcement). Ex post enforcement is particularly associated with the regulatory strategy referred to as 'command-and-control-regulation.'[647] Command-and-control-regulation tends not to intervene too much when it comes to *how* the regulated entities ensure compliance with applicable laws. In a pure (theoretical) command-and-control mode of regulation, it does not matter whether regulated entities comply with the law by sticking to carefully systematized internal procedures for ensuring compliance or as a matter of random luck. Legislators relying on this strategy typically set out the rules that regulated entities are expected to comply with and attach sanctions to the breach of those rules. Hence, the law seeks to ensure compliance through the threat of sanctions. Enforcement in the context of command-and-control regulation tends to be reactive rather than preventive.

When laws require that certain preventive measures must be taken to ensure compliance with applicable requirements or objectives, for instance by demanding that various assessments are conducted, this is not typical of command-and-control regulation. Rather, preventive measures can be viewed as part of a regulation strategy that responds to the typical limitations of command-and-control-regulation: 'meta-regulation.'[648] Meta-regulation is characterised by

---

[645] In accordance with Black's definition of 'regulation,' which emphasises the attempt to influence behaviour: Black (2001) 142.

[646] For an overview: Neil Gunningham, "Enforcement and Compliance Strategies," in *The Oxford Handbook of Regulation* (Oxford: Oxford University Press, 2010).

[647] Cary Coglianese and Evan Mendelson, "Meta-Regulation and Self-Regulation," in *The Oxford Handbook of Regulation* (Oxford: Oxford University Press), 146.

[648] Coglianese and Mendelson (2010).

the tendency to force regulated entities to conduct certain activities for the purpose of ensuring their own compliance, and to document or report the results of such activities.[649] Hence, meta-regulation involves an element of self-regulation.[650] However, meta-regulation typically refers to contexts where the self-regulatory efforts are mandated by law, in which case one may speak of 'enforced self-regulation.'[651] In a meta-regulation regime, the role of regulatory authorities is primarily to assess the internal compliance measures implemented by the regulated entities.[652] These compliance measures are often designed to produce information based on which the regulator may assess compliance, if it chooses to access the information.[653]

The pre-deployment discrimination assessment requirements discussed in this chapter can be viewed as part of a meta-regulatory approach to the regulation of AI-CDS systems. One reason why meta-regulation might – under the right circumstances – be effective at enhancing compliance, is because the regulatory activities are carried out by stakeholders that understand the regulated product or process well and has access to relevant information.[654] This is probably one of the reasons why mechanisms associated with meta-regulation are relied on in the AI Act.

The AI Act arguably relies on a combination of meta-regulation and command-and-control regulation. The focus on making AI providers and deployers responsible for conducting preventive compliance measures is typical of meta-regulation. At the same time, the articulation of specific requirements and threats of sanctions for non-compliance entail a

---

[649] Christine Parker, *The Open Corporation: Effective Self-Regulation and Democracy* (Cambridge University Press, 2002), 245-46; Gunningham (2010) 135.

[650] The term 'self-regulation' is used in various ways in regulation literature, but it typically refers to regulatory activities conducted without direct steering from governmental regulatory agencies: Black (2001) 113-14.

[651] John Braithwaite, "Enforced Self-Regulation: A New Strategy for Corporate Crime Control," *Michigan Law Review* 80, no. 7 (1982): 1470-71, https://doi.org/10.2307/1288556.

[652] Gunningham (2010) 135.

[653] The production of information of interest to the regulator is emphasised by Parker: Parker (2002) 245.

[654] Victoria I Daskalova and Michiel A Heldeweg, "Challenges for Responsible Certification in Institutional Context: The Case of Competition Law Enforcement in Markets with Certification," in *Certification–Trust, Accountability, Liability* (Springer, 2019), 39.

considerable element of command-and-control regulation.[655] Another command-and-control aspect of the AIA is the empowerment of supervisory authorities and notified bodies not only to assess an AI provider's compliance efforts, but also to conduct independent assessments of AI systems.[656]

The use of pre-deployment assessments, including discrimination assessments, can also be seen as part of an entity's 'internal control' measures. 'Internal control' is a term that broadly refers to an entity's efforts aimed at ensuring that its operations meet various expectations, not only pertaining to legal compliance but also other objectives such as ensuring a certain quality and safety of operations, economic and social governance, data governance, etc.[657] The methodological elements for discrimination assessment that this thesis develops could be operationalised as part of an internal control system within a healthcare institution or AI provider.

7.3.3 The EU's 'Risk-Based Approach' to the Regulation of AI

The notion of 'risk' plays a central role in the AIA. It is designed on the basis of a risk assessment, through which the EU legislature anticipates the risks posed by AI systems.[658] The anticipated risks posed by AI systems constitute the catalyst and the justification for the AI Act. The European Commission refers to its approach to the regulation of AI as a "risk-based regulatory approach" and a "proportionate risk-based approach."[659] This proportionate, risk-based regulatory approach entails that the scope and substance of the requirements set out in the AIA are designed to address AI-related risks in a proportionate manner, i.e., in a

---

[655] Gellert also notes that rather than a (risk-based) partially flexible and discretionary system of compliance, the proposed AIA seems to feature the logic of rules abidance that is characteristic of command and control regulation: R. M. Gellert, "The Role of the Risk-Based Approach in the General Data Protection Regulation and in the European Commission's Proposed Artificial Intelligence Act. Business as Usual ?," *Journal of Ethics and Legal Technologies* 3 (2021): 26.

[656] Annex VII AIA (EC), Section 4.4. ("Whenever the notified body is not satisfied with the tests carried out by the provider, the notified body shall directly carry out adequate tests, as appropriate.")

[657] e.g., Will Kenton, "Internal Controls: Definition, Types, and Importance," Julius Mansa and Suzanne Kvilhaug eds. *Investopedia*, 24 May, 2023, https://www.investopedia.com/terms/i/internalcontrols.asp; "Internkontroll," in *Store norske leksikon*. https://snl.no/internkontroll.

[658] European Commission (21 April 2021).

[659] AIA (EC), Explanatory Memorandum, 3.

manner that does not provide unjustified barriers to the conduct of business, innovation and trade.

The AIA's risk classification scheme serves as the primary tool in ensuring a proportionate, risk-based regulation.[660] According to this scheme, the majority of requirements in the AIA are only applicable to 'high-risk' AI systems. The risk classification scheme is designed to ensure that AI systems are classified as 'high-risk' if they "have a significant harmful impact on the health, safety and fundamental rights of persons in the Union."[661] As noted, AI-CDS systems will typically fall into the 'high-risk' category.[662] The classification of an AI system as 'high-risk' is not based on an individual risk assessment of a specific AI system, but rather by the AIA's general risk classification rules. However, the AIA mandates that individual risk assessments must still be undertaken by the provider. This is a part of the risk management system that the provider is obliged to establish as per Article 9 AIA.[663]

7.3.4 Conformity Assessment: The AIA's Overarching Compliance Measure

Before an AI system can be placed on the market or put into service in accordance with the AIA, the provider of the system is obligated to ensure that a 'conformity assessment' has been carried out.[664] To signify that the conformity assessment has been conducted, a CE (Conformité Européenne) mark shall be included in the AI system's accompanying documentation.[665] The conformity assessment is the AI Act's overarching regulatory measure to ensure that AI systems and their providers comply with the requirements stipulated for

---

[660] Tobias Mahler, "Between Risk Management and Proportionality: The Risk-Based Approach in the Eu's Artificial Intelligence Act Proposal," in *Law in the Era of Artificial Intelligence*, ed. Liane Colonna and Stanley Greenstein, Nordic Yearbook of Law and Informatics (Stockholm: The Swedish Law and Informatics Research Institute, 2022), 267-68.

[661] Recital 27 AIA (EC).

[662] Section 6.4.1.

[663] Article 9 AIA (EP).

[664] Article 16 AIA (EP).

[665] While the main rule is that a physical CE mark shall be visibly affixed on high-risk AI systems, AI-CDS systems are stand-alone software systems without a physical embodiment on which to place a mark. In such cases, Article 49 AIA (EP) foresees that the CE mark shall be included in the accompanying documentation or created as a digital CE mark.

high-risk AI systems.[666] Before the European Commission presented the AI Act proposal in April 2021, the idea of a pre-deployment certification scheme for AI systems had already gained momentum in the public discourse on regulatory strategies to ensure safety and fundamental rights protection.[667] As an example, the Committee of Ministers of the Council of Europe had recommended the adoption of certification schemes based on regional and international standards to ensure the origin and quality of datasets and models.[668]

Within the EU's existing product safety framework, conformity assessment is a well-established procedure. A conformity assessment is understood within this framework as a process that "demonstrates whether a product, service, process, claim, system or person" meets applicable requirements.[669] Due to the overarching role of conformity assessment as a compliance measure therein, the AIA is often characterised as adopting a product safety approach to AI regulation.[670]

In relation to some products within the existing EU product safety framework, the conformity assessment may be part of a pre-deployment *approval* scheme, requiring pre-deployment

---

[666] Jakob Mökander et al., "Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation," *Minds and Machines* 32, no. 2 (2022): 241, https://doi.org/10.1007/s11023-021-09577-4.

[667] A significant role for certification of AI systems is suggested by Mantelero: Alessandro Mantelero, *Report on Artificial Intelligence (Convention 108)*, Council of Europe Consultative Committee of the Convention for the Protection of Individuals with Regard to Auotmatic Processing of Personal Data (2019), 13-14, https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6. Furthermore, an overview of proposed certification schemes for AI is provided by Cihon et al: Peter Cihon et al., "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries," *IEEE Transactions on Technology and Society* 2, no. 4 (2021): 203.

[668] Committee of Ministers, *Recommendation of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems*, Council of Europe (8 April 2020), https://rm.coe.int/09000016809e1154.

[669] "Conformity Assessment," International Organization for Standardisation (ISO) (web page) accessed 12 November, 2023, https://www.iso.org/conformity-assessment.html; *Iso/Iec 17000:2020(En),* (International Organization for Standardisation (ISO)), section 4.1-4.2.

[670] e.g., Lilian Edwards, *Regulating AI in Europe: Four Problems and Four Solutions*, Ada Lovelace Institute (March 2022), 6, https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/; Hauglid and Mahler (2023) 5.

authorisation from a governing body.[671] However, most products under EU law, including medical devices, do not require such authorisation. The objective of the conformity assessment within the AIA (and MDR) is not to persuade an authority to authorise an AI system, but rather to assert to users, consumers, patients, and authorities that the product manufacturer has met the relevant provisions. In other words, the intention of AIA certification is to cultivate trust in AI systems,[672] as it signals to the public and authorities that the AI system complies with the AIA's stipulations for high-risk AI systems.[673]

While conformity assessment is a well-established compliance measure in EU product safety legislation, the AIA makes use of this compliance measure in an expanded manner compared to existing laws. This expansion is attributed to the fact that the AIA relies on conformity assessment to ensure compliance not only with safety requirements, but also with fundamental rights. In comparison, the conformity assessment under the MDR primarily ensures that medical devices are fit for their intended purposes, capable of achieving the manufacturer's stated performance level, and pose an acceptable risk to health and safety.[674]

Despite expanding the scope of conformity assessment to the area of fundamental rights, the AIA is not meant to impact "the specific logic, methodology or general structure of conformity assessment" under the MDR.[675] Instead, it anticipates a unified conformity assessment procedure where an AI-CDS system's provider shall demonstrate conformity with both the AIA and MDR requirements.[676] Conformity assessment procedures for AI-CDS

---

[671] Unlike in the US, EU law does not require pre-market approval of medical devices. In comparison, EU law requires pre-market approval for certain other product categories, such as medicinal products. When there is a pre-market approval scheme, a conformity assessment may be a step in the process towards approval.

[672] Recital 62 AIA (EP).

[673] Recital 67 AIA (EP).

[674] Annex I MDR, section 1.

[675] Recital 63 AIA (EP).

[676] Article 43(3) AIA (EP). This provision outlines certain additional steps compared to the procedures that already follow from the MDR. Moreover, to prepare the ground for this type of combined MDR/AIA conformity assessment, the AIA facilitates the authorization of notified bodies that are operating under the MDR so that the same notified bodies may also assess conformity to AIA requirements, in addition to the MDR requirements, provided that they fulfil the AIA's requirements for notified bodies.

systems will regularly involve a 'notified body.'[677] As noted in section 6.4.1, the involvement of a notified body according to the relevant MDR procedure has a specific legal ramification under the AIA: it triggers the classification of an AI-CDS system as 'high-risk' under the AIA.

As detailed in the following sections, the conformity assessment covers provisions obligating AI providers and, to some extent, deployers to conduct certain pre-deployment assessments, including discrimination assessments. Consequently, the methodological elements developed in this thesis could be utilised by providers, deployers and notified bodies in the context of conformity assessments. However, the conformity assessment is an overarching mechanism to ensure that the various compliance measures specified within the AIA are carried out by the regulated entities. The measures of principal interest are the ones that require a pre-deployment assessment, in one form or another, and encompasses aspects of discrimination. The following sections identify and explore such preventive compliance measures within the relevant legal framework.

7.4 Pre-Deployment Discrimination Assessment Requirements in the AI Act

7.4.1 The AIA's Risk Assessment Requirement

Article 9 AIA describes the required risk management system that providers of high-risk AI systems are obliged to establish as a "continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system."[678] The exact scope and formulation of the requirements pertaining to the risk management system are subject to ongoing legislative negotiations at the time of writing.[679] One may reasonably assume, based on the different negotiation versions, that the risk management system shall entail a process comprising the

[677] Section 6.5.1.

[678] Article 9(2) AIA (EP). This corresponds to the definition of risk management under the MDR: Annex I MDR, Chapter 1, Section 3.

[679] The risk management provisions in the negotiation version adopted by the European Parliament contains several changes compared to the initial proposal from the European Commission.

following four steps: risk identification, risk estimation, risk evaluation, and risk management measures aimed at addressing any identified risks.[680]

The aforementioned steps align closely with the risk management requirements that manufacturers of medical devices are subject to under the MDR.[681] This correlation is logical because the providers of AI medical devices should be enabled to integrate the AIA risk management requirements into their existing MDR-based risk management systems. Moreover, corresponding steps can be found in the International Standardisation Organisation's (ISO) risk management standard, ISO 31000. This standard uses the term 'risk assessment' to encompass the process of 'risk identification,' 'risk analysis' (which encompasses what the AIA refers to as 'risk estimation'), and 'risk evaluation.'[682] Consequently, 'risk assessment' refers to three of the four steps in the risk management system required under the AIA. In this thesis, the term 'risk assessment' is employed to refer to these three steps, in accordance with the terminology used in ISO 31000. It is worth noting that the ISO also adopts this terminology in a dedicated risk management standard for AI systems, based on ISO 31000.[683]

While the European Commission's AIA proposal does not specify the scope of the risk assessment requirement, the Compromise Text adopted by the European Parliament specifies that the risk assessment shall encompass risks to the health and safety of natural persons as well as risks to their fundamental rights.[684] The reference to fundamental rights suggests that the mandatory risk assessment must encompass aspects of discrimination. This requirement therefore constitutes a pre-deployment discrimination assessment requirement. It entails that the 'risk' of algorithmic discrimination must be assessed, thus aligning with the articulation of the AIA's non-discrimination objective in the Explanatory Memorandum accompanying the

---

[680] These steps are included in the Commission's proposal as well as in the version adopted by the Parliament.

[681] Annex I MDR, Chapter 1, Section 3.

[682] *ISO 31000:2018 (E),* (International Organization for Standardisation (ISO)), 11-12.

[683] *ISO/Iec 23894:2023(E),* (International Organization for Standardisation (ISO)).

[684] Article 9(1)(a) AIA (EP). The version adopted by the Parliament mentions "equal access and opportunities" among the fundamental rights to be considered in the assessment. While it is not clear exactly what this refers to, it seems reasonable to assume that it refers to the non-discrimination principle and its underpinning objectives, cf. section 4.2.

Commission's proposal, which states that the AIA aims "to minimise the risk of algorithmic discrimination."[685] Consequently, it represents a type of discrimination assessment according to which the deployment of an AI-CDS system is determined on the basis of the risk of discrimination posed by the system.

7.4.2 Data Bias Examination

Article 10(f) AIA establishes that training, validation and testing datasets (which are collectively referred to as 'training data' within this thesis) shall be subject to an "examination in view of possible biases."[686] While the initial AIA proposal from the European Commission does not further specify the scope of such an examination, the Parliament's position clarifies that there shall be *examination in view of possible biases that are likely to lead to discrimination prohibited under Union law*.[687] This provision entails an obligation to assess specific components of an AI system, to determine whether those components contain biases likely to lead to discrimination. This assessment constitutes a pre-deployment discrimination assessment.

The particular pre-deployment discrimination assessment required by Article 10(f) AIA is oriented towards the *probability* that discrimination might occur. It is reasonable to interpret this provision as indicating that biases hinder deployment of an AI-CDS system if they have a higher probability of resulting in discrimination rather than not. If biases are likely to lead to discrimination, this requirement implies that an AI system cannot be deployed unless the biases are mitigated to such an extent that they are no longer 'likely' to lead to discrimination.

7.4.3 Discrimination Impact Assessment?

While undecided, it is currently possible that the AI Act might include requirements amounting to a fundamental rights impact assessment, in addition to the risk assessment requirement in Article 9 AIA. In the preparatory works accompanying the European Commission's AIA proposal (a document which is itself called an 'impact assessment'), the requirement of a fundamental rights impact assessment is considered as a potential compliance measure.[688] In support of such a requirement, the preparatory works refer to

---

[685] AIA (EC), Explanatory Memorandum, 4.

[686] Article 10(2)(f) AIA (EC).

[687] Article 10(f) AIA (EP).

[688] European Commission (21 April 2021): 58-59.

recommendations from, e.g., the Council of Europe and the Fundamental Rights Agency.[689] The CoE recommends that AI systems should be subject to a human rights impact assessment "in order to identify the risks of rights-adverse outcomes."[690] The Fundamental Rights Agency particularly highlights the need for contextualised fundamental rights impact assessments of AI systems due to the differences between the areas where they may be applied as well as the contextual nature of fundamental rights.[691] A fundamental rights impact assessment for AI systems is also advocated for in the EU-HILEG AI Ethics Guidelines from 2019.[692] Moreover, fundamental rights impact assessment methodologies for AI systems have been suggested in academic literature.[693]

Despite several recommendations, a fundamental rights impact assessment requirement is not included in the European Commission's AI Act proposal, "because users of high-risk AI systems would normally be obliged to do a Data Protection Impact Assessment (DPIA) that already aims to protect a range of fundamental rights of natural persons and which could be interpreted broadly, so new regulatory obligation was considered unnecessary."[694] The GDPR's DPIA requirement is discussed in section 7.5.1 below. It is suggested therein that the assumption that the DPIA requirement adequately addresses a range of fundamental rights may be contested. Therefore, if the intention is to subject AI providers to a comprehensive fundamental rights impact assessment, an explicit requirement to that effect should be

---

[689] Ibid.

[690] Committee of Ministers (2020): 10.

[691] European Agency for Fundamental Rights (FRA), *Getting the Future Right - Artificial Intelligence and Fundamental Rights* (Luxembourg, 2020), 87 and 96, https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf.

[692] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission (Brussels, 8 April 2019), 15, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[693] Heleen L Janssen, "An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making," *International Data Privacy Law* 10, no. 1 (2020), https://doi.org/10.1093/idpl/ipz028; Alessandro Mantelero and Maria Samantha Esposito, "An Evidence-Based Methodology for Human Rights Impact Assessment (Hria) in the Development of AI Data-Intensive Systems," *Computer Law & Security Review* 41, no. 105561 (2021), https://doi.org/10.1016/j.clsr.2021.105561; Mantelero (2022).

[694] European Commission (21 April 2021): 59.

included in the AIA. Such a requirement is indeed included in the Compromise Text adopted by the European Parliament.[695]

The Compromise Text explicitly introduces a requirement for deployers of high-risk AI systems to conduct a fundamental rights impact assessment requirement prior to deployment.[696] This can be understood as requiring a pre-deployment discrimination assessment. However, upon examination, the proposed text lacks clarity in relation to the scope, content, and methodology of this impact assessment. It refers to various assessment types, requiring a "verification that the use of the system is compliant with relevant Union and national law on fundamental rights,"[697] an assessment of "the reasonably foreseeable impact on fundamental rights of putting the high-risk AI system into use,"[698] as well as "specific risks of harm likely to impact marginalised persons or vulnerable groups." If this provision becomes enacted, there is a need to clarify how discrimination in AI systems should be assessed in accordance with the provision.

Moreover, it is worth mentioning that although the Commission's AIA proposal does not specifically mention a fundamental rights impact assessment in the operational provisions, the technical documentation requirements in the proposal imply that a discrimination impact assessment should be conducted prior to deployment. The technical documentation requirements are specified in Annex IV AIA.[699] This annex requires, according to the Commission's proposal, that the technical documentation shall include "the metrics used to measure […] potentially discriminatory impacts."[700] Notably, this does not say anything about which metrics should be used to measure discriminatory impacts. The provision leaves stakeholders with considerable uncertainty regarding the appropriate process and criteria to apply when measuring 'discriminatory impacts.'[701] In the absence of further guidance, the provision might interpreted by AI providers as a reference to technical fairness metrics, as

---

[695] Article 29 a AIA (EP); Recital 58 a AIA (EP).

[696] Article 29 a AIA (EP); Recital 58 a AIA (EP).

[697] Article 29 a(1)(d) AIA (EP).

[698] Article 29 a(1)(e) AIA (EP).

[699] Article 11 AIA (EC).

[700] Annex IV AIA, Section 2(g) (EC).

[701] Bellamy et al. (2019) 77. ("[…] clarity on which bias metrics […] are most appropriate for different contexts is yet to be achieved.")

such metrics are often referred to as potential solutions to the issue of bias in computer science literature.[702] However, a discrimination impact assessment in the context of EU law must be understood as an assessment based on the non-discrimination principle in EU law, rather than technical fairness metrics. In this context, relying on fairness metrics can be a fallacy because those metrics are rarely based on the definition of discrimination as recognised by EU law.[703] The application of EU non-discrimination law often involves the consideration of qualitative factors and the delicate task of reconciling conflicting arguments and interests. These considerations are not easily translated into the quantitative measurements that are usually implied by fairness 'metrics.'

In summary, it remains possible that the AIA might end up including a discrimination impact assessment requirement. Such a requirement would constitute a pre-deployment discrimination assessment. The provisions that have been suggested at this point would be liable to create considerable uncertainty regarding the methodological approach to such an assessment. They provide very little guidance in terms of what it would mean to assess discrimination. Moreover, if the Parliament's suggestion is adopted, the fundamental rights impact assessment requirement would apply only to deployers. This is a challenging proposal – deployers would be in a better position to conduct a fundamental rights impact assessment if providers were also under the same obligation, because deployers need to rely on relevant information and assessments from the provider.

7.5 Pre-Deployment Discrimination Assessment Requirements in Other Legislation

7.5.1 The GDPR's Data Protection Impact Assessment Requirement

Unlike the AIA and MDR, the GDPR does not formally require the establishment of a risk management system. However, it does require that preventive compliance measures are

---

[702] Meike Zehlike et al., "Beyond Incompatibility: Interpolation between Mutually Exclusive Fairness Criteria in Classification Problems," *arXiv preprint arXiv:2212.00469* (2022); Cynthia Dwork et al., "Fairness through Awareness" (Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, Massachusetts, Association for Computing Machinery, 2012); Castelnovo et al. (2022); María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante, "Addressing Fairness in Artificial Intelligence for Medical Imaging," *Nature Communications* 13, no. 4581 (2022), https://doi.org/10.1038/s41467-022-32186-3; Bellamy et al. (2019) 77.

[703] Wachter, Mittelstadt, and Russell (2021 B) 29.

conducted before certain data processing activities are commenced. One such measure is the Data Protection Impact Assessment (DPIA) required according to Article 35 GDPR. Before commencing a type of processing that "is likely to result in a high risk to the rights and freedoms of natural persons," Article 35(1) GDPR places the data controller under the obligation to "carry out an assessment of the impact of the envisaged processing operations on the protection of personal data." As noted in chapter 6, an AI developer is the data controller for development activities, whereas a healthcare institution acting as deployer is typically the data controller for putting an AI-CDS system into service.

The obligation to conduct a DPIA applies prior to the development process for AI-CDS systems (provided that the systems are trained on personal data)[704] as well as prior to the putting into service and use of the systems.[705] Formally, Article 35 GDPR is articulated as applying to "processing operations" and not to products or systems as such. In practice, however, to properly address risks posed by the processing operations encompassed by development or deployment of an AI-CDS system, a DPIA must be conducted of the AI-CDS system as such.

The use of an AI-CDS system entails a series of processing operations which are repeated over time. Upon implementation, the deployer of an AI-CDS system may conduct one "single assessment" to consider the impact of the overall processing operations that will take place when the system is used, as permitted by the final sentence of Article 35(1) GDPR.[706] In the assessment, the deployer must consider risk scenarios arising from the context of use that is

---

[704] If the development relies exclusively on synthetic data, this would mean that the developers do not have to conduct a DPIA.

[705] Collection and use of health data for training and other development purposes require a DPIA, according to Article 35(3)(b) GDPR (referring to "processing on a large scale of special categories of data"). To deploy and use an AI-CDS system would usually entail "a type of processing in particular using new technologies," which means that a DPIA is required if an initial risk assessment suggests a high level of risks to the rights and freedoms of natural persons, cf. Article 35(1) GDPR.

[706] "A single assessment may address a set of similar processing operations that present similar high risks." The Article 29 Data Protection Working Party recognises that the DPIA can be used to assess technology products as such: Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (Dpia)* (13 October 2017), 8, https://ec.europa.eu/newsroom/article29/items/611236. However, for all practical purposes, the user must carry out a product-oriented DPIA before implementing AI-CDS systems.

envisaged for the system.[707] To carry out a DPIA of an AI-CDS system, some contributions from the AI provider is often necessary, unless the provider and the deployer are the same entity.

The mandatory content of a DPIA is outlined in Article 35(7) GDPR. It shall include a description of the processing activity and its purpose, an assessment of the necessity and proportionality of the processing, and – importantly – an assessment of the risks to individuals' "rights and freedoms" and measures to address these risks.[708] Thus, risk assessment is a central component of the DPIA requirement. Moreover, the risk assessment requirement in Article 35(7) is where Article 35 GDPR specifically refers to an assessment that might be interpreted as encompassing discrimination aspects. Whether this risk assessment requirement constitutes a pre-deployment discrimination assessment requirement depends on how one interprets the reference to "rights and freedoms." This reference determines the mandatory scope of the risk assessment required under Article 35 GDPR. However, there is considerable debate in academic literature regarding the extent of this reference and, consequently, the scope of the DPIA requirement. Some argue that the requirement in Article 35 GDPR encompasses all fundamental rights enshrined in the EU charter,[709] while others contend that the DPIA only pertains to the data protection obligations

[707] Thus, there are similarities between the situation when the user of an AI-CDS system conducts a DPIA of the system and the situation when the developer conducts a risk assessment of the system as a medical device, in accordance with risk management obligations in the MDR.

[708] Article 35(7) GDPR. Somewhat confusingly, Article 35(1) speaks of the rights and freedoms of "natural persons," whereas Article 35(7) speaks of the rights and freedoms of "data subjects." However, the latter provision's explicit reference to Article 35(1) makes it clear that the discrepancy can be disregarded when interpreting the provision.

[709] Katerina Demetzou, "Gdpr and the Concept of Risk," in *Privacy and Identity Management: Fairness, Accountability, and Transparency in the Age of Big Data*, ed. Eleni Kosta et al. (Cham, Switzerland: Springer, 2018), 143; Katerina Demetzou, "Risk to the 'Rights and Freedoms'," in *Data Protection and Privacy: Data Protection and Democracy*, ed. Dara Hallinan et al., Computers, Privacy and Data Protection (Oxford: Hart, 2020), 140; Dara Martin Hallinan, Nicholas, "Fundamental Rights, the Normative Keystone of Dpia," *European Data Protection Law Review* 6, no. 2 (2020): 179, https://doi.org/10.21552/edpl/2020/2/6.

and principles specified within the GDPR itself.[710] To date, the issue remains unresolved in CJEU case law.

According to Article 1(2) GDPR, the objective of the regulation is to protect "fundamental rights and freedoms of natural persons and *in particular* their right to the protection of personal data" (my italicisation). In a statement that appears to be anchored in this wording, the Article 29 Data Protection Working Party suggests that the reference to rights and freedoms "primarily concerns the rights to data protection and privacy but *may also involve* other fundamental rights such as freedom of speech, freedom of thought, freedom of movement, prohibition of discrimination, right to liberty, conscience and religion" (my italicisation).[711] The statement suggesting that the assessment *may* involve other fundamental rights than those relating to privacy and data protection does not clarify the extent to which these other fundamental rights *must* be assessed.

While the mandatory scope of the DPIA requirement could be clarified in future case law, it currently remains undetermined whether and to what extent a DPIA *must* include an assessment of the risk of discrimination. Although these issues present interesting research questions, they are not the focus of this thesis. However, the methodological elements developed in this thesis can be applied in the context of a DPIA, regardless of whether Article 35 GDPR necessitates an assessment of discrimination risks. There are compelling reasons for those conducting a DPIA to consider the risk of algorithmic discrimination as part of their assessment, even if it is not mandatory under Article 35 GDPR. Recognising this opportunity, legal scholarship has shed light on the potential of compliance measures within data protection law to contribute to compliance with the non-discrimination principle by including

---

[710] Gellert (2020) 199; Raphaël Maurice Gellert, "Why the Gdpr Risk-Based Approach Is About Compliance Risk, and Why It's Not a Bad Thing," *Trends and Communities of legal informatics: IRIS 2017 - Proceedings of the 20th International Legal Informatics Symposium* (2017): 529.

[711] This is based on the wording of Article 1(2) GDPR, which refers "in particular" to the right to data protection: Article 29 Data Protection Working Party, *Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks* (30 May 2014), 4, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf; Article 29 Data Protection Working Party (2017): 6.

aspects of discrimination when conducting a DPIA.[712] This scholarship emphasises the shortcomings of relying on ex post enforcement measures in non-discrimination law – an issue that was discussed in section 7.2 above.

7.5.2 The 'Activity Duty' as Risk Management in Norwegian and Swedish Non-Discrimination Law

As established in section 7.2, the Equality Directives do not currently lay down any pre-deployment discrimination assessment requirements. These Directives primarily rely on ex post enforcement measures. However, it is worth acknowledging that risk assessment as a preventive measure to ensure compliance is not completely absent in non-discrimination law at the national level. The EU Equality Directives were, for the most part, enacted in the early 2000s. Since then, considerable advancements have been made in non-discrimination law in many countries. These advancements reflect an increased recognition of the shortcomings associated with traditional ex post enforcement measures. As a result, there has been a proliferation of preventive enforcement strategies aimed at addressing these limitations.

Of particular importance is the adoption of variations of an 'activity duty' (*aktivitetsplik*t) in some countries' national non-discrimination law frameworks.[713] The activity duty is typically a duty to make proactive, targeted, and systematic efforts to promote equality and prevent discrimination.[714] In some cases, this duty encompasses a risk assessment obligation.

In Norway, the activity duty applies to public authorities and employers of a certain size. In Sweden, it applies to employers and education providers. As one of several components of the activity duty in those countries, Norwegian and Swedish non-discrimination laws prescribe

---

[712] Goodman (2016) 506; Hacker (2018) 1184; Laurens Naudts, "How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them," in *Data Protection and Privacy: The Internet of Bodies*, ed. Ronald Leenes et al., Computers, Privacy and Data Protection (Oxford: Hart, 2019), 83-84; Giacomo Capuzzo, "A Comparative Study on Algorithmic Discrimination between Europe and North-America," *Comparative Law Review* 10, no. 2 (Fall 2019): 142.

[713] In the context of the UK Equality Act 2010, the public sector activity duty has indeed been framed as a step towards 'meta-regulation' or 'enforced self-regulation': Bob Hepple, "Enforcing Equality Law: Two Steps Forward and Two Steps Backwards for Reflexive Regulation," *Industrial Law Journal* 40, no. 4 (2011): 319, https://doi.org/10.1093/indlaw/dwr020.

[714] This is based § 26 of the Norwegian Equality and Non-Discrimination Act.

risk management requirements aimed at preventing discrimination (in Norway the risk assessment component of the activity duty only applies to employers).[715] These requirements include steps that correspond to 'risk assessment' as understood within this thesis, cf. ISO 31000.[716]

There exists limited methodological guidance to how discrimination risk assessments should be conducted within the Norwegian and Swedish frameworks. Some guidance can be found in a report from the law committee that proposed the risk assessment requirement in Swedish non-discrimination law. According to this report, risk mitigation measures should be adopted whenever risks are identified which might disadvantage a protected group.[717] Moreover, the committee mentions some recommended ways of identifying risks: Review of procedures, guidance documents and policy documents within the organisation is recommended,[718] as well as surveys or interviews, group interviews, and other forms of dialogue.[719] It can be observed that these methodological recommendations reflect the fact that the risk assessment requirement relates to the operations and practices of the entities subject to the activity duty. These methodological recommendations arguably do not contemplate a risk assessment of specific products, such as AI systems. Hence, they do not offer assistance in the development of a pre-deployment discrimination assessment methodology specifically tailored to AI-CDS systems. In Norway, the legislature intentionally left the methodological aspects of discrimination risk assessment for the regulated entities to determine. It is reasoned in the

---

[715] Ibid; Swedish Act 5 June 2008 on Non-Discrimination (*Diskrimineringslag* (2008:567), chapter 3, § 2. Although variations of an 'activity duty' can be found also in other European countries, such as Finland, Denmark and the UK, the activity duty requirements are not articulated in terms of 'risk' in the laws of those countries: To varying degrees, these countries (and probably many other countries) have non-discrimination laws which require systematic activities from certain entities. The UK Equality Act 2010 provides a public sector equality duty and empowers the Minister to impose specific duties on authorities: Hepple (2011) 318.

[716] Norwegian Equality and Non-Discrimination Act, § 36; *Iso 31000:2018 (E)*, 11-12.

[717] *Nya Regler Om Aktiva Åtgärder Mot Diskriminering (Sou 2014:41)*, 243. ("Att de ska motsvara ett faktiskt behov innebär däremot inte att det måste finnas någon individ i verksamheten som kan drabbas, utan att det i verksamheten finns risker eller hinder som skulle kunna drabba någon av de grupper som diskrimineringsgrunderna ska skydda.")

[718] Ibid.

[719] Nya Regler Om Aktiva Åtgärder Mot Diskriminering (Sou 2014:41) 244.

preparatory works that the activity duty applies to a diverse category of entities (employers) and that it would therefore be challenging to specify the duty in more detail.[720]

Although Swedish and Norwegian non-discrimination laws include provisions for discrimination risk assessment, they do not offer clear guidance on the methodology for conducting such assessments, especially not in relation to AI-CDS systems. However, the methodological knowledge developed in this thesis could contribute to developing assessment methodologies for risk assessments required under the activity duty at the national level. This would particularly be the case if the risk assessment requirement should be extended to healthcare institutions in the future.

While there is little methodological guidance specifically regarding the discrimination risk assessment that is part of the activity duty in Norwegian law, the Norwegian Equality and Anti-Discrimination Ombud in November 2023 published a general guidance document titled '*innebygd diskrimineringsvern.'*[721] The title may be translated into 'embedded protection against discrimination' or, perhaps (less accurately), 'non-discrimination by design.'[722] The guidance document aims at facilitating the detection and prevention of discrimination in AI systems. The document, which constitutes a high-level introduction to the topic of preventing discrimination in AI systems, focusses on outlining relevant discrimination risks and the questions that stakeholders ought to address and discuss during development of AI systems.[723] For entities that are subject to the activity duty, consideration of the questions proposed in the Ombud's guidance may be part of their systematic work to prevent discrimination. For example, the proposed questions could be implemented into a risk assessment methodology.

Hence, both the Ombud's guidance and this thesis develop considerations which may be integrated into various assessment methodologies and used to assess discrimination in an AI

---

[720] Prop.81 L (2016-2017), 281.

[721] Likestillings- og diskrimineringsombudet, *Veileder for Innebygd Diskrimineringsvern* (2022), https://ldo.no/globalassets/_ldo_2019/_bilder-til-nye-nettsider/ki/ldo.-innebygd-diskrimineringsvern.pdf.

[722] The Ombud's guidance is partially inspired by Sloot et al.'s handbook on 'non-discrimination by design': Likestillings- og diskrimineringsombudet (2022): 19; Bart van Der Sloot et al., *Non-Discrimination by Design* (2023), https://www.tilburguniversity.edu/about/schools/law/departments/tilt/research/handbook.

[723] Likestillings- og diskrimineringsombudet (2022): 4.

system before its deployment. However, compared to the Ombud's guidance, this thesis takes a slightly different perspective. The guidance develops questions primarily for AI developers and stakeholders to consider during the development of an AI system. It is therefore structured around the development process.[724] In comparison, this thesis has a stronger orientation towards the *assessment* of an AI system: It primarily contemplates the situation where a developer, deployer or third-party assesses an AI system for the purpose of determining whether it can be deployed. Thus, in relation to the AI development process,[725] the most direct implications of this thesis concern the testing of AI models, because testing is a crucial aspect of assessing a system and determining whether it can or should be deployed. The structure in which the methodological elements developed in this thesis are presented differs, therefore, from the structure of the Ombud's guidance. However, the methodological elements developed in this thesis also illuminate the questions that developers should raise during the various stages of the development process, to ensure that an AI system will be positively assessed at a later stage.

Similarly, the Norwegian Ombud's guidance sheds some light on the *assessment* of AI systems, even though its principal focus is on the *development* of these systems. However, the Ombud's guidance is structured around the AI development process. This thesis and the Ombud's guidance complement each other. When it comes to the methodological elements of pre-deployment discrimination assessments, the thesis develops more detailed questions and directions on how to assess the implications of each question. Moreover, it identifies specific technical methods that may potentially be used to assess discrimination in practice. At the same time, the Ombud's guidance highlights how some similar considerations to the ones developed in this thesis may be integrated into an AI development process.

7.6 What is Required?

7.6.1 Do the Pre-Deployment Discrimination Assessment Requirements Refer to the Equality Directives or Article 21 of the Charter?

Pre-deployment discrimination assessment requirements have been identified in the AIA, cf. section 7.4. However, it is not immediately clear whether they should be understood as referring to Article 21 of the EU Charter or to the Equality Directives. As noted in section

[724] Likestillings- og diskrimineringsombudet (2022): 11.

[725] Section 1.5.9.

6.3.2, the distinction does matter: Unlike the Equality Directives, Article 21 of the Charter does not explicitly distinguish between direct and indirect discrimination, and the CJEU's jurisprudence seems to permit objective justification of potential direct discrimination in cases based directly on Article 21 of the Charter. There is a difference between basing a methodology for pre-deployment discrimination assessments on the general rule that direct discrimination is never justifiable except for specific statutory exceptions and basing it on a provision according to which potential direct discrimination can be justified. According to the European Commission's proposal, the AI Act "complements existing Union law on non-discrimination with specific requirements that aim to minimise the risk of algorithmic discrimination."[726] Notably, the phrase "Union law on non-discrimination" refers to EU non-discrimination law in a broad sense, presumably encompassing the Equality Directives as well as Article 21 of the Charter. Similarly, recital 44 AIA refers to "discrimination prohibited by Union law."[727] The various negotiation versions of the AIA also include references to the "right to non-discrimination" and specifically to Article 21 of the Charter.[728] On the other hand, however, the Parliament's Compromise Text also emphasises the need to address bias that creates direct or indirect discriminatory effects.[729] As noted, the distinction between direct and indirect discrimination is only explicitly acknowledged by the expression of the non-discrimination principle in the Equality Directives.[730]

Quite possibly, the EU legislature has not reflected on the difference between referring to Article 21 of the Charter and referring to the Equality Directives. As noted in section 6.3.1, they both mirror the same fundamental principle. To date, the difference that was highlighted in section 6.3.2 has not received much attention in literature or practice. In summary, it is most likely that all three negotiation versions of the AIA must be interpreted as referring to the non-discrimination principle as such, rather than categorically referring to Article 21 or

---

[726] AIA (EC), Explanatory Memorandum, 4.

[727] Recital 44 AIA (EP). Moreover, the EU Council's negotiation version (the General Approach), as well as the Parliament's version, include a reference to "discrimination prohibited under Union law" in Article 10 AIA. The Parliament's Compromise Text also includes a reference to "discriminatory impacts and unfair biases that are prohibited by Union or national law": Article 4(a) AIA (EP).

[728] Recitals 15, 17, 39 AIA (EC); Recital 16a AIA (EP).

[729] Recital 44 AIA (EP).

[730] Section 6.3.2.

the Equality Directives. The consequence of this interpretation is that the AIA refers to the provisions of EU non-discrimination law that are applicable to the underlying subject matter. Section 6.3.2 has already concluded that the RED and GSED most likely apply to clinical decision-making. Hence, if ethnic discrimination or sex discrimination occurs after the deployment of an AI-CDS system, there is a violation of these Directives. This implies that the pre-deployment discrimination assessment requirements in the AIA should also be interpreted as referring to the same Directives, as far as AI-CDS systems are concerned.

7.6.2 Implications of the Different Types of Pre-Deployment Assessments

7.6.2.1 What it Means to Assess Discrimination Before Deployment

The AIA's emphasis on assessing discrimination before deployment of an AI system is a novel development in EU law. This development is part of the EU's risk-based approach to AI regulation, in which meta-regulation is combined with command-and-control regulation. The pre-deployment discrimination assessments required in the AIA involve the application of the non-discrimination principle in a preventive context that differs from the ex post enforcement contexts in which this principle is traditionally applied. In an ex post enforcement context, such as in judicial proceedings, a judiciary concludes on whether the disputed factual events amount to discrimination.

In contrast, the pre-deployment discrimination assessments explored in this chapter do not involve assessing whether discrimination has occurred. Rather, a pre-deployment discrimination assessment aims to determine whether an AI-CDS system can be deployed. This is a decision that must be made at a time when discrimination has not actually occurred - there is arguably no violation of the non-discrimination principle before an AI-CDS system is deployed. The question is, therefore, what it means to assess discrimination before deployment. This is a question of what the discrimination assessment requirements identified in this chapter require before an AI-CDS system can be deployed.

However, before discussing the implications of the different assessment requirements, it is worth noting that the Equality Directives' prohibitions on direct and indirect discrimination in themselves suggest that the very deployment of an AI-CDS system might be unlawful, regardless of any pre-deployment discrimination assessment requirements. Indirect discrimination occurs as soon as a practice exists that "would put" persons from a protected

group at a particular disadvantage.[731] It follows from the wording that an actual victim of discrimination need not be identified. However, the "would put" wording does not suggest that a practice amounts to indirect discrimination if there is merely a chance that the system might place persons from a protected group at a particular disadvantage. "Would" is a word that normally implies a relatively high level of probability. It is not clear exactly how probable the occurrence of a particular disadvantage must be under EU law, for indirect discrimination to arise. Nevertheless, the deployment of an AI-CDS system can, in theory, amount to a violation of the prohibition on indirect discrimination as of the moment it is deployed, if the system causes a particular disadvantage that is not objectively justified. The AIA's discrimination assessment requirements are intended to prevent discrimination from occurring upon deployment or during the use of an AI system. The probability that a system "would put" persons from a protected group at a particular disadvantage is one aspect that needs to be considered when conducting a pre-deployment discrimination assessment.

CJEU jurisprudence indicates that the prohibition on direct discrimination could also be violated regardless of whether any specific individual has been harmed by a practice.[732] The *Feryn* case concerned an employer that publicly declared that it would not recruit persons of a certain ethnic origin. The CJEU's ruling remarks that such a statement is likely to dissuade certain candidates from applying and therefore constitutes direct discrimination.[733] The *Feryn* case refers to publicly outspoken ethnic discrimination, and the argument from that case is not easily adaptable to the context of AI-CDS systems. It cannot be ruled out that such a system may constitute direct discrimination as of its moment of deployment. For example, one may envisage a system that relies, unlawfully, on a protected characteristic among its feature variables and this is communicated through publicly accessible system documentation. This could in theory dissuade certain persons from seeking healthcare. However, according to the discrimination assessment requirements in the AIA, as further discussed below, the point is not to assess whether the act of deployment will in itself constitute discrimination. Rather, these requirements are oriented towards the probability that discrimination may occur at some point during the use of the system in practice.

---

[731] Article 2(2)(b) RED; Article 2(b) GSED.

[732] Judgment of 10 July, 2008, Feryn, C-54/07, ECLI:EU:C:2008:397, paras. 25 and 28.

[733] Feryn, C-54/07, para. 25.

The standard for deployment, when assessing discrimination in an AI-CDS system, ultimately depends on the type of discrimination assessment one applies. 'Assessing discrimination' means different things depending on whether the assessment is a discrimination risk assessment, bias examination, or discrimination impact assessment. The implications of these respective discrimination assessment types are further discussed in the following.

7.6.2.2 Discrimination Risk Assessment

Section 7.4.1. identified the AIA's risk assessment requirement as a pre-deployment discrimination assessment requirement. This raises the question of what an assessment of the 'risk' of discrimination implies for the decision to deploy an AI-CDS system. The main implication is that the decision on whether to deploy an AI-CDS system is determined with reference to 'risk.' However, 'risk' is a word denoting multiple meanings in different contexts.[734] Garland points out that the body of works sometimes called 'risk literature' actually covers "different projects, different forms of inquiry, and different conceptions of their subject matter, all linked tenuously together by a tantalizing, four-letter word…"[735] However, it is usually accepted that risk is "something which might or might not occur";[736] that risks are important because risks have consequences;[737] and that risk is assessed by considering the combination of their probability of occurrence and the consequences of their occurrence.[738]

The management and assessment of risks is a distinct, yet diverse, field of knowledge and practice – in which methodologies for identifying, assessing, and mitigating risks have been developed.[739] While there are many areas where risk assessment methodologies are well-

---

[734] Consider, for example, the overview provided by Aven and Renn: Terje Aven and Ortwin Renn, *Risk Management and Governance: Concepts, Guidelines and Applications*, ed. Jeryl L. Mumpower and Ortwin Renn, Risk, Governance and Society, (Heidelberg: Springer, 2010), 2-3.

[735] David Garland, "The Rise of Risk," *Risk and morality* 1 (2003): 49.

[736] David Hillson and British Standards Institution, *The Risk Management Universe: A Guided Tour*, 2nd ed. (London: BSI, 2007), 378.

[737] Ibid.

[738] *Iso 31000:2018 (E)*, 6; Aven and Renn (2010) 2-3.

[739] e.g., Hillson and British Standards Institution (2007); Terje Aven et al., *Risikoanalyse: prinsipper og metoder, med anvendelser*, 2nd ed. (Oslo: Universitetsforlaget, 2017), 19; Aven and Renn (2010) 121; Centre for Information Policy Leadership (CIPL), *Risk, High Risk, Risk Assessments*

established, the nearest point of reference in the context of AI-CDS systems is health and safety-oriented risk assessments.[740] In this area, risk assessments are traditionally oriented towards the quantification of foreseeable risks and the protection of health and safety from those risks. The MDR's risk assessment requirement is a relevant example.[741] In the MDR, risk assessment is a measure aimed at ensuring the safety and performance of medical devices before they are placed on the EU market. This type of risk assessment aims at assessing the risk posed by a medical device to the health and safety of patients. It does not aim at assessing the risk of non-compliance with laws. In contrast, risk assessment as discrimination assessment essentially involves assessing the risk of non-compliance with the non-discrimination principle.

The safety of a medical device is often assessed based on data from clinical investigations (also called 'clinical trials').[742] The purpose of clinical investigations is to produce data (statistics) that can be used for risk assessment purposes, among other purposes, so that it can be determined whether the risks associated with a device are acceptable "when weighed against the benefits to be achieved by the device."[743] For instance, two hundred patients

---

and Data Protection Impact Assessments under the Gdpr (21 December 2016), 10, https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_21_december_2016.pdf.

[740] The development within EU law, where risk assessment has been extended from highly technical areas to broader social and societal issues, can be understood by reference to the German sociologist Ulrich Beck's theory of the 'risk society': Ulrich Beck. *Risk Society: Towards a New Modernity*. London: Sage Publications, 1992. Weimer attributes to Beck's risk society the fact that "the notion of risk is being associated not only with highly specialized technical areas but also with broader societal questions about the 'way of life' in late modernity": Maria Weimer, *Risk Regulation in the Internal Market: Lessons from Agricultural Biotechnology*, ed. Paul Craig and Gráinne de Búrca, Oxford Studies in European Law, (Oxford: Oxford University Press, 2019), 20. Similarly, Hood, Rothstein and Baldwin note that, in addition to risks which are knowable through scientific investigation, "Beck and associated thinkers would add other cases they see as typical of 'advanced modernity', like the risks associated with genetically modified organisms, reproductive technology, or computer failures that also potentially impact on a wide range of everyday life": Christopher Hood, Henry Rothstein, and Robert Baldwin, *The Government of Risk: Understanding Risk Regulation Regimes* (Oxford: Oxford University Press, 2010), 4.

[741] Article 52(1) MDR.

[742] Article 61 MDR and Part A of Annex XIV MDR.

[743] Article 62(1)(c) MDR.

undergoing knee replacement surgery may opt to enrol in a clinical investigation where they get a new type of knee prosthesis instead of the standard prosthesis they would otherwise be offered in normal clinical care. The clinical investigation study follows the patients for 12 months to see if the undesirable event of prosthesis failure occurs within this period. If ten out of the two hundred patients experience prosthesis failure, this suggests that the probability of prosthesis failure within 12 months is 10/200 = 5 %. Depending on how dangerous prosthesis failure is for patients, it may be decided that the risk is outweighed by the benefits of the prosthesis. If so, the risk is deemed to be acceptable, and the prosthesis can be placed on the market.

As other scholars have pointed out, it is not obvious how risk assessment methodologies can be aligned with the notion of fundamental rights, or how an assessment of 'risks to rights' is properly conducted. These are questions that have, particularly, been raised in the context of the GDPR's DPIA requirement, which requires a consideration of risks to fundamental rights, as discussed in section 7.5.1.[744] Macenaite notes that EU data protection law operationalises the fundamental rights to privacy and data protection and is therefore different from laws that primarily address more traditional types of 'risk,' typically health and safety risks.[745] Gonçalves underscores that risks conventionally presuppose some calculation (which implies quantification), typically related to physical or technical impacts.[746] Van Dijk, Gellert and Rommeltveit note that risks to "rights and freedoms" (as per Article 35 GDPR) as the object of risk assessment may obfuscate the application of traditional risk management methodologies, including that of the ISO framework.[747] Moreover, Yeung and Bygrave note that "the relationship between risk and fundamental rights is poorly theorized in human rights

[744] Section 7.4.2.

[745] Milda Macenaite, "The "Riskification" of European Data Protection Law through a Two-Fold Shift," *European Journal of Risk Regulation* 8, no. 3 (2017): 533, https://doi.org/10.1017/err.2017.40.

[746] Maria Eduarda Gonçalves, "The Risk-Based Approach under the New Eu Data Protection Regulation: A Critical Perspective," *Journal of Risk Research* 23, no. 2 (2020): 145, https://doi.org/10.1080/13669877.2018.1517381.

[747] Niels van Dijk, Raphaël Gellert, and Kjetil Rommetveit, "A Risk to a Right? Beyond Data Protection Risk Assessments," *Computer Law & Security Review* 32, no. 2 (2016), https://doi.org/10.1016/j.clsr.2015.12.017.

scholarship generally."[748] Thus, it is widely recognised in legal scholarship that the merger of risks and rights is a challenging one.

Indeed, there is a potential misalignment between traditional notions of 'risk' and the nature of fundamental rights.[749] This misalignment arises from the idea that fundamental rights are inviolable, which leads to the argument that practices that even have a low probability of violating these rights should be considered unacceptable.[750] In contrast, when risk assessment is relied on as a type of discrimination assessment, it implies that the deployment of an AI-CDS system can be deemed acceptable, even if the potential for discriminatory outcomes for individuals is not eliminated. This would be similar to how the deployment of a medical device in accordance with the MDR, following risk assessment, does not mean that the device will never cause injury to patients; and a security breach in a data processing system does not necessarily constitute a violation of the requirements relating to the security of processing in Article 32 GDPR.[751]

In general, the purpose of risk assessment is typically to form a basis for a decision on whether or not to do something. When risk assessment is relied on as a type of pre-deployment discrimination assessment, the decision revolves around whether to deploy an AI system. Unless all identified risks can be eliminated, the decision to do deploy the system involves the taking of risk and, thus, the acceptance of risk. This is the fundamental nature of risk assessment. Consequently, based on risk assessment, an AI-CDS system can be deployed even though it carries a certain risk of algorithmic discrimination and, thus, interferes with the

---

[748] Karen Yeung and Lee A. Bygrave, "Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship," *Regulation & Governance* 16: 144, https://doi.org/doi.org/10.1111/rego.12401.

[749] Yeung and Bygrave (2022) 146.

[750] There is a conflict between risks and rights if, as Hansson states, "it could be claimed that if I have a right that you do not bring about a certain outcome, then I also have a right that you do not perform any action that has a non-zero probability of leading to that outcome": Sven Ove Hansson, "Ethical Criteria of Risk Acceptance," *Erkenntnis* 59, no. 3 (2003): 291 and 98, https://doi.org/10.1023/A:1026005915919; Claudia Quelle, "Does the Risk-Based Approach to Data Protection Conflict with the Protection of Fundamental Rights on a Conceptual Level? (Preprint)," Available at SSRN 2726073 (2015): 3.

[751] Lee A. Bygrave, "Security by Design: Aspirations and Realities in a Regulatory Context," *Oslo Law Review* 8, no. 3 (2021): 167, https://doi.org/10.18261/olr.8.3.2.

right to non-discrimination. In other words, the acceptable risk posed by an AI-CDS system in a pre-deployment setting implies that deployment would constitute an *interference* with, but not an *infringement* of, the right to non-discrimination.

According to the ISO's risk management methodology, the acceptance of risk requires the establishment of 'risk criteria,' i.e., "terms of reference against which the significance of a risk is evaluated."[752] The methodological elements proposed in this thesis could be used as a starting point for developing specific risk criteria for assessing algorithmic discrimination risk in an AI-CDS system. Further development of methodological elements of discrimination assessment *tailored to the risk assessment context* is a task worth pursuing, which requires efforts beyond the scope of this thesis. However, the methodological elements developed in this thesis could serve as a foundation for such endeavours. Starting with these elements as a basis, future research could consider in further detail how the probability of discrimination risks and the severity of their consequences may be assessed.

7.6.2.3 Discrimination Impact Assessment

Section 7.4.3 found that a discrimination impact assessment might become included as a pre-deployment assessment requirement in the AIA. Such a discrimination impact assessment could be envisaged as a separate requirement for AI-CDS systems or it could be fitted within existing proposals for a 'fundamental rights impact assessment.' If it is not included in the AIA, it could become part of future legislation, as there is considerable support for the use of impact assessments to ensure effective protection of fundamental rights in the context of AI technologies.

The main implication of assessing discrimination as part of a pre-deployment impact assessment, is that deployment of an AI-CDS system is decided with reference to the 'impact' of deploying the system on the fundamental right to non-discrimination. However, the existing proposal to include a discrimination impact assessment requirement in the AIA is ambiguous in terms of how such an assessment could be conducted in practice.[753] As noted in section 7.4.3, the fundamental rights impact assessment proposed by the European Parliament

---

[752] *Iso 31000:2018 (E)*, 10-11. Similarly, risk management literature sometimes refer to the definition of an entity's "risk tolerance": Aven and Renn (2010) 170-81 and 223.

[753] Section 7.4.3.

refers to risks as well as reasonably foreseeable impacts of deploying an AI system.[754] However, the proposed requirement does not clarify which considerations an assessor should take into account in a pre-deployment assessment to determine the impact of an AI system on the right to non-discrimination.

Like risk assessments, impact assessments can serve many different objectives and, thus, measure 'impact' by reference to different objects. For example, there are environmental impact assessments, health impact assessments, and social impact assessments.[755] Consequently, there may be many different impact assessment methodologies. Moreover, the relationship between impact assessment and risk assessment is not always clear. Risk assessment is often a component of an impact assessment, in which case the impact assessment encompasses broader considerations in addition to risk assessment.[756] A risk assessment may also be utilised as a preliminary assessment to determine whether, subsequently, a broader impact assessment should be conducted. In relation to the Data Protection Impact Assessment requirement in Article 35 GDPR, risk assessment is framed both as a part of a broader impact assessment and as a preliminary assessment determining the need for an impact assessment. Section 7.5.1 focussed on the risk assessment component of Article 35 GDPR because the risk assessment requirement is the part of Article 35 GDPR which may be interpreted as encompassing discrimination aspects.

Much like how the notion of 'risk' is challenging to align with the fundamental right to non-discrimination, the idea of deploying an AI-CDS system based on an assessment of its discriminatory impacts also finds itself in a certain tension with the inviolable right to non-discrimination. If a discrimination impact assessment becomes required for AI-CDS systems, now or in the future, there is a need to develop a methodology for such an assessment beyond the contributions of this thesis. However, the contributions from this thesis may be worth considering as a starting point for the development of an assessment methodology tailored to

---

[754] Article 29 a(1)(e) AIA (EP).

[755] Kendyl Salcito et al., "Experience and Lessons from Health Impact Assessment for Human Rights Impact Assessment," *BMC international health and human rights* 15, no. 24 (2015): 1, https://doi.org/10.1186/s12914-015-0062-y; Mantelero (2022) 25.

[756] For example, the human rights impact assessment model developed by Mantelero includes risk assessment as a component: Mantelero (2022) 17.

the fundamental rights impact assessment context.[757] In this context, the considerations and principles/criteria developed in this thesis may be used to determine the 'impact' of an AI-CDS system on the right to non-discrimination.

7.6.2.4 Data Bias Examination

The risk- and impact assessment requirements discussed in this chapter can to some extent be understood in the light of existing methodologies developed for various risk- and impact assessments not related to discrimination. In contrast, the AIA's data bias examination requirement, discussed in section 7.4.2, is a type of pre-deployment assessment for which the underlying assessment methodology is less established. However, out of the three types of discrimination assessments identified within the relevant legal framework, the data bias examination requirement is the most straightforward one in terms of its implications for the decision to deploy an AI-CDS system. This type of discrimination assessment concentrates on the probability that bias in training data might lead to discrimination if an AI-CDS system is deployed. It is implied that a system can be deployed if non-discrimination is more likely than discrimination.

A data bias examination should particularly consider the sources of data bias discussed in section 4.4.2 of this thesis and assess the probability that they might lead to discrimination. The considerations that should be taken into account when assessing the probability of discrimination arising from these sources are not immediately evident. The relevant considerations require careful interpretation and contextualisation of the non-discrimination principle. This thesis offers one step in the direction of developing the relevant considerations, principles, criteria, and methods that should be applied, based on the non-discrimination principle in EU law. However, this thesis focusses on developing methodological elements that can be applied across different underlying assessment methodologies, including risk and impact assessments. Thus, while it provides certain pointers, it does not aim to fully address the specific issue of exactly how to estimate the

[757] Mantelero and Esposito note that "In considering the impact of AI on human rights, the dominant approach in many documents is mainly centred on listing the rights and freedoms potentially impacted rather than operationalising this potential impact and proposing assessment models": Mantelero and Esposito (2021) 10. Furthermore, they argue that there is a need for development of "a more tailored model of impact assessment, at the same time avoiding mere theoretical abstractions based on generic decontextualised notions of human rights": Mantelero and Esposito (2021) 8.

probability of discrimination. Further adapting the methodological contributions of this thesis to the specific issue of probability estimation should therefore be considered as a potential topic for future research.

7.7 Conclusion

The most important context for application of a pre-deployment discrimination assessment methodology is where a pre-deployment discrimination assessment is legally mandated by EU law. In this context, an assessment shall be conducted by the provider of an AI-CDS system or by the deployer of the system. In addition, notified bodies involved in conformity assessment may review the discrimination assessment conducted by the provider of an AI-CDS system, and may conduct an independent discrimination assessment if deemed necessary. It can also be envisaged that supervisory authorities or other external auditors may conduct pre-deployment discrimination assessments. Furthermore, the development of a discrimination assessment methodology is relevant for providers of AI-CDS systems intending to assess their models voluntarily at different stages of the development process. While the AIA requires that discrimination is assessed before an AI-CDS system is placed on the market or put into service, providers might want to assess discrimination also at earlier stages of development, to address compliance issues proactively and avoid discovering discrimination issues closer to deployment.

This chapter has identified three types of pre-deployment discrimination assessments that are mandatory according to the Parliament's Compromise Text: discrimination risk assessment, discrimination impact assessment, and data bias examination. If all these requirements are not included in the final version of the AIA, all pre-deployment assessment types discussed in this chapter are nonetheless encouraged regardless of the scope of mandatory obligations.

The assessment of discrimination in an AI-CDS system, and the subsequent decision to deploy an AI-CDS system, do not rule out the possibility of discrimination arising after deployment. Even if testing conducted as part of a pre-deployment discrimination assessment indicates that a patient in the test dataset is discriminated against (or rather, would have been discriminated against if the test scenario had occurred after deployment), this does not necessarily mean that an AI-CDS system cannot be deployed. The benchmark for deployment depends on the underlying assessment methodology that one applies. This chapter has highlighted the types of benchmarks that are referred to in the relevant provisions laying down pre-deployment discrimination assessment requirements. These provisions imply that

the decision to deploy an AI-CDS system may be determined by reference to the risk of discrimination (risk assessments), the impact on the right to non-discrimination (impact assessments), or the probability of discrimination (biases 'likely' to lead to discrimination).

Consequently, to 'assess discrimination in an AI-CDS system' in accordance with these provisions does not mean that a system is deemed discriminatory or not discriminatory. Rather, the *probability of discrimination* emerges as a central benchmark for determining whether an AI-CDS system can be deployed based on a pre-deployment assessment. In addition to being the centre of attention for the AIA's data bias examination, probability is also a central aspect of risk assessment.

In addition to probability, the other main benchmark for deployment, according to the pre-deployment assessment requirements identified in this chapter, is *impact*. Impact, in this context, means the same as 'consequence.' A discrimination risk assessment requires that deployment is decided based on a combination of probabilities and consequences. Moreover, a discrimination impact assessment is predominantly oriented towards impact. Hence, the decision to deploy an AI-CDS system in accordance with the relevant legal framework hinges on the probability that discrimination might occur, and/or the consequences deployment might have for the right to non-discrimination.

How to measure the probability of discrimination and the consequences/impact on the right to non-discrimination, when contemplating the deployment of an AI-CDS system, is not obvious. As this chapter has highlighted, if discrimination occurs when a system is deployed or during its use, this would be considered a very serious consequence. If an AI-CDS system is found likely to cause direct discrimination, it therefore seems clear that the system cannot be deployed based on the types of discrimination assessments examined in this chapter. The decision on whether to deploy a system is more difficult where there is a lower probability of discrimination, or where the system is likely to interfere with the right to non-discrimination without infringing this right. This could, for example, occur where there is a particular disadvantage, but the assessor considers that the disadvantage is objectively justified (see chapters 9 and 11).

By developing methodological elements for pre-deployment discrimination assessments, this thesis outlines important considerations, principles/criteria, and methods that should be included when assessing the probability of discrimination and the consequences of deploying

an AI-CDS system. However, it does not emphasise the further development of a risk assessment methodology, a bias examination methodology, or an impact assessment methodology, specifically. Further attuning the methodological elements developed herein to the respective types of discrimination assessments that are required now or in the future, remains a task for future research efforts. With this in mind, Part IV turns to the non-discrimination principle in EU law, aiming to develop methodological elements of pre-deployment discrimination assessment based on this principle.

PART IV: DISCRIMINATION

8 Direct and Indirect Algorithmic Discrimination

8.1 Introduction to Part IV

Part III introduced the relevant legal framework and identified pre-deployment discrimination assessment requirements within this framework. The pre-deployment discrimination assessment requirements are found outside of non-discrimination law, but they are intended to ensure non-discrimination. However, these provisions do not define what discrimination is. To understand what discrimination is and, thus, what it is that one needs to consider when 'assessing discrimination,' Part IV analyses the non-discrimination principle in EU law and develops methodological elements of assessing discrimination based on this principle, in a pre-deployment setting. The analysis concentrates on the non-discrimination principle as it is expressed in the Equality Directives, given that Part III found the RED and GSED to be applicable to discrimination in clinical decision-making. While the Directives prohibit different types of discrimination, direct and indirect discrimination are most important in the context of this thesis.

In Part IV, chapters 9-11 each analyse specific conditions that would have to be met for direct or indirect discrimination to be found in an ex post enforcement context. Based on the analyses, methodological elements of assessing discrimination in a pre-deployment setting are developed in each chapter. First, chapter 8 takes a slightly different approach by exploring the distinction between direct and indirect discrimination, rather than analysing a specific condition for finding discrimination. It is argued that this distinction should be reflected by a pre-deployment discrimination assessment methodology, due to its importance in EU law.

8.2 Importance of Distinguishing Between Direct and Indirect Discrimination: Purpose of Chapter 8

The distinction between direct and indirect discrimination is crucial in EU law, because direct discrimination normally cannot be justified according to the Equality Directives, whereas indirect discrimination does not arise if there is an objective justification for the measure at

issue.[758] Under the RED, direct discrimination can only be justified by reference to necessary occupational requirements.[759] Occupational requirements are relevant in matters concerning employment and working life, but not in the context of clinical decision-making. Consequently, the RED does not permit justification of direct discrimination in this context. Under the GSED, which prohibits sex discrimination, there are two possible justifications for treatment that would otherwise be seen as direct discrimination. First, the Directive permits "favourable provisions concerning the protection of women as regards pregnancy and maternity."[760] Second, it accepts that goods or services may be provided "exclusively or primarily to members of one sex" when this is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.[761] The latter justification option suggests that AI-CDS systems may be developed and provided specifically for men or women. However, it does not apply to the more common situation where an AI-CDS system is intended for men and women alike.

In ex post litigation based on the Equality Directives, the distinction between direct and indirect discrimination is crucial because the Directives provide very different possibilities of justification depending on whether the disputed measure constitutes direct discrimination or potential indirect discrimination.[762] While direct discrimination cannot be justified save for specific exceptions provided for by the applicable Directive, it follows from the definition of indirect discrimination that indirect discrimination does not arise if there is an objective justification for the measure at issue.[763]

Given its importance, a methodology for pre-deployment discrimination assessments should operationalise the distinction between direct and indirect discrimination. Specifically, such

---

[758] This follows from the definition of indirect discrimination: Article 2(2)(b) RED; Article 2(b) GSED.

[759] Article 4 RED.

[760] Article 4(2) GSED.

[761] Article 4(5) GSED.

[762] Christa Tobler, *Limits and Potential of the Concept of Indirect Discrimination*, European Commission Directorate-General for Employment, Social Affairs and Equal Opportunities (Luxembourg: Office for Official Publications of the European Communities, 2008), 48; Henrard (2019) 107. (noting how the distinction in EU law "translates into radically different justification mechanisms.")

[763] Judgment of 18 November, 2010, Kleist, C-356/09, ECLI:EU:C:2010:703, para. 41.

assessments should consider the extent to which there is a potential for direct discrimination in an AI-CDS system and take into account the non-justifiability of this type of discrimination. Moreover, as regards the *severity* of risks or impacts, the occurrence of direct discrimination should, arguably, be treated as a particularly severe consequence of deploying an AI-CDS system. The restricted scope for justifying direct discrimination pursuant to the Equality Directives suggests that this type of discrimination is considered particularly grave.

However, the distinction between direct and indirect discrimination is a debated topic in EU non-discrimination law.[764] The wording of the Equality Directives provides little guidance, and the CJEU tends not to clearly articulate the general rules it relies on to draw the distinction. This chapter nonetheless considers the CJEU's jurisprudence (section 8.5) for the purpose of adapting the lessons learned from this jurisprudence to the context of pre-deployment discrimination assessment of an AI-CDS system. The analysis leads to insights that are key to operationalising this dichotomy within a pre-deployment discrimination assessment: It is argued that such an assessment should involve an examination of whether an AI-CDS system incorporates any protected characteristics or inextricably linked features variables, because this would indicate that the system might cause direct discrimination. The criteria for determining whether a factor is inextricably linked to a protected characteristic are outlined based on an analysis of CJEU jurisprudence in cases where it has expanded the scope of the prohibition on direct discrimination, compared to the traditional starting point that direct discrimination is where a protected characteristic is explicitly relied on in a decision-making process or there is a discriminatory intent or motive involved. To set the stage for the analysis, section 8.3 reiterates the definitions of direct and indirect discrimination in EU law and section 8.4 traces their historical origins. The CJEU's case law is analysed in section 8.5,

---

[764] Christopher McCrudden, "The New Architecture of Eu Equality Law after Chez: Did the Court of Justice Reconceptualise Direct and Indirect Discrimination?," *European Equality Law Review* forthcoming (2016): 6; Hugh Collins and Tarunabh Khaitan, "Indirect Discrimination Law: Controversies and Critical Questions," in *Foundations of Indirect Discrimination Law*, ed. Hugh Collins and Tarunabh Khaitan (Oxford: Hart, 2018), 18-19. Henrard points out how the dividing line between direct and indirect discrimination is particularly difficult to draw in cases of "systemic discrimination, and deep-seated prejudices against a group": Henrard (2019) 114. See also Sandra Fredman, "Direct and Indirect Discrimination: Is There Still a Divide?," in *Foundations of Indirect Discrimination Law*, ed. Hugh Collins and Tarunabh Khaitan (Oxford: Hart, 2018); Adams-Prassl, Binns, and Kelly-Lyth (2023) 156.

before section 8.6 discusses the implications of that analysis and develops the thesis's first methodological elements of assessing discrimination in an AI-CDS system before deployment.

Legal scholarship on algorithmic discrimination, although still in its relatively nascent stages at the time of writing, has often been guided by the presumption that the rule of indirect discrimination is more apt to capture algorithmic discrimination, than the rule of direct discrimination.[765] It is typically assumed that the prohibition of direct discrimination does not significantly contribute to combatting algorithmic discrimination, because direct discrimination only arises where a decision procedure relies explicitly on a protected characteristic or an intent or motive is involved.[766] A valuable corrective to the tendency to play down the importance of direct algorithmic discrimination has been provided by Adams-Prassl, Binns and Kelly-Lyth.[767] The present chapter partially aligns with their contrarian stance, suggesting that the prohibition of direct discrimination in EU law does, indeed, hold substantial relevance in the context of algorithmic discrimination from AI-CDS systems.

8.3 Reiteration of the Definitions

Direct discrimination occurs where a person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of a protected characteristic.[768]

---

[765] Tischbirek assumes that the doctrine of direct discrimination in EU non-discrimination law rarely will apply to AI: Alexander Tischbirek, "Artificial Intelligence and Discrimination: Discriminating against Discriminatory Systems," in *Regulating Artificial Intelligence*, ed. Thomas Wischmeyer and Timo Rademacher (Springer Nature Switzerland AG, 2020), 114. Similarly, Hacker assumes that "[i]n machine learning contexts, direct discrimination will be rather rare": Hacker (2018) 1151. (Hacker reasons that the discriminating measure "must either be directly motivated by the protected criterion or the decision maker must directly relate to it.") Furthermore, Wachter, Mittelstadt and Russell assume that "direct discrimination cases are in all likelihood going to be rare in automated systems": Wachter, Mittelstadt, and Russell (2021 B) 19-20. See also: Gerards and Xenidis (2021): 67.

[766] Hacker claims that "direct discrimination does not cover indirect proxy discrimination (via correlations with a seemingly neutral criterion[…])": Hacker (2018) 1151. Similarly, Wachter, Mittelstadt and Russell assert that, for direct discrimination to occur, "the rule, practice or action alleged to be discriminatory must explicitly refer to a protected characteristic": Wachter, Mittelstadt, and Russell (2021 B) 8; Wachter, Mittelstadt, and Russell (2021 B) 15-16.

[767] Adams-Prassl, Binns, and Kelly-Lyth (2023).

[768] Article 2(2)(a) RED; Article 2(1) GSED.

For example, direct ethnic discrimination occurs if an Asian job applicant does not get a job because the employer believes that Asians are less effective workers than other people. In healthcare, a clear example of direct discrimination would be to prioritise the treatment of a woman over a similarly situated man because a clinician generally practices a 'women first' policy. The prohibition on direct discrimination entails, at its core, that individuals must be treated equally regardless of the characteristics that are protected by the relevant non-discrimination laws.

Indirect discrimination occurs, according to the RED, where "an apparently neutral provision, criterion or practice would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons."[769] In the GSED, the definition of indirect discrimination is the same, except that it compares "persons of one sex" to "persons of the other sex."[770] The phrase "provision, criterion or practice" is abbreviated as 'PCP' in the following.

As noted in section 4.2, the prohibition on indirect discrimination promotes substantive equality, whereas the prohibition on direct discrimination primarily ensures formal equal treatment. Thus, indirect discrimination could encompass instances where formal equal treatment would put a protected group at a particular disadvantage. Indirect discrimination laws are typically characterised as being concerned with the effects of a practice or decision, rather than their appearance.[771]

Unsurprisingly, the Equality Directives do not make any specific mention of AI or ML algorithms. This lack of explicit reference can be attributed to the fact that these technologies were not widely used in decision-making practices at the time when the Directives were drafted. However, the definition of indirect discrimination provided in the Directives is sufficiently broad to encompass AI systems, as they can easily be categorized as a "provision, criterion or practice."[772] Similarly, the rule against direct discrimination is technology-neutral and applicable to any measure that treats an individual less favourable than another based on a

[769] Article 2(2)(b) RED.

[770] Article 2(b) GSED.

[771] Fredman (2022) 280-82.

[772] Tobler notes that "the legal concept of indirect discrimination relates to measures in the broadest meaning of the word": Tobler (2008): 29. See also: Allen and Masters (2020) 592.

protected characteristic. Consequently, discriminatory AI systems fall within the purview of both rules.

8.4 Historical Origins, Traditional Starting Points, and Progressive Expansion of the Rule Against Direct Discrimination

8.4.1 Historical Origins of Direct and Indirect Discrimination Rules

Since the foundation of the European Economic Community (EEC, now EU), the non-discrimination principle (then referred to as the 'equal treatment principle') followed from the Rome Treaty, the original constitution of the EEC.[773] In the Rome Treaty, discrimination on grounds of nationality was prohibited according to Articles 7 and 48, while Article 119 demanded equal pay for men and women. These provisions prohibited differential treatment between comparable cases and did not expressly address indirect discrimination.[774]

A doctrine of indirect discrimination, according to which facially neutral measures could violate the equal treatment principle, was eventually developed by the CJEU. In academic literature, it is often asserted that the CJEU's development of an indirect discrimination rule emerged after the US Supreme Court's establishment of the 'disparate impact' doctrine in *Griggs v Duke Power* (1971) and the UK's subsequent adoption of a statutory indirect discrimination rule.[775] However, Tobler refers to the *Ugliola* ruling (1969) concerning

---

[773] European Economic Community, Treaty Establishing the European Economic Community (Rome Treaty), 25 March 1957.

[774] Christa Tobler, *Indirect Discrimination under Directives 2000/43 and 2000/78*, European Commission Directorate-General for Justice and Consumers (Luxembourg: Publications Office of the European Union, 2022), 50 and 53. The principle of equal treatment in relation to nationality is now expressed in Article 21(2) of the EU Charter and Article 18 TFEU. The part of the non-discrimination principle that specifically addresses protected characteristics other than nationality is now enshrined in Article 21(1) of the EU Charter and Article 19 TFEU.

[775] Judgment of the US Supreme Court of 8 March, 1971, Griggs V Duke Power Co., 401 U.S. 424. Hunter and Shoben describe how the introduction of indirect discrimination in UK law occurred after a visit to the US from the British Home Secretary: Rosemary C Hunter and Elaine W Shoben, "Disparate Impact Discrimination: American Oddity or Internationally Accepted Concept," *Berkeley Journal of Employment and Labor Law* 19, no. 1 (1998): 109 and 15. See also: Joseph Seiner, "Disentangling Disparate Impact and Disparate Treatment: Adapting the Canadian Approach," *Yale Law & Policy Review* 25, no. 1 (Fall 2006): 117.

discrimination based on nationality as an early sign of an indirect discrimination rule in EU law.[776] Indeed, in *Ugliola,* the Court holds that a Member State would violate the equal treatment principle by "indirectly introducing discrimination in favour of their own nationals."[777]

Although a trace of indirect discrimination reasoning occurs in the abovementioned *Ugliola* ruling, an indirect discrimination rule under Article 119 of the Rome Treaty emerged later. *Jenkins* (1981), concerning different payment for part-time and full-time work, is often reckoned as the first application of indirect discrimination under Article 119 of the Rome Treaty.[778] In this case, the Court holds that different hourly pay-rates for the two categories of workers can violate the equal treatment principle if "a considerably smaller percentage of women than of men" qualifies for the higher rates (those offered to full-time workers).[779]

In statutory EU law, indirect discrimination was first introduced when a definition of indirect sex discrimination was included in Directive 97/80/EC (Burden of Proof Directive).[780] Thereafter, a slightly different articulation of indirect discrimination was adopted in the Employment Equality Directive (EED) and the RED in 2000. The GSED followed suit in 2004. When interpreting the direct and indirect discrimination rules in the context of these

---

[776] Judgment of 15 October, 1969, Ugliola, C-15/69, ECLI:EU:C:1969:46; Tobler (2022). When other commentators tend not to mention the Ugliola case in the context of indirect discrimination, this might be because the principle of equal treatment based on nationality is seen as a specific aspect of equal treatment, with a different legal basis than the non-discrimination principle in relation to other protected characteristics. Ever since the foundation of the EEC, there has been a separate statutory expression of the equal treatment principle in relation to nationality (articles 7 and 48 if the Rome Treaty) and the equal treatment principle in relation to sex (Article 119 of the Rome Treaty).

[777] Ugliola, C-15/69, para. 6. Years later, the Court also notes in *Sotgiu* (1974) that "the rules regarding equality of treatment forbid not only overt discrimination by reason of nationality but also all covert forms of discrimination which, by the application of other criteria of differentiation, lead in fact to the same result": Judgment of 12 February 1974, C-152/73, Sotgiu, ECLI:EU:C:1974:13, para. 3.

[778] Judgment of 31 March 1981, C-96/80, Jenkins, ECLI:EU:C:1981:80; Hunter and Shoben (1998) 122; Adams-Prassl, Binns, and Kelly-Lyth (2023) 149.

[779] C-96/80, Jenkins, para. 13.

[780] Council Directive 97/80/EC of 15 December 1997 on the Burden of in Cases of Discrimination Based on Sex (Burden of Proof Directive), Article 2(2). The articulation of indirect discrimination here differs somewhat from how indirect discrimination is defined in the Equality Directives.

Directives, case law from before their adoption is relevant, because the directives are specific expressions of the general non-discrimination principle within the directives' areas of application.[781]

8.4.2 Direct Discrimination: Traditional Starting Points and Progressive Expansion

The traditional area of application of the rule against direct discrimination in EU non-discrimination law is PCPs that explicitly refer to protected characteristics or are motivated by a desire to differentiate between individuals based on their protected characteristics. In contrast, the rule against indirect discrimination is traditionally applied in cases where the disputed PCP does not explicitly rely on protected characteristics and no motivation to discriminate is established.[782]

An exception from this starting point has been introduced by the CJEU in cases concerning discrimination on grounds of pregnancy.[783] In *Dekker* (1990), the Court holds that there is direct discrimination on grounds of sex if an unfavourable treatment is formally based on pregnancy.[784] The decisive reasoning is that "only women can be refused employment on grounds of pregnancy and such a refusal therefore constitutes direct discrimination on grounds of sex."[785] According to the reasoning applied in *Dekker*, a PCP that exclusively disadvantages persons from a protected group is directly discriminatory.[786] However, following *Dekker,* the CJEU has hesitated to turn this into a general rule applicable outside of pregnancy discrimination. In *Schnorbus* (2000), the Court takes an approach to direct

---

[781] Section 6.3.1.

[782] e.g., Kimberly Liu and Colm O'Cinneide, *The Ongoing Evolution of the Case-Law of the Court of Justice of the European Union on Directives 2000/43/EC and 2000/78/EC*, European Commission Directorate-General for Justice and Consumers (Luxembourg: Publications Office of the European Union, November 2019), 8.

[783] Kees Waaldijk and Christa Tobler, "Case C–267/06, Tadao Maruko V. Versorgungsanstalt Der Deutschen Bühnen, Judgment of the Grand Chamber of the Court of Justice of 1 April 2008," *Common Market Law Review* 46, no. 2 (2009): 738.

[784] Judgment of 8 November, 1990, Dekker, C-177/88, ECLI:EU:C:1990:383.

[785] Dekker, C-177/88, 12.

[786] Tobler (2022): 80; Justyna Maliszewska-Nienartowicz, "Direct and Indirect Discrimination in European Union Law–How to Draw a Dividing Line," *International Journal of Social Sciences* 3, no. 1 (2014): 43.

discrimination that aligns more with the traditional starting point, according to which the prohibition on direct discrimination applies only to PCPs that rely on protected characteristics explicitly or as a matter of motivation.[787]

The provisions at issue in *Schnorbus* gave priority to applicants who had completed compulsory military or civilian service, which women were not required to do under the relevant national legislation. In its ruling, the CJEU notes that women therefore could not benefit from the disputed measure.[788] In other words, women were exclusively and entirely negatively impacted by the provisions at issue.[789] The CJEU nonetheless holds that there is no direct discrimination based on sex so long as a PCP does not "apply differently according to the sex of the persons concerned."[790] The ruling therefore confirms that the exception for pregnancy discrimination was, at that time, not applicable in other cases of women being exclusively disadvantaged by a certain practice.

Further expansion of the prohibition on direct discrimination has been developed by the CJEU starting with the *Nikoloudi* ruling (2005). After this ruling, the CJEU's approach to direct discrimination gradually shifts, moving beyond a focus solely on the *form* of a disputed PCP to also consider its *effects*.[791] Nevertheless, the CJEU has remained cryptic in providing guidelines for determining when the rule against direct discrimination is applicable in cases where a protected characteristic is not explicitly relied upon.

---

[787] Judgment of 7 December, 2000, Schnorbus, C-79/99, ECLI:EU:C:2000:676. The view presented herein regarding the interpretation of *Schnorbus* is concurrent with Waaldijk and Tobler: Waaldijk and Tobler (2009) 738. A slightly different interpretation is offered by Eriksson: Andrea Eriksson, "European Court of Justice: Broadening the Scope of European Nondiscrimination Law," *International Journal of Constitutional Law* 7, no. 4 (2009): 743-44, https://doi.org/10.1093/icon/mop025.

[788] Schnorbus, C-79/99, para. 38.

[789] This assumes that all men in a comparable situation would be able to complete the compulsory military service for men, which is a reasonable assumption given that men who could not complete the service because of relevant circumstances such as health-related reasons, would not be in a comparable situation.

[790] Schnorbus, C-79/99, paras. 33-34.

[791] Waaldijk and Tobler (2009) 737-40.

To develop methodological elements of assessing discrimination in an AI-CDS system, it is necessary to understand the distinction between considerations which may indicate direct discrimination in an AI-CDS and considerations which may indicate indirect discrimination. CJEU jurisprudence illuminating the distinction between direct and indirect discrimination is therefore analysed in the following. Section 8.6 connects the general insights established in section 8.5 with specific knowledge and considerations related to AI-CDS systems, thereby developing methodological elements of assessing discrimination in these systems.

8.5 The CJEU Equates Certain Criteria with Protected Characteristics Based on Their Effects

8.5.1 *Nikoloudi* (2005), *Maruko* (2008), etc: Criteria That Exclusively Disadvantage a Protected Group

In *Nikoloudi* the CJEU finds that a practice working to the detriment only of women constitutes direct discrimination, even though the disputed practice does not refer directly to workers' sex.[792] This ruling is based on Directive 76/206/EC on equal treatment between men and women in the working life. At the time, this directive did not explicitly distinguish between direct and indirect discrimination, but such a distinction had been introduced in the CJEU's jurisprudence based on the general non-discrimination principle.

Slightly simplified, the disputed practice in *Nikoloudi* was based on provisions in two collective agreements between Nikoloudi's (the claimant) employer and its federation of workers.[793] The relevant provisions concerned the conditions for being given the position of 'established staff' in the company, a position to which certain benefits were connected. The relevant provisions did not mention the employees' sex and are described by the CJEU as "ostensibly neutral as to the worker's sex."[794] However, the provisions referred to a category of workers (part-time cleaners) which could only consist of women because of the company's recruitment policy – the company only recruited women as part-time cleaners. Given this context, the CJEU holds that the provisions on promotion to 'established staff' constitutes direct discrimination on grounds of sex. It appears that the *exclusivity* of the negative effects

[792] Judgment of 10 March, 2005, Nikoloudi, C-196/02, ECLI:EU:C:2005:141, para 36.

[793] More precisely, the disputed practice arose from a reading of the two collective agreements in conjunction with the company's nationwide staff regulations.

[794] Nikoloudi, C-196/02, paras. 36 and 40.

caused by the disputed PCP is the main reason why it is deemed as constituting direct discrimination. The CJEU notes that, if men are also disadvantaged by the disputed PCP, there is a case of potential *indirect* discrimination.[795]

Retrospectively, the importance of the *Nikoloudi* ruling is that it added another deviation (in addition to the one for pregnancy discrimination, cf. *Dekker)* to the traditional starting point that direct discrimination only applies where a protected characteristic is explicitly relied on or there is a motivation to discriminate. As in *Dekker,* the essential reason for the deviation is a PCP's *exclusive detrimental effects* on women.

While *Nikoloudi* concerns Directive 76/206/EC, the subsequent ruling in *Maruko* (2008) confirms the expansion of the rule against direct discrimination also under the EED. The case concerned a German law that denied a widower's pension to a man whose same-sex life partner had passed away. The claimant, Mr. Maruko, had not been married to his partner, and according to German law, marriage was a necessary requirement for eligibility for a widower's pension. Pursuant to the traditional starting point that direct discrimination applies where a protected characteristic is explicitly relied on, the marriage requirement would be considered an apparently neutral criterion, and not a case of direct discrimination based on sexual orientation. Adhering to the traditional starting point, Advocate General Ruiz-Jarabo notes in his Opinion that "[t]he refusal to award the pension is not based on the sexual orientation of the insured and therefore there is no direct discrimination contrary to Article 2 of Directive 2000/78."[796] In contrast, the CJEU's ruling takes a different stance, holding that such a pension scheme does indeed constitute direct discrimination.[797]

The CJEU's ruling in *Maruko* does not provide a clear and generalizable explanation as to why the facts presented in that case would result in a finding of direct discrimination. However, it can be inferred that the crucial factor is that the widower's pension in question was only accessible to individuals who were married, excluding those who were in registered

---

[795] Judgment (GC) of 1 April, 2008, Maruko, C-267/06, ECLI:EU:C:2008:179, para. 24; Waaldijk and Tobler (2009) 739.

[796] Opinion of AG Ruiz-Jarabo Colomer of 6 September, 2007, Maruko, Case C-267/06, ECLI:EU:C:2007:486, para 96. The AG deems the disputed rule as indirect discrimination, cf. para. 102 of the Opinion.

[797] Maruko, C-267/06, para. 72.

partnerships. Consequently, a reasonable interpretation is that the CJEU applies the same rationale as in the *Nikoloudi* ruling*:* If a PCP *exclusively* disadvantages persons from a protected group, it cannot be considered apparently neutral.

An alternative explanation of why the disputed PCP in the *Maruko* case is seen as direct discrimination could be that it effectively *excluded all comparable persons within a protected group from receiving an advantage.*[798] The marriage requirement in German law resulted in the exclusion of all homosexual individuals in formalised relationships from the widower's pension scheme, which was the relevant advantage at issue).[799] However, it is important to note that this explanation differs from the rationale relied upon in the *Nikoloudi* case. If the Court in the *Maruko* case indeed relies on this alternative rationale, it has the opportunity to explicitly state it.[800] As reverted to below, the CJEU does not explicitly acknowledge the possibility of direct discrimination when a PCP excludes an entire protected group from receiving an advantage, until the *Frédéric Hay* ruling (2013).[801]

A similar factual situation to that in *Maruko* arose again in the case of *Römer,* which also concerned the unequal treatment of individuals in registered partnerships compared to those

---

[798] Upon closer examination, a subtle difference emerges between the facts underlying the questions referred to the CJEU in Maruko and those in Nikoloudi. While both cases concern measures potentially working exclusively to the disadvantage of a protected group, the situation in Maruko differed from Nikoloudi in the following way: In Nikoloudi, the disputed PCP did not disadvantage all comparable persons in a protected group. The practice at issue only disadvantaged women, because all part-time workers were women. However, women were also in the group that was not negatively affected by the PCP (the advantaged group) because there were female full-time workers in the company, and the Court accepted comparison between part-time and full-time workers.

[799] Notably, for the purposes of comparison between groups, the CJEU considers only couples who have decided to formalise their relationship.

[800] Waaldijk and Tobler have argued, based on *Nikoloudi* and *Maruko*, that "direct discrimination now also includes cases where reliance on a formally neutral criterion in fact only affects one protected group, be it by nature or because of a rule that has the force of law": Waaldijk and Tobler (2009) 740. Their reference to 'nature' relates to the case law on pregnancy discrimination. The reference to "a rule that has the force of law" is based on Nikoloudi and Maruko, because in these cases the disputed provisions were given by national laws.

[801] Section 8.5.2.

who were married under German law.[802] Due to the similarities in the facts, the CJEU's ruling in *Römer* relies directly on the *Maruko* ruling without stating the reasoning used to distinguish between direct and indirect discrimination .[803] The same can be said of the circumstances in the case of *Kleist*.[804]

The *Kleist* ruling concerned a national law that allowed an employer to dismiss an employee if the employee was entitled to retirement pension. According to the laws of the Member State, women were generally eligible to retire at the age of 60, while men were not eligible for retirement until the age of 65. Thus, the situation arose where women could be dismissed at a younger age than men. The disputed provisions did not refer explicitly to sex. Yet, they relied on a criterion that exclusively disadvantaged women and placed all women in the relevant comparison group at a disadvantage. Judging by the approach taken in *Nikoloudi, Maruko,* and *Römer,* these circumstances amount to direct discrimination. Thus, the CJEU rightfully notes that the disputed criterion was "inseparable from the workers' sex" and that there is, consequently, "a difference in treatment that is directly based on sex."[805]

8.5.2 *Frédéric Hay* (2013): Criteria That Exclude an Entire Group

The *Frédéric Hay* (2013) case concerned a collective agreement that was beneficial for married employees compared with employees in registered partnerships.[806] This ruling is interesting in the light of the case law considered in previous sections because, this time, the registered partnership was available to different-sex couples as well as same-sex couples under the law of the respective Member State.[807] Marriage, on the other hand, was only available to different-sex couples. Because the collective agreement in this case disadvantaged all employees who were in a registered partnership, and registered partnership was available regardless of sexual orientation, the 'marriage' criterion did not *exclusively* disadvantage same-sex couples. Consequently, the reasoning applied by the CJEU in

---

[802] Judgment (GC) of 10 May, 2011, Römer, C-147/08, ECLI:EU:C:2011:286. The case revolved around provisions which entailed that pensioners who were in a registered partnership received lower retirement pensions than married pensioners.

[803] Römer, C-147/08, para. 52.

[804] Kleist, C-356/09.

[805] Kleist, C-356/09, 31.

[806] Judgment of 12 December, 2013, Frédéric Hay, C-267/12, ECLI:EU:C:2013:823.

[807] Frédéric Hay, C-267/12, para. 43; Liu and O'Cinneide (2019): 56.

Nikoloudi, which appears to have been relied on also in the other cases mentioned in the previous section, would not indicate that the facts in the *Frédéric Hay* case amounted to direct discrimination.[808] The CJEU nonetheless holds:

> The difference in treatment based on the employees' marital status and not expressly on their sexual orientation is still direct discrimination because only persons of different sexes may marry and homosexual employees are therefore unable to meet the condition required for obtaining the benefit claimed.[809]

It appears that the decisive fact in this ruling is the fact that same-sex couples *could not* obtain the benefit, because marriage was a necessary condition for it. The ruling therefore suggests that there is direct discrimination if receiving an advantage is contingent upon a criterion that *excludes an entire protected group*. The *Frédéric Hay* ruling therefore represents an additional expansion of the direct discrimination rule.[810]

8.5.3 Preliminary Summary

The series of CJEU rulings analysed above, arguably indicates that direct discrimination is not restricted to cases where the disputed PCP explicitly refers to a protected characteristic or there is a motivation to discriminate (the traditional starting point). Furthermore, the expansion from the traditional starting point is not limited to discrimination based on pregnancy. In addition to cases where a PCP explicitly refers to a protected characteristic or there is a motivation to discriminate, it can be argued that the rule against direct discrimination in EU law is applicable in cases where:

(i) a PCP works exclusively to the disadvantage of persons from a protected group; or
(ii) a PCP relies on a criterion that would exclude all persons in a protected group from receiving an advantage.

---

[808] In *Frédéric Hay,* the CJEU did not refer to *Nikoloudi.* Rather, the Court notes that the facts of the case differed from the facts of *Maruko* and *Römer*, because those cases concerned measures working exclusively to the disadvantage of a protected group: Frédéric Hay, C-267/12, para. 43.

[809] Frédéric Hay, C-267/12, para. 44.

[810] In a more general manner, Liu and O'Cinneide note that the ruling affirms that direct discrimination "sometimes applies to differential treatment that is not explicitly and overtly based on one of the protected grounds": Liu and O'Cinneide (2019): 55.

8.5.4 CHEZ (2014): Equivalent Factors, Stereotyping, or Prejudice

In the *CHEZ* case, the CJEU was asked about the practice of a provider of electricity meters for consumers. According to the facts of the case, the provider had placed the electricity meters in one specific neighbourhood significantly higher than in other neighbourhoods, thus making it more difficult for consumers in the former neighbourhood to monitor their own electricity usage.[811] The neighbourhood where the meters had been placed higher up, was predominantly inhabited by persons of Roma origin, and the question for the CJEU was whether such a practice constituted discrimination of an ethnic group.

In its ruling, the CJEU instructs the referring court on how to determine whether the practice is "apparently neutral," but does not conclude on the matter in relation to the facts of the case. In its instructions, the CJEU notes that a PCP shall be considered to be apparently neutral where it has "regard to factors different from and not equivalent to the protected characteristic."[812] Thus, if a PCP has regard to factors equivalent to a protected characteristic, the PCP is not apparently neutral. However, what makes a factor 'equivalent' to a protected characteristic? Arguably, this language can be understood in the context of the CJEU's earlier case law on the scope of the direct discrimination rule. Based on the analysis in the preceding sections, one could say that a factor is equivalent to a protected characteristic if it either works *exclusively* to the disadvantage of persons from a protected group or *entirely excludes* a protected group from being advantaged. This understanding implies that whether an equivalent factor is relied on, can be determined by considering the effects of relying on that factor.

Considering the facts referred to the CJEU in *CHEZ,* it appears that those facts did not amount to direct discrimination based on their effects. Like in *Frédéric Hay*, the disputed practice in *CHEZ*, i.e., the placement of electricity meters higher in one neighbourhood compared to other neighbourhoods, did not exclusively disadvantage people with a protected characteristic. It disadvantaged everyone who lived in the concerned neighbourhood.[813]

---

[811] CHEZ, C-83/14, para. 22.

[812] CHEZ, C-83/14, para. 29.

[813] In fact, the claimant in the case was not herself of Roma origin. The Court confirmed that the victim of discrimination does not have to belong to the protected group. This often called

However, in contrast to the situation in *Frédéric Hay,* the disputed practice did not disadvantage *all* individuals in the protected group. Not all persons of Roma origin who fell within the relevant comparison range were disadvantaged by the disputed practice, because not all customers of the electricity company who were of Roma origin resided in the concerned neighbourhood.[814] Therefore, the facts described by the referring court may not be deemed as direct discrimination on the basis of the effects of the disputed practice. The CJEU does not explicitly elaborate on this reasoning in its ruling, as it appears to primarily focus on the possibility of finding direct discrimination based on the presence of a discriminatory motivation or intent. However, when considering the facts presented in *CHEZ* in the light of the lessons derived from previous case law, as summarised in section 8.5.3, it appears that a finding of direct discrimination in *CHEZ* would have to be based on demonstrating that the treatment was explicitly based on ethnicity or motivated by ethnicity.

Because there was no explicit provision or criterion referring to ethnicity in the facts of *CHEZ,* the CJEU carefully instructs the referring court that direct discrimination arises "if it is apparent that a measure which gives rise to a difference in treatment has been introduced for reasons relating to racial or ethnic origin."[815] It is reasonable to infer from this that the CJEU suspects that a discriminatory motive is present. Particularly, it notes that facts indicating that the practice is based on ethnic stereotypes or prejudices are relevant.[816] This indicates that differential treatment based on stereotyping or prejudice may be deemed as direct discrimination. These types of cognitive, human biases are potential sources of equality-related biases in AI-CDS systems, cf. section 4.4.2.6.

8.5.5 "Equivalent," "Inseparable," and "Inextricably Linked"

The CJEU's expansion of the scope of the rule against direct discrimination means that direct discrimination can be found beyond the cases where a PCP refers to protected characteristics or there is a motivation to discriminate, if there is an exclusive disadvantage or a protected group is excluded from receiving an advantage. These are *effects* of a PCP which may lead to

---

'discrimination by association' in academic literature, an aspect that is not further considered in this thesis.

[814] CHEZ, C-83/14, paras. 88 and 90. CHEZ, C-83/14, paras. 88 and 90; CHEZ, C-83/14, para. 88 and 90; CHEZ, C-83/14; CHEZ, C-83/14.

[815] CHEZ, C-83/14, para. 95.

[816] CHEZ, C-83/14, para. 82.

a finding of direct discrimination, even if the PCP does not explicitly refer to protected characteristics. When the use of a certain factor or criterion has such effects, one could say that it is 'equivalent' to a protected characteristic, inspired by the language used by the CJEU in *CHEZ*.[817]

In other cases, most recently joint cases C-804/18 and C-341/19 WABE, the question of whether reliance on a certain factor should be deemed as direct discrimination has been framed as a question of whether that factor is "inextricably linked" to a protected characteristic.[818] Moreover, in *Kleist*, it is noted that the applied factors are "inseparable" from a protected characteristic.[819] These phrases are used in different contexts in CJEU case law. Sometimes, it appears as if they are used to denote a causation requirement.[820] Indeed, as chapter 10 returns to, there is a close connection between the causation requirement and the distinction between direct and indirect discrimination. However, leaving the causation requirement aside for now, it is noteworthy that the notions of 'equivalence,' 'an inextricable link,' and 'inseparability' arguably refer to the same core issue: the issue of determining whether one is dealing with a PCP that constitutes potential direct discrimination, rather than being 'apparently neutral.'

Throughout the remainder of this thesis, the phrase 'inextricably linked' is used to denote that the use of a certain factor potentially leads to direct discrimination even though the factor is not defined as a protected characteristic. Due to this thesis's focus on sex and ethnicity as protected characteristics, particular attention is given to the possibility that AI-CDS systems might include feature variables that are inextricably linked to sex or ethnicity.

---

[817] CHEZ, C-83/14, para. 109.

[818] Judgment (GC) of 15 July, 2021, WABE, Joined Cases C-804/18 and C-341/19, ECLI:EU:C:2021:594, paras. 52-53 and 73. Earlier case law mostly concerns age discrimination: Judgment (GC) of 12 October, 2010, Ingeniørforeningen I Danmark, C-499/08, ECLI:EU:C:2010:600; Judgment of 7 June, 2012, Tyrolean Airways, C-131/11, ECLI:EU:C:2012:329; Judgment of 14 February, 2019, Horgan and Keegan, C-154/18, ECLI:EU:C:2019:113; Judgment (GC) of 26 January, 2021, Szpital Kliniczny, C-16/19, ECLI:EU:C:2021:64.

[819] Kleist, C-356/09.

[820] This applies, e.g., to the ruling in WABE, Joined Cases C-804/18 and C-341/19.

8.5.6 Concluding Discussion

The purpose of the case law analysis in the previous sections was to establish an understanding of the distinction between direct and indirect discrimination in EU law (as stated in section 8.4.2). This knowledge serves the objective of the thesis because the methodological elements of assessing discrimination in a pre-deployment setting should include considerations which may indicate direct discrimination and considerations which may indicate indirect discrimination.

The case law analysis has suggested that the prohibition on direct discrimination is applicable beyond cases where a PCP explicitly relies on protected characteristics or where there is a motive to discriminate. The following table categorises the facts referred to the CJEU in the cases analysed above and indicates on what basis the CJEU appears to find direct discrimination in each case. None of the cases concerned PCPs that relied explicitly on a protected characteristic:

Case	Excludes all comparable persons in a protected group	Exclusively disadvantages a protected group	Involves intent, motivation, stereotyping, or prejudice
Nikoloudi	-	**X**	-
Maruko	X	**X**	-
Römer	X	**X**	-
C-356-09 Kleist	X	**X**	-
Frédéric Hay	**X**	-	-
CHEZ	-	-	Possibly (for the referring court to determine)

Table 1: The table indicates why the disputed PCP might be seen as directly discriminatory. A bold, underlined "X" indicates the rationale that the CJEU appears to have relied on.

Based on this understanding of the CJEU's jurisprudence, direct discrimination occurs where a PCP exclusively disadvantages a protected group or entirely excludes a protected group receiving an advantage. In other words, direct discrimination can in these circumstances be found on the basis of the effects of a PCP on a protected group. When such effects arise, one should assume that the PCP relies on factors or criteria that are *inextricably linked* to a protected characteristic. However, there might also be additional arguments worth considering when determining whether a certain factor should be considered inextricably linked with a protected characteristic. Arguably, there may be factors or criteria that are so intertwined with a protected characteristic that they are practically inseparable and thus inextricably linked according to a qualitative, common-sense judgment. For example, academic literature recognises that a person's skin colour is probably a feature that must be considered inextricably linked with ethnicity.[821]

In addition to the cases where it is argued that there is an inextricable link between a protected characteristic and a different, yet related factor, the table above indicates that direct discrimination may arise where intent, motivation, prejudice or stereotyping is involved. This thesis does not consider instances where AI developers or users intend to discriminate. However, the notion that the involvement of prejudice and stereotyping in a decision-making process may indicate direct discrimination, is kept in mind throughout the remainder of the thesis.

[821] Ellis and Watson (2012) 167.

8.6 Implications for Pre-Deployment Discrimination Assessment of an AI-CDS System

8.6.1 Overarching Methodological Elements

This chapter has considered several CJEU rulings for the purpose of establishing an understanding of the distinction between direct and indirect discrimination in EU law. Due to the absence of methodological clarity in the CJEU's approach to this distinction, this chapter has reasoned about different rationales that might explain the results that the CJEU has arrived at. The first methodological elements of assessing discrimination in an AI-CDS system before deployment are developed based on the present chapter.

Methodological Elements

- □ Examine whether the model includes PILFs as feature variables (see section 8.6.2)?
  - ➢ Yes =potential direct discrimination
  - ➢ No = apparently neutral, unless bias is caused by stereotyping/prejudice
- ? Is bias caused by stereotyping/prejudice (see section 10.4.7)?
  - ➢ Yes = potential direct discrimination
  - ➢ No = apparently neutral

The first step of the assessment should be to examine whether a model utilised by an AI-CDS system incorporates any protected characteristics or factors that are inextricably linked to such characteristics. For the sake of convenience, the term 'Protected or Inextricably Linked Factor' will be abbreviated as 'PILF' throughout the remainder of this thesis. If PILFs are present in a model, there is a potential for direct discrimination. The criteria used to determine whether a feature variable should be considered a PILF are discussed in the subsequent section.

If a model does not include PILFs, there is one other relevant consideration which may indicate whether the model could lead to direct discrimination: It should be considered whether biases in the model are caused by stereotyping or prejudice. This chapter has posited that if bias in an AI-CDS system arises from stereotyping or prejudice, direct discrimination may occur. To further assess the likelihood of a model leading to direct discrimination, one must establish whether there is a sufficient causal link between the stereotyping/prejudice and the outputs of the model. Thus, the consideration of stereotyping or prejudice as a source of

bias falls within the purview of the direct causation assessment, the methodological elements of which are developed in chapter 10.[822]

8.6.2 Criteria Determining Whether a Model Includes PILFs as Feature Variables

The case law analysed in this chapter supports that a certain factor is inextricably linked to a protected characteristic if it has one of these two effects when applied in a PCP:

(i) it works exclusively to the disadvantage of a protected group;

(ii) it entirely excludes a protected group from an advantage.

In addition, it has been suggested that an inextricable link can be found based on a qualitative, common-sense assessment according to which a factor is so intertwined with a protected characteristic that they are practically inseparable. The example of a person's skin colour being inextricably linked with ethnicity was mentioned in section 8.5.6. As an extension of this argument, one could argue that the melanin level in a person's skin could be inextricably linked with ethnicity.

Methodological Elements: Feature Identification

? Are protected characteristics among the feature variables?
? Might the model exclusively disadvantage a protected group?
? Might the model entirely exclude a protected group from receiving an advantage?
? Are there feature variables which are otherwise inseparable from a protected characteristic?

➢ If at least one of the above questions is answered in the affirmative, the assessment should proceed to consider whether there is direct causation between a PILF and a disadvantageous treatment (see chapters 9-10).

Consequently, four questions should be addressed in order to determine if a model relies on PILFs. The first question pertains to the inclusion of protected characteristics among a

[822] The view that the influence of prejudice is an indication of direct discrimination is supported in academic literature. Henrard argues that when "ingrained prejudice" against a particular ethnic group is at the root of disadvantageous treatment, the treatment be seen as occurring "on grounds of" ethnicity, even if ethnicity is not explicitly relied on. Henrard emphasises that ethnicity plays a *causal* role in these cases and that it influences decision-making in a manner that goes beyond mere "effect" and lies "closer to intent": Henrard (2019) 108.

model's feature variables.[823] The remaining three questions aim to ascertain whether a model incorporates factors that are inextricably linked to protected characteristics. This part of a discrimination assessment, which aims to ascertain whether a model relies on PILFs, can be referred to as 'Feature Identification,' because the objective is to determine whether certain features are incorporated in a model. The following sections discuss the implications of the criteria that determine whether a PILF should be deemed to be present in a model and develops further considerations for the Feature Identification process.

8.6.3 Models Excluding an Entire Protected Group from an Advantage

An AI-CDS system that *excludes* an entire protected group, i.e., a system that makes it impossible for persons with a certain characteristic to receive an advantageous output, is not very likely to occur. Such an exclusive detrimental effect would arise, for example, if a diagnostic model for some reason misdiagnoses all female patients. It does not seem likely that such a drastically underperforming model would be used or brought to the market. It is nonetheless worth checking, in a pre-deployment discrimination assessment, whether there are any protected groups that only receive disadvantageous outputs. In practice, the criterion that a model cannot exclude an entire group from advantageous outputs could be implemented as one of several performance criteria to consider during pre-deployment testing and scrutinization of an AI-CDS system's behaviour.

8.6.4 Models Exclusively Disadvantaging a Protected Group

Might there, on the other hand, be cases where a model *exclusively disadvantages* a protected group without relying directly on a protected characteristic? It is possible, in theory, that a disadvantage occurs only for women or only for a particular ethnic group (or ethnic minorities as a group).[824] However, while PCPs exclusively disadvantaging a protected group have been an issue in several cases considered by the CJEU, models employed in AI-CDS systems are significantly more complex. Unlike the PCPs considered by the CJEU, which may

---

[823] Adams-Prassl, Binns and Kelly-Lyth note that a significant number of AI systems in different sectors have been found to rely directly on protected characteristics: Adams-Prassl, Binns, and Kelly-Lyth (2023) 157.

[824] Whether EU non-discrimination law requires that there must be one particular ethnic group being disadvantaged is discussed in section 9.7.

exclusively disadvantage a particular protected group, the disadvantages related to AI-CDS systems are not likely to impact only individuals with a specific characteristic.

AI-CDS systems rely on a multitude of feature variables to generate personalised outputs for individual patients. These outputs are not based on binary categorisations, such as distinguishing between part-time workers and full time workers,[825] married persons and registered life partners,[826] or individuals who can be dismissed at the age of 60 versus those who cannot be dismissed until the age of 65.[827] Instead, AI-CDS systems utilize a much more intricate network of variables, reflecting the fundamental concept of personalized medicine, which seeks to tailor medical treatment to the unique characteristics of each patient.

Due to the many feature variables and target variables that may be involved, the group of persons who receive disadvantageous outputs from an AI-CDS system are not likely to be limited to persons with one particular protected characteristic. However, if, by some rare occurrence, test data reveals an exclusive disadvantage, this should be taken as a strong indication that direct discrimination may occur if the model is deployed.

8.6.5 Feature Identification in Opaque Models – Hidden Inferences[828]

Certain AI models may have the ability to be easily interpreted, allowing assessors to observe the included feature variables within the model.[829] However, due to the 'black box' problem,[830] which is particularly associated with complex neural networks based on deep learning,[831] there exist models which may not be as readily interpretable, making it difficult to

---

[825] e.g., Nikoloudi, C-196/02.

[826] e.g., Maruko, C-267/06; Römer, C-147/08; Frédéric Hay, C-267/12.

[827] Kleist, C-356/09.

[828] See section 4.4.4.

[829] Been Kim et al., "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (Tcav)" (paper presented at the International conference on machine learning, 2018), 2; Jane R Bambauer, Tal Zarsky, and Jonathan Mayer, "When a Small Change Makes a Big Difference: Algorithmic Fairness among Similar Individuals," *UC Davis Law Review* 55, no. 4 (2021): 2414. ("In the most straightforward scenario, an algorithmic decision-making system uses an interpretable model, where an analyst can directly inspect the model and understand its behavior.")

[830] Section 1.6.2.

[831] Section 1.5.4.

identify the feature variables. Typically, such models are designed to handle complex and unstructured data, and may even process data of different types at once, such as medical images and clinical notes in free text format.

Importantly, even if PILFs are removed from the training data and not used as input data, it is still possible that a model might infer PILFs from a set of correlated features that are observable in the training data and input data.[832] If this occurs, the model could end up relying directly on PILFs as feature variables, without the knowledge of the developers.[833] This would arguably imply a potential for direct discrimination. While this potential for direct discrimination is alluded to by Adams-Prassl, Binns and Kelly-Lyth,[834] it is often underestimated in legal scholarship on algorithmic discrimination.[835] However, in technical ML literature, especially relating to deep neural networks, it is recognised that hidden inferences may occur which correspond to the direct reliance on a factor that is not observable in training data: as noted by Bau et al:

> Observations of hidden units in large deep neural networks have revealed that human-interpretable concepts sometimes emerge as individual latent variables within those networks (…).[836]

What happens, in practice, when such hidden or 'latent' variables occur, is that a neural network creates an internal representation that corresponds to a certain concept.[837] While occurrence of these hidden or 'latent' variables have predominantly been studied in neural networks trained on images or graph-structured data, more recent research emphasises the

---

[832] Banerjee et al. (2021); Hauglid (2022). Paulus and Kent (2020) 6.

[833] Section 4.4.4; Strümke, Slavkovik, and Stachl (2023) 11; Adams-Prassl, Binns, and Kelly-Lyth (2023) 160.

[834] Adams-Prassl, Binns, and Kelly-Lyth (2023) 159. (noting that learning algorithms may create proxies that are indissociable from protected characteristics.)

[835] Section 8; Hacker (2018) 1151; Tischbirek (2020) 114. Wachter, Mittelstadt, and Russell (2021 B) 19-20; Gerards and Xenidis (2021): 67.

[836] David Bau et al., "Network Dissection: Quantifying Interpretability of Deep Visual Representations," *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017): 6541.

[837] Strümke, Slavkovik, and Stachl (2023) 1.

possibility of neural networks learning such variables also from tabular data.[838] This is particularly relevant to consider in a discrimination assessment, because tabular data collected in a medical context may contain several pieces of information which are correlated, to varying degrees, with protected characteristics.[839] The presence of such information increases the likelihood that neural networks may create internal representations of PILFs. The ML literature considered for this chapter does not specifically discuss the extent to which PILFs may be inferred and used as latent variables in neural networks. However, a contribution from Strümke, Slavkovik and Stachl suggests that ML algorithms have the capability to infer various vulnerable states.[840] For example, they consider 'feeling sad or down' as a vulnerable state that may become a latent variable. If hidden inference of such concepts are possible, it is conceivable that PILFs such as sex or ethnicity could also become latent variables.

Hidden inferences may have played a role in the case of Simon Tesfay, as depicted in the fictional case study provided in section 5.1. It is plausible that the AI-CDS system employed by Simon's hospital, UHS, included a variable that correlated with ethnic minority status, and that it associated such status with lower utilization of health services. This correlation could potentially explain why the system is less likely to predict that ethnic minority patients will surpass the threshold for being offered the Preventive Care Program, which is based on predicted future health needs.[841] Consequently, it would appear that all ethnic minority patients have a lower probability of being enrolled in the Preventive Care Program, compared to similarly situated patients from the ethnic majority.[842] This lower probability does not in itself constitute direct discrimination, cf. section 8.5.6, but there may be direct discrimination if there is direct causation between this lower probability and the ethnic minority status. The

---

[838] Mateo Espinosa Zarlenga et al., "Tabcbm: Concept-Based Interpretable Neural Networks for Tabular Data," *Transactions on Machine Learning Research* 7 (2023): 3.

[839] Section 4.4.2.1.

[840] Strümke, Slavkovik, and Stachl (2023).

[841] Section 5.1.

[842] The *lower probability of being selected* is arguably a disadvantage in itself: Fredman thus criticises the *Essop* judgment in the UK because the court seems to consider as disadvantaged only those who experienced the concrete negative consequences of a measure that placed all BME candidates at heightened risk of failure: Sandra Fredman, "The Reason Why: Unravelling Indirect Discrimination," *Industrial Law Journal* 45, no. 2 (2016): 238.

assessment of causation is addressed in chapter 10. For now, it is crucial to underscore the importance of considering the possibility of hidden inferences during the Feature Identification process proposed in this thesis.

Methodological elements: feature identification in opaque models

The following applies to models where feature variables are not observable to the assessor.

? Consider the possibility that PILFs might be relied on through hidden inferences.

□ Relevant technical methods:
- Examine input data for information from which sex or ethnicity might be inferred
- Methods for studying model behaviour
- Testing with Concept Activation Vectors (for neural networks)

The importance of searching for hidden inferences highlights the need for development of technical methods that can identify feature variables in opaque models. Strümke, Slavkovik and Stachl suggest that the training/input data should be carefully examined to identify information from which these vulnerable states can potentially be inferred.[843] Furthermore, in addition to examining the training/input data, they argue that it might be feasible to identify whether vulnerable states are relied on by studying a model's behaviour during pre-deployment testing.[844] While the authors refer to various states they define as 'vulnerable,' the same methods could be applied to identify PILFs within an AI-CDS system. By scrutinising the correlation between a model's outputs and the sex and ethnicity of patients (provided that this information is available during testing), it may be feasible to determine whether it is more likely than not that the model relies on these characteristics.[845]

A more specific technical method to consider is to "probe the internal state of a neural network for concepts," recognising the possibility that a neural network may create internal representations corresponding to PILFs.[846] For instance, Kim et al. propose a technique called 'Testing with Concept Activation Vectors' (TCAV) as a potential way of estimating whether

[843] Strümke, Slavkovik, and Stachl (2023) 12.

[844] Ibid.

[845] Strümke, Slavkovik, and Stachl (2023) 12-13.

[846] Strümke, Slavkovik, and Stachl (2023) 14.

a given concept is influential on the outputs produced by a neural network.[847] In theory, one could define a protected characteristic such as 'sex' or 'ethnicity' as a concept and apply TCAV to determine whether these characteristics have been inferred and turned into latent variables in a neural network. This avenue is worth exploring through further interdisciplinary research.

8.7 Conclusion

In the context of a pre-deployment discrimination assessment, one could say that 'apparently neutral' refers to an AI-CDS system that is not likely to cause direct discrimination.[848] It is arguably in line with the approach taken by the CJEU if such an assessment concentrates, first, on assessing direct discrimination, before turning to potential indirect discrimination.[849] Feature Identification has been proposed as the first step of a pre-deployment discrimination assessment, cf. section 8.6.1. Feature Identification is the process through which one assesses the possibility that a model relies on PILFs as feature variables. Given that probability is central to the types of pre-deployment assessments found in the relevant legal framework,[850]

---

[847] Kim et al. (2018).

[848] AG Jäaskinen's opinion in *Römer* suggests that the correct methodological approach is to ask first whether there is direct discrimination, and then consider indirect discrimination only if the first question is answered in the negative: Opinion of Advocate General Jäaskinen of 15 July, 2010, Römer, C-147/08, ECLI:EU:C:2010:425, para 100-01. Henrard also finds that the method of the Court in *CHEZ* entails that direct discrimination is checked before indirect discrimination: Henrard (2019) 114.

[849] I also interpret Tobler's thorough analysis of the indirect discrimination rule to be in line with this view, as Tobler defines an apparently neutral PCP as "[a] measure that is not directly or inextricably linked to a protected ground": Tobler (2022): 59. However, in my view, the most appropriate definition of an "apparently neutral" PCP is a measure that does not treat someone unfavourably based on a PILF. In other words, an apparently neutral PCP is a PCP that is not directly discriminatory. If a PCP is linked to a PILF in a way that does not treat a person less favourably than another, the PCP maintains its apparent neutrality. This view is arguably supported by the AG's Opinion in Römer: Römer, C-147/08, para. 101. ("The question of the interpretation of Article 2(2)(b) of Directive 2000/78, relating to the concept of indirect discrimination, is raised only if it is found that there has not been direct discrimination, either at the end of the examination of comparability of the situations carried out by the Court itself if it considers it is able to do so, as the Com- mission suggests, or after the analysis of this kind which will be left for the national court to carry out.")

[850] Section 7.6.2.

Feature Identification may consider the likelihood that PILFs are incorporated in a model. Important to this consideration, this chapter has developed the criteria for determining whether a given factor is inextricably linked to a protected characteristic. Such an inextricable link is established if a model either excludes an entire protected group from receiving an advantage or exclusively disadvantages a protected group, or if a factor must be deemed inseparable from a protected characteristic based on a common-sense assessment.

The methodological elements developed in this chapter for Feature Identification in opaque models are particularly important in relation to the potential for hidden inferences in neural networks. This chapter has elucidated the fact that hidden inferences, which may occur unintentionally and unbeknownst to the developers, can lead to direct discrimination. ML literature on the creation of latent variables in neural networks was referred to, suggesting that the occurrence of direct discrimination in AI systems may be more prevalent than commonly assumed in the existing legal scholarship on algorithmic discrimination.[851] Relevant technical methods for scrutinising potential hidden inferences were identified in technical literature. As proposed in this chapter, future interdisciplinary research should further explore the feasibility of applying these methods in the context of a pre-deployment discrimination assessment based on the non-discrimination principle.

In addition, this chapter has posited that the finding of biases stemming from stereotyping or prejudice would be an indication of potential direct discrimination. This is, therefore, an aspect that should be addressed as part of the direct discrimination assessment. This particular aspect of direct discrimination assessment is further discussed in chapter 10.

The following methodological elements of assessing discrimination in an AI-CDS system before its deployment have been developed in this chapter:

[851] Contrary to Hacker's argument that, in most cases in which bias is an accidental feature of the data processing, unfavorable treatment does not occur "on grounds of" group membership and therefore does not amount to direct discrimination: Hacker (2018) 1152.

Feature Identification

Objective: Examine whether PILFs are included as feature variables in a model

Considerations for highly interpretable models

- ? Are protected characteristics included?
- ? Are inextricably linked factors included?
  - Disadvantage exclusively affecting a protected group;
    - ➢ Technical method: Testing of model behaviour
  - Exclusion of entire protected group from being advantaged;
    - ➢ Technical method: Testing of model behaviour
  - Factors otherwise inseparable from sex or ethnicity.
    - ➢ Method: common-sense assessment

Considerations for opaque models

- ? Consider the possibility that PILFs might be relied on through hidden inferences
  - Technical methods:
    - ➢ Examine input data for information from which sex or ethnicity might be inferred
    - ➢ Testing of model behaviour
    - ➢ Testing with Concept Activation Vectors (for neural networks)

No PILFs = apparently neutral model (unless bias is caused by stereotyping or prejudice, cf. chapter 10)

9 Disadvantage and Comparison

9.1 Introduction

AI-CDS systems are developed and deployed to improve clinical decisions to the benefit of patients, clinicians, and society at large. However, the benefits of AI-CDS systems may not be equally distributed across groups defined by protected characteristics such as ethnicity or sex. Moreover, there might be instances where an AI-CDS system treats one person less favourably than another based directly on a protected characteristic.[852] EU non-discrimination law entails certain thresholds for how such comparative disadvantages may be distributed across protected groups, with the implication that comparative disadvantages exceeding the law's thresholds potentially constitute discrimination. This chapter examines those thresholds, even though they are not clearly defined in the law. Furthermore, it explores how the extent of comparative disadvantage may be measured. Lessons from CJEU jurisprudence on comparison and disadvantage measurement in ex post enforcement contexts are adapted and developed into considerations, principles, criteria, and methods to be included in a pre-deployment assessment discrimination assessment of an AI-CDS system.[853]

This chapter unfolds as follows: section 9.2 explains the notion of a 'disadvantage' in EU non-discrimination law. Section 9.3 identifies the most salient disadvantages that patients might encounter when biased AI-CDS systems are used, building on the knowledge from chapters 3 and 4 about how equality-related biases manifest in AI-CDS systems. Section 9.4 then details the distinct disadvantage requirements that pertain to direct and indirect discrimination, respectively. In section 9.5, criteria for determining comparability in the context of clinical decisions are developed based on an analysis of the comparability requirement in EU non-discrimination law. Such criteria are needed in order to measure the extent of disadvantage, a problem that is further discussed in section 9.6. Given this thesis's focus on the pre-deployment discrimination assessment context, the development of criteria for comparison and disadvantage measurement concentrates on this context. Finally, section 9.7 discusses an important controversy in EU law related to measuring disadvantages

[852] As discussed in chapter 8 and section 10.4.

[853] Regarding the method of developing these methodological elements, see chapter 2.

experienced by groups defined by ethnicity, before section 8.8 summarises the chapter's key findings and summarises the methodological elements developed in this chapter.

9.2 The notion of a 'Disadvantage' in EU Non-Discrimination law

In a general, commonplace understanding, the word 'disadvantage' can denote any burden or loss experienced by an individual or group, either in absolute terms or in comparison to others. In this general context, being 'disadvantaged' does not necessarily imply a worsening of one's prior condition. It could denote a missed opportunity for improvement, such as the failure to secure a promotion or university admission - a disadvantage in absolute terms.[854] *In comparative terms,* a disadvantage arises when a person or a group finds themselves in an inferior position relative to others in respect of a specific value, resource, or metric. This could occur if one person incurs an absolute disadvantage that another does not, or if one person is treated more favourably than another without the latter experiencing a disadvantage in absolute terms.

There are many types of disadvantages that can lead to a finding of discrimination. In CJEU case law, statements about the relevance of various disadvantages are rare. To the best of my knowledge, among cases where the CJEU concludes that the facts presented by the referring court does not constitute discrimination, there are no cases where this conclusion is grounded in the absence of a disadvantage. Rather, it is more common that claimants who are disadvantaged are found not to be in a comparable situation to those who are better off, or it is determined that the disadvantage does not affect a protected group to such an extent that they are "particularly disadvantaged" as required in relation to indirect discrimination. For instance, in the *Maniero* case, the CJEU does not dispute that there was a disadvantaged group and an advantaged group.[855] However, it holds that the disadvantage was not particular to an ethnic group,[856] mirroring its approach in the earlier case of *Jyske Finans*.[857] In the latter case, a disadvantage was clearly experienced by persons who were required to produce

---

[854] Hellborg (2018) 242.

[855] The CJEU notes that there is an advantaged group (those who satisfy the requirement of having successfully completed the First State Law Examination, and a disadvantaged group (all persons who do not satisfy that requirement): Maniero, C-457/17, para. 49.

[856] Maniero, C-457/17, para. 50.

[857] Jyske Finans, C-668/15.

additional proof of identification where other persons could simply rely on their driver's license. The CJEU finds that there was no ethnic discrimination because the criterion was applicable to all persons born outside of the EU (implicating that because a wide variety of ethnicities were affected, the criterion was not inextricably linked to ethnicity).[858]

Examples of disadvantages that have either been explicitly recognised as such or, at least, not been questioned as such by the CJEU include being refused as a blood donor,[859] being subjected to a security check,[860] having more difficult access to electricity metres,[861] not being eligible for an education scholarship,[862] being discouraged from[863] or excluded from applying for a job or position,[864] being dismissed from a job or position,[865] having to present additional proof of identification in order to access services,[866] the partial withholding of occupational pension payments,[867] not being granted an allowance, pension, etc.,[868] and being placed lower than others on a scale that determines salary or remuneration.[869] These examples primarily concern disadvantages of a rather direct and tangible nature. They involve situations where individuals are denied certain resources, burdens are imposed upon them, or their opportunities to receive certain benefits are less compared to others.

---

[858] Jyske Finans, C-668/15, paras. 20-21.

[859] Léger, C-528/13.

[860] Judgment (GC) of 15 April, 2021, Braathens Regional Aviation, C-30/19, ECLI:EU:C:2021:269.

[861] CHEZ, C-83/14.

[862] Maniero, C-457/17.

[863] Feryn, C-54/07; Judgment of 25 April, 2013, Asociaţia Accept, C-81/12, ECLI:EU:C:2013:275; Judgment (GC) of 23 April, 2020, Rete Lenford, C-507/18, ECLI:EU:C:2020:289.

[864] Judgment of 2 April, 2020, Comune Di Gesturi, C-670/18, ECLI:EU:C:2020:272.

[865] e.g., Judgment (GC) of 17 July, 2008, Coleman, C-303/06, ECLI:EU:C:2008:415.

[866] Jyske Finans, C-668/15.

[867] YS, C-223/19.

[868] Szpital Kliniczny, C-16/19; Maruko, C-267/06.

[869] Horgan and Keegan, C-154/18; Judgment of 7 February, 2019, Escribando Vindel, C-49/18, ECLI:EU:C:2019:106. (In *Escribando Vindel,* the CJEU instructs the referring court of the importance of examining whether the members of the disadvantaged group belonged to a particular age group that was different from the age group that was advantaged by the disputed measure.)

Relevant disadvantages in EU non-discrimination law can also encapsulate more intangible, 'representational' harms,[870] such as the stigmatising effect of a measure. More tangible disadvantages probably appear more often in litigation, given that victims are more likely to pursue them.[871] However, the CJEU holds in *CHEZ* that the RED precludes a national law requiring that the relevant harm must consist of prejudice to rights or legitimate interests for a measure to violate the non-discrimination principle.[872] This ruling suggests that a disadvantage does not have to infringe upon a person's rights or even their legitimate interests to be relevant under EU non-discrimination law. In *CHEZ,* the CJEU considers both the difficulty of checking the electricity metres and the stigmatising effect of the practice of placing them higher in a certain neighbourhood compared to other neighbourhoods.[873]

In academic literature, there is a rich variety of explanations of the 'disadvantages' or 'harms' of inequality and discrimination.[874] They typically include tangible/allocational disadvantages related to how resources, positions and opportunities are distributed, as well as more representational types of harm. Indeed, the notion of a disadvantage in EU non-discrimination law has a broad interpretation.

A crucial issue in practice is often to determine whether the situation of a person or group experiencing a disadvantage is sufficiently comparable to that of those who are at an advantage. This assessment of comparability, and how to conduct it in a pre-deployment context, is discussed in section 9.5. A prerequisite for an appropriate comparison and disadvantage measurement, however, is the definition of what it means to be advantaged and disadvantaged in relation to a specific decision or practice. The following section summarises the disadvantages that are most likely to be inflicted by AI-CDS systems, with reference to

---

[870] Section 4.2.

[871] Nilsson notes in relation to ECtHR jurisprudence that the ECtHR's discussion of harms caused by a disputed measure usually concentrates on the claimant's ability to enjoy a right. However, the *Konstantin Markin* case is mentioned as an example of a case where the ECtHR emphasised stereotyping as a harmful effect: Nilsson (2020) 139.

[872] CHEZ, C-83/14, para. 129; Liu and O'Cinneide note that the CJEU applies a broad interpretation of the less favourable treatment requirement in the *CHEZ* ruling: Liu and O'Cinneide (2019): 55.

[873] Henrard (2019) 113.

[874] e.g., Moreau (2004); Khaitan (2015); Fredman (2016 B); Nilsson (2020) 138.

the discussion of equality-related biases and their sources in chapter 4. When the following section lists three categories of disadvantages, this does not mean that AI-CDS systems cannot cause other disadvantages. The three categories are selected in order to concentrate the analysis around the most relevant disadvantages that should be considered as part of an assessment of discrimination in an AI-CDS system.

9.3 Disadvantages of Biased AI-CDS Systems

9.3.1 Disparate Performance

The concern about inequality in the standard of care offered to patients from different groups relates to the technical issue of *disparate performance*, i.e., the situation where an AI-CDS system does not perform equally well for patients from different groups.[875] For classification tasks (e.g., diagnosis), performance in AI models is usually measured in terms of predictive accuracy, a measure that concentrates on error rates. If a patient receives an erroneous clinical assessment, this is a relevant disadvantage. In relation to diagnosis-setting, for instance, a disadvantage is incurred if a clinical assessment leads to a false positive or false negative conclusion.

In the case of a false negative diagnosis, there is an imminent risk that the patient will not receive correct or timely treatment. A false positive diagnosis, on the other hand, can lead to an unnecessary intervention. Practically all medical interventions are associated with risk.[876] Being exposed to this risk, and the burden of undergoing treatment, is a disadvantage if there is no proper medical justification for it. To understand how the extent of this disadvantage can be measured, section 9.6 takes a closer look at how disparate performance can be measured in AI-CDS systems. For now, it suffices to note that receiving an inaccurate output is often a disadvantage (in absolute terms) in the context of healthcare and, thus, receiving less accurate outputs than others (disparate performance) is a disadvantage in comparative terms.

When disparate performance occurs in an AI-CDS system involved in determining the allocation of scarce resources,[877] it may be more fitting to say that disparate performance relates to the *access* to care rather than the *standard* of care. These concerns are intertwined.

---

[875] Section 4.3.3.

[876] Kåre I. Birkeland et al., *Indremedisin : 1*, vol. 1 (Drammen: Vett & Viten, 2017), 184.

[877] Section 1.7.5.

Both types of impact (i.e., on access to care and standard of care), are disadvantages capable of amounting to discrimination, and they are associated with disparate performance of AI-CDS systems. Therefore, it is important to consider disparate performance between protected groups when assessing discrimination in an AI-CDS system.

9.3.2 Resource Denial (Regardless of Performance)

Importantly, a disadvantage may be incurred by a patient even if an AI-CDS system does not generate a less accurate output for that patient compared to others. If a system causes a benefit to be denied for a patient, this is a disadvantage regardless of the system's accuracy. This is particularly important in the context of decisions concerning the allocation of scarce resources. For example, a high-performing AI-CDS system might suggest that resources should not be allocated to a patient because the patient does not meet a minimum threshold or because another patient is considered more in need of those resources. The resource denial is, as such, a disadvantage even if the AI-CDS system is performing as expected.

9.3.3 Stigma and Prejudice

In chapter 4, stereotyping and prejudice were discussed as potential *sources* of bias in AI-CDS systems. The present chapter explores the disadvantage requirement in EU non-discrimination law. In this context is important to recognise that prejudice and stigma can be considered disadvantages potentially *resulting from* the use of an AI-CDS system.[878] Sources of bias that may result in stigma and prejudice have already been discussed in chapter 4 and are not reiterated here.

It is recognised in CJEU case law that stigmatisation is a relevant disadvantage under EU non-discrimination law.[879] If an AI-CDS system perpetuates stereotypes or prejudices, it arguably contributes to stigma. Thus, it effectively reproduces a relevant disadvantage. Stigma and prejudice may arise in addition to more tangible disadvantages (allocational harms), or they may constitute disadvantages in themselves (representational harms).[880]

---

[878] The stigma or prejudice that results from the use of an AI-CDS system may be caused by pre-existing stigma or prejudice, or it may be introduced during the development process.

[879] CHEZ, C-83/14, para. 87.

[880] Section 4.2.

While stereotyping refers to a generalisation that implies assumptions about an individual or group,[881] stigma refers to the metaphorical 'mark' that is imprinted on persons associated with a particular group.[882] In the context of discrimination, stigma is an undeserved attribute that is attached to the persons being discriminated against.[883] Prejudice can be understood, as per section 3.2.1, as any attitude, emotion or behaviour towards a group or member of a group that creates or maintains hierarchical status relations between groups.

9.4 Disadvantage Requirements for Direct and Indirect Discrimination

9.4.1 "Less Favourable Treatment"

The finding of a disadvantage is a prerequisite both for direct and indirect discrimination, although there are certain differences between the disadvantage requirements under the two rules, according to the Equality Directives. For direct discrimination to arise, the Equality Directives do not require that a person is disadvantaged in absolute terms. The definition of direct discrimination only requires that a person is treated "less favourably" than another person is, has been, or would be treated in a comparable situation. This implies that a person can be subject to direct discrimination even if the discriminatory treatment improves a person's situation, as long as it improves another person's situation more significantly. For instance, a person who receives a 3 % salary raise is treated less favourably than a person who receives a 10 % salary raise. Similarly, although the deployment of an AI-CDS system improves the services provided to one individual, the system might benefit another individual even more. If such a comparative disadvantage occurs "on grounds of" (i.e., a causation requirement) a protected characteristic, there could be direct discrimination. The causation requirement is analysed in chapter 10, while the present chapter concentrates on the disadvantage requirement.

---

[881] Section 3.2.1.

[882] Stacey Hannem and Chris Bruckert, "Introduction," in *Stigma Revisited: Implications of the Mark*, ed. Stacey Hannem and Chris Bruckert (Ottawa: University of Ottawa Press, 2012), 2-3; In a non-discrimination law context, Solanke views stigma essentially as a "mark," referring to the Greek origins of the word, which means "to stick": Solanke (2017) 18.

[883] Stigma is often defined with reference to Erving Goffman as "an attribute that is deeply discrediting": Matthew Clair, "Stigma," in *Core Concepts in Sociology*, ed. J. Michael Ryan (Wiley-Blackwell, 2018), 318.

In clinical decision-making, the starting point is that differential treatment is necessary and desirable – it is the norm rather than the exception. There are many clinical decision contexts where a difference in treatment does not indicate that one patient is treated *less favourably* than another. For instance, in the case of diagnosis, a difference in diagnosis does not indicate that one patient is treated less favourably than another if both patients receive an accurate assessment.

Similarly, in relation to treatment recommendations and preventive interventions,[884] a difference in treatment does not indicate that one patient is treated less favourably than another, if there are medically relevant reasons necessitating the difference in treatment. Moreover, in such cases patients would not be in comparable situations, because they would differ in aspects that are relevant to the decision.[885] On the other hand, if an AI-CDS system generates a correct diagnosis for one patient and an erroneous diagnosis for another patient that actually has the same disease, this amounts to a less favourable treatment of the latter patient compared to the former. These examples illustrate how the inquiry into whether less favourable treatment exists is intertwined with the question of whether the patients are in comparable situations, particularly when it comes to diagnoses, treatment selection, or predictive interventions. When such clinical decisions entail differential treatment, the differential treatment does not necessarily imply a disadvantage as long as the patients are in situations that are not comparable with regard to the specific clinical decision at hand. This means that non-discrimination law does not prohibit differential treatment of patients based on PILFs for the purposes of these decisions, if the use of PILFs is medically justified. The issue of assessing whether the use of protected characteristics is medically justified is further discussed in section 9.5.

In relation to clinical decisions concerning the allocation of scarce resources, the situation is slightly different. When differential treatment between two patients occurs in this context, one of them will experience a disadvantage regardless of whether the situations are comparable or not; even if there are medically relevant reasons for the difference in treatment, receiving a resource such as a hospital bed or ventilator is usually more beneficial than not receiving it. In this context, therefore, it can be assumed that some patients are treated more favourably than

---

[884] Section 1.7.

[885] The comparability requirement is discussed in section 9.5.

others, regardless of whether the cases are comparable. When a resource denial, in this sense, occurs on the basis of a protected characteristic, there may be direct discrimination.

9.4.2 "Particular Disadvantage"

Indirect discrimination requires that persons in a protected group are put at a "particular disadvantage."[886] The types of disadvantages that may be relevant (see section 9.2) are the same as for direct discrimination. However, in relation to indirect discrimination, the *extent* of disadvantage is measured at the group level. Unlike in direct discrimination where the treatment of one individual compared to another is the crux, indirect discrimination focuses on the extent to which a disadvantage is more likely to affect one protected group compared to others. For instance, the fact that one person suffers the disadvantage of receiving a false negative diagnostic assessment is not enough to establish indirect discrimination. To determine whether there is a "particular disadvantage," a comparison between groups is necessary. Like in the context of direct discrimination, the relevant disadvantage in the context of indirect discrimination may be a disadvantage in absolute terms or a relative disadvantage (i.e., being treated less favourably than someone else).[887] Thus, it is arguably possible for a group to be put at a "particular disadvantage" by the deployment of an AI-CDS system even though deployment of the system improves the quality of clinical assessments for the group on average, compared to the situation before the AI-CDS system was deployed (currently, this usually means the situation where decisions are made without the assistance of AI).

For a practice to constitute indirect discrimination according to the RED, specifically, it is required that the measure "would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons."[888] Under the GSED, indirect discrimination is when a practice "would put persons of one sex at a particular disadvantage compared with

---

[886] Article 2(2)(b) RED; Article 2(b) GSED.

[887] The fact that a relative disadvantage is sufficient to trigger indirect discrimination is evident, e.g., from the ruling in the *Brachner* case (based on Directive 79/9). Here, the claimant, Ms Brachner, received a 1,7 % increase in her pension. Because this increase was lower than the increase received by certain other persons, the CJEU noted that persons in Ms Brachner's situation suffer a disadvantage: Judgment of 20 October, 2011, Brachner, C-123/10, ECLI:EU:C:2011:675, para. 57.

[888] Article 2(2)(b) RED.

persons of the other sex."[889] The requirement that the disadvantage must be *particular* to a protected group signifies an assessment that, at its core, is oriented towards a quantitative threshold: It is "particularly" persons of a protected group who must be disadvantaged.[890]

Over time, the CJEU has used a variety of more or less synonymous phrases to denote the quantitative threshold that determines whether a "particular disadvantage" is present. Traditionally, the Court has suggested that the disputed practice must disadvantage a "far greater number,"[891] or "far more"[892] persons from a protected group compared to other persons, that it must disadvantage "a significantly higher percentage"[893] of persons from a protected group, or that the advantageous outcomes must be available to a "much lower proportion of women" compared to men.[894] In recent CJEU case law, the thresholds are more consistently described as being met if persons from a protected group are disadvantaged in "considerably" or "significantly" greater proportions or numbers than other people.[895] This articulation of the threshold is arguably the one that best represents how the particular disadvantage requirement is understood, currently. In *CHEZ*, the Court clarifies that the requirement of a "particular disadvantage" does not imply a qualitative assessment of the disadvantage experienced by each person within the disadvantaged group. As with direct discrimination, it is sufficient that there is a less favourable treatment, which does not have to be "serious, obvious or particularly significant."[896]

---

[889] Article 2(1) GSED.

[890] CHEZ, C-83/14, para. 109; Jyske Finans, C-668/15, para. 27; Maniero, C-457/17, para. 47.

[891] Judgment of 13 July, 1989, Rinner-Kühn, C-171/88, ECLI:EU:C:1989:328, para. 16.

[892] Brachner, C-123/10, para. 56.

[893] Brachner, C-123/10, para. 63.

[894] Judgment of 13 May, 1986, Bilka, C-170/84, ECLI:EU:C:1986:204, para. 29. More examples of similar expressions from the CJEU are given by Wachter, Mittelstadt and Russell: Wachter, Mittelstadt, and Russell (2021 B) 13.

[895] CHEZ, C-83/14, para. 107, cf. paras. 99-02; Judgment of 8 May, 2019, Villar Láiz, C-161/18, ECLI:EU:C:2019:382, para. 38; Judgment of 3 October, 2019, Schuch-Ghannadan, C-274/18, ECLI:EU:C:2019:828, para. 45; Judgment of 28 August, 2020, Bvaeb, C-405/20,para. 49; YS, C-223/19, para. 49.

[896] CHEZ, C-83/14, paras. 98 and 109.

The reluctancy to indicate concrete, numerical thresholds is a deliberate choice from the CJEU. This choice facilitates appreciation of contextual and societal differences between situations in which EU non-discrimination law applies.[897] Arguably, although a disadvantage does not have to be particularly serious, it is permissible to consider the degree of seriousness of the relevant disadvantage when determining where the "particular disadvantage" threshold should be placed in a given case. Furthermore, the CJEU has suggested that the threshold may be lowered if a disparity between groups is "persistent and relatively constant" over a long period of time.[898]

It is also important to note that, although the CJEU uses the term 'significant,' this does not necessarily refer to the notion of 'statistical significance' that is used in empirical sciences. In academic literature, Fredman suggests that a difference between groups which is statistically significant is sufficient to establish a case of potential indirect discrimination.[899] However, application of the concept of statistical significance does not immediately clarify where to draw the threshold for a particular disadvantage. Statistical significance is contingent on the chosen significance level, which is often referred to as the 'p-value' in scientific contexts. The most commonly used p-value is 0,05. Nevertheless, if the particular disadvantage threshold can be translated into a p-value, hypothesis testing (the method of determining statistical significance) might be useful as a technical method for disadvantage measurement in a pre-deployment discrimination assessment. The question of exactly how hypothesis testing could be utilised in this context should be considered in future research efforts.

---

[897] Marc De Vos, "Substantive Formal Equality in Eu Non-Discrimination Law," in *The European Union as Protector and Promoter of Equality*, ed. Thomas Giegerich (Cham: Springer International Publishing, 2020), 253; Wachter, Mittelstadt and russel therefore argue that the abstract articulation of the quantitative threshold makes it difficult to build the prohibition of indirect discrimination into an AI system and, thus, that "fairness cannot be automated": Wachter, Mittelstadt, and Russell (2021 B) 28.

[898] Judgment of 9 February, 1999, Seymour-Smith, C-167/97, ECLI:EU:C:1999:60, para. 61.

[899] Fredman (2022) 291.

9.5 Comparability, Before and After Deployment

9.5.1 The Importance of Comparability in Connection with Direct and Indirect Discrimination

The non-discrimination principle is frequently described as inherently comparative.[900] To establish direct or indirect discrimination, a comparison is required, although there is a distinction between the group-oriented comparison requirement in relation to indirect discrimination and the individual comparison that is required in relation to direct discrimination. In both cases, the comparison requirement is closely connected with the requirement of a disadvantage, as discussed in the previous section.[901]

The Equality Directives' definition of direct discrimination explicitly requires the finding of a disadvantage based on *comparability*: Direct discrimination only arises if one person is treated less favourably than another person in a "comparable" situation. In contrast, a requirement of comparability is not explicit in the definition of indirect discrimination. In non-discrimination law scholarship, there are mixed views on whether is a requirement of comparability in relation to indirect discrimination.[902] Some authors argue that there is no requirement of comparability in relation to indirect discrimination, because consideration of the comparability between cases is not necessary to consider the negative impacts on a group.[903] The practical consequence of this view is that an assessment of potential indirect discrimination can concentrate on the distribution of outcomes from a certain practice, eliminating the need to scrutinise whether individuals affected differently by it are in

---

[900] Réaume (2013) 7; McColgan (2014) 101; Wachter, Mittelstadt, and Russell (2021 B) 10. The comparative aspect of the non-discrimination principle has been emphasised in case law long before the enactment of the Equality Directives. For instance, the CJEU notes in the *Sermide* ruling that "comparable situations must not be treated differently and different situations must not be treated in the same way unless such treatment is objectively justified": Judgment of 13 December, 1984, Sermide, C-106/83, ECLI:EU:C:1984:394, para. 28.

[901] Henrard summarises harm as a "less favourable treatment or disadvantage, with an element of comparison or comparability": Henrard (2019) 106.

[902] Waaldijk and Tobler openly disagree with each other on the matter: Waaldijk and Tobler (2009) 744-45.

[903] Dagmar Schiek, "Indirect Discrimination," in *Materials, Cases and Text on National, Supranational and International Non-Discrimination Law* (Hart Publishing, 2007), 468-71. This opinion is shared by Waaldijk: Ibid.

comparable situations. For example, if a hiring practice is assessed, one would simply compare the proportion of male applicants hired with the proportion of female applicants hired, regardless of whether the applicants within each group differ in important ways.

Other authors find, however, that the cases to be compared must indeed be 'comparable' also in the context of indirect discrimination.[904] Tobler argues that comparability is an integral part of the general non-discrimination principle and that it applies to indirect discrimination as well as direct discrimination.[905] Tobler's position is arguably supported by CJEU case law that is more recent than the literature discussing the issue of comparability in indirect discrimination. In *CHEZ,* the principles guiding comparison in EU non-discrimination law are expressed by the CJEU in the following way:

> … the requirement relating to the comparability of the situations for the purpose of determining whether there is a breach of the principle of equal treatment must be assessed in the light of all the elements which characterise them.[906]

The *CHEZ* ruling leaves it to the referring court to finally assess whether the disputed practice amounts to direct or indirect discrimination. The reference to the general "principle of equal treatment" appears as a reference to the fundamental principle of non-discrimination, which is often called the 'equal treatment' principle. It implies that the requirement of comparability is a fundamental part of the non-discrimination principle, without distinction between direct or indirect discrimination.[907] Arguably, when the CJEU refers to the requirement of comparability in the context of the general "principle of equal treatment,"[908] this indicates that

---

[904] Tobler in Waaldijk and Tobler (2009) 744-45; Bell also finds that the CJEU does apply the comparability requirement in cases concerning indirect discrimination, although he criticises the focus on comparability from a more policy-oriented viewpoint: Mark Bell, "The Principle of Equal Treatment: Widening and Deepening," in *The Evolution of Eu Law*, ed. Paul Craig and Gráinne de Búrca (Oxford: Oxford University Press, 2011), 632-33.

[905] This argument is put forward by Tobler: Waaldijk and Tobler (2009) 744-45.

[906] CHEZ, C-83/14, para. 89.

[907] Waaldijk and Tobler note that the CJEU expresses as much, for example, in the *Schnorbus* case: Waaldijk and Tobler (2009) 740.

[908] CHEZ, C-83/14, para. 89; See also, e.g., Maruko, C-267/06, paras. 67-69; Römer, C-147/08, para. 42; Frédéric Hay, C-267/12, para. 33; Escribando Vindel, C-49/18, para. 50; Judgment of 19 July, 2017, Abercrombie & Fitch Italia, C-143/16, ECLI:EU:C:2017:566, para. 25.

the requirement applies both to cases of direct and indirect discrimination. However, Bell's observation that the CJEU does not follow a consistent line on this matter remains apt.[909] The Court rather seems to concentrate on comparability to different degrees depending on the circumstances of each case. According to Bell, it uses the comparability requirement tactically in cases where it thus avoids "explosive politics."[910]

While there may well be different views on the matter, the following proceeds with the assumption that, as a starting point, comparability is relevant when assessing direct as well as indirect discrimination. However, the following discussion leads to some modification of this starting point as far as the indirect discrimination assessment is concerned. It is argued in section 9.5.4 that comparability should play an important role when measuring the disadvantage of resource denial, whereas consideration of comparability is not necessary to measure the disadvantage of disparate performance. This is primarily because, in relation to the latter type of disadvantage, consideration of comparability does not add anything of importance to the assessment (this argument is elaborated in section 9.5.4.3). The following section considers what 'comparability' entails under the non-discrimination principle.

9.5.2 What Makes Cases Comparable?

To find that cases are comparable, the CJEU has provided the following guidance:

> … it is required not that situations be identical, but only that they be comparable and, on the other hand, whether the situations are comparable must be determined not in a global and abstract manner, but in a specific and concrete manner in the light of the benefit concerned …[911]

As reflected in this quote, an assessment of comparability is oriented towards *situations* and not *persons*. While the situations need not be identical, the question is what level of similarity is required for them to be considered comparable. In general terms, the requirement is understood to mean that two persons must be in "materially similar circumstances."[912] The question is, however, what makes two cases materially similar. This problem is challenging in

---

[909] Bell (2011) 633.

[910] Bell (2011) 633.

[911] Maruko, C-267/06, paras. 67-69; Römer, C-147/08, para. 42; Frédéric Hay, C-267/12, para. 33.

[912] Liddell and O'Flaherty (2018) 44-45.

traditional ex post enforcement contexts,[913] as well as in the context of a pre-deployment discrimination assessment. Fredman notes that the choice of comparator itself requires a complex value judgment as to which of the myriad differences between two individuals are relevant and which are irrelevant.[914] The necessity of making value judgments in a specific and concrete manner means that criteria for comparability are heavily context-dependent. For example, in the *CHEZ* case, the CJEU states that:

> … all final consumers of electricity who are supplied by the same distributor within an urban area must, irrespective of the district in which they reside, be regarded as being, in relation to that distributor, in a comparable situation so far as concerns the making available of an electricity meter intended to measure their consumption and to enable them to monitor changes in their consumption.[915]

While this statement is tailored to the circumstances of the *CHEZ* case, it suggests that, as a starting point, all recipients of goods or services from the same provider should be considered comparable. Then there may be individual differences between service recipients that render them non-comparable. In *CHEZ*, the CJEU refers to service recipients "within an urban area," indicating that in this particular case the service recipients in rural areas are not comparable to those in urban areas. Specific guidance on comparison in a clinical decision-making context does not exist in CJEU case law. However, the preamble of the GSED, which prohibits sex discrimination, points out the importance of comparability in relation to medical care:

> To prevent discrimination based on sex, this Directive should apply to both direct discrimination and indirect discrimination. Direct discrimination occurs only when one person is treated less favourably, on grounds of sex, than another person in a comparable situation. Accordingly, for example, differences between men and women in the provision of healthcare services, which result from the physical differences between men and women, do not relate to comparable situations and therefore, do not constitute discrimination.[916]

As this recital suggests, physical differences between men and women can render two persons incomparable. This suggests a guiding principle for clinical decisions: two individuals may

---

913 Khaitan (2015) 71; Fredman (2022) 254-55.

914 Fredman, ibid.

915 CHEZ, C-83/14, para. 90.

916 Recital 12 GSED.

not be comparable if their differences are relevant to the clinical decision in question. While CJEU case law supports that comparison begins with all patients receiving services from a specific healthcare institution or from an AI provider should be considered as a starting point, a closer examination of individual patients may reveal that they find themselves in different medical situations. With this in mind, the following sections delve into comparability could be determined and how comparison could be conducted in relation to clinical decisions.

9.5.3 Comparison Pool and Comparability in an Ex Post Enforcement Context

In the context of a post-deployment assessment of whether an AI-CDS system has violated the non-discrimination principle, e.g., in a court trial, a claimant might want to provide evidence that a similarly situated person has been treated more favourably by the disputed AI-CDS system. In the fictitious case of Tesfay v. Storevik University Hospital,[917] Simon Tesfay's case is based on comparison with how the hospital treats Lars Holm. When assessing a discrimination claim based on evidence of an actual comparator person, the court must consider whether the comparator person is in circumstances materially similar to the claimant, in accordance with the guidance provided by the CJEU as described above. However, the claimant does not necessarily need to refer to an actual comparator person. It is sufficient if the claimant can show that a hypothetical comparator person would have been treated more favourably.[918] When assessing a claim based on a hypothetical comparator, the court needs to consider whether the hypothetical comparison is appropriately constructed. Finding or constructing an appropriate person to compare with can be challenging because actual comparators may not exist and it can be difficult to decide how a hypothetical comparator should be constructed to be realistic and allow for meaningful comparison to an alleged victim of discrimination.

Irrespective of whether comparison is conducted based on actual or hypothetical comparator cases, identifying the relevant persons to consider for the purposes of comparison may not be immediately evident. This necessitates defining how to delimit the relevant pool of cases for

---

[917] Section 5.1.

[918] In relation to direct discrimination, this follows from how direct discrimination is defined in the Equality Directives: Article 2(2)(a) RED and Article 2(a) GSED. Given that this thesis argues that a comparability requirement also applies for indirect discrimination, it is assumed that hypothetical comparators can also be used to demonstrate comparability in relation to indirect discrimination.

comparison, before the question of *comparability* (i.e., whether the cases are materially similar) between individual cases can be considered.[919] When it comes to how to delimit the relevant pool of potential comparators, the guidance provided in the CJEU's case law is articulated in a manner that is rather specific to the individual cases that have occurred before the court.[920] As mentioned in the previous section, all recipients of the same services from the same provider should be considered. Another example is that when the disputed measure is a national law regulating the working life, the relevant group of persons for comparison consists of "all those workers subject to the national legislation in which the difference in treatment has its origin."[921] While this provides little guidance in the specific context of AI-CDS systems, it indicates that all persons that are somehow affected by the disputed practice may be considered.

The trajectory of court proceedings is typically that a claimant refers to some actual or hypothetical persons who have been or would be treated more favourably than the claimant. Next, the court determines whether the comparators referred to by the claimant are comparable or not. *Ex post*, a court does not have to consider other comparators than those referred to by the parties to the specific dispute. However, a court might find that the claimant casts too wide a net by asking the court to consider how the disputed measure impacts persons within a very broadly defined pool. For instance, if a healthcare institution uses a version of an AI-CDS system with a model specifically trained and tailored to the patient population served by the institution, it is primarily relevant to consider patients on whom this specific version has been used. If a claimant wants to run the model on data from other areas, the fact that the model is not intended to be used in those areas would give the court a sound argument in favour of denying such comparison.

---

[919] Connolly underscores the importance of "choosing the pool" for comparison: Connolly (2011) 165; Khaitan frames it as a problem of determining "the relevant pool within which the two comparator groups" (i.e., the advantaged and disadvantaged group) are to be identified: Khaitan (2015) 75; Wachter, Mittelstadt and Russell frame it as a question of "the reach of the contested rule": Wachter, Mittelstadt, and Russell (2021 B) 8-9; Fredman sees it as a problem of defining "the relevant pool of comparison": Fredman (2022) 290.

[920] Wachter, Mittelstadt, and Russell (2021 B) 3.

[921] YS, C-223/19, paras. 52 and 72-74.

However, at this point, it is necessary to distinguish between claims based on direct and indirect discrimination. The fact that a model is intended to be deployed in a healthcare institution serving a specific geographical area does not render *individuals* from other areas entirely irrelevant as comparators. For direct discrimination to be found ex post, it is sufficient that one person would be treated less favourably than another. Even though a person lives in a different area, it is of course possible that this person might receive healthcare in the area where an AI-CDS system is deployed. Because an individual case of differential treatment is enough to find direct discrimination ex post, EU non-discrimination law arguably entitles a claimant to refer to individual cases outside of the intended geographical area of use, for the purpose of establishing direct discrimination. As long as one relevant comparator patient can be identified or constructed, direct discrimination can be found ex post. Without reference to such a comparator, direct discrimination cannot be found ex post. In contrast to the crucial role of comparison in relation to direct discrimination ex post, this thesis argues that consideration of the comparability between patients in a test dataset is less important when assessing direct discrimination in a pre-deployment setting. The reason for this is that the Feature Identification (chapter 8) and the assessment of direct causation (chapter 10) probably give sufficient indications of the likelihood that an AI-CDS system will cause direct discrimination.

In relation to indirect discrimination, the question of whether comparison based on data from outside of the population that an AI-CDS system is intended for, is a more complex issue. If a system is intended for a specific, geographically defined population, it would be less relevant to show that a "particular disadvantage" might have occurred if the system had been deployed outside of this population. A comparison in relation to indirect discrimination in an ex post enforcement context should arguably concentrate on the intended target population, with the consequence that a court may refuse comparison of group-level disadvantages based on data from other geographical areas. The importance of considering the intended target population in the context of a pre-deployment discrimination assessment is discussed below.

9.5.4 Comparison Pool and Comparability in a Pre-Deployment Assessment Context

9.5.4.1 Defining the Relevant Data for Pre-Deployment Comparison

Methodological element

Step 1:

- □ Determine how to construct an appropriate dataset for comparison purposes.

When assessing an AI-CDS system in a pre-deployment context, it has yet to be used in a clinical setting. Therefore, the delimitation of the relevant comparator pool cannot be based on a distinction between patients on whom the system has been used and patients on whom it has not been used. Moreover, in a pre-deployment setting, there is no individual claimant whose case serves as the starting point, or base case, for the assessment. Instead, a pre-deployment assessment necessitates comparison based on retrospective data.[922] Within the relevant dataset it is necessary to determine which cases are *comparable* to each other. This issue is further explored in the following sections. First, the present section discusses the definition of the relevant population from which to collect data for the purposes of the comparison in a pre-deployment setting. In theory, a dataset intended for testing an AI model may encompass hundreds of thousands, or even millions, of individual cases. To conduct a comparison as necessary for assessing discrimination, an important step would involve determining how to construct an appropriate dataset for comparison purposes.

[922] It is assumed that testing of an AI-CDS system as part of pre-deployment discrimination assessments in practice will regularly rely on retrospective data. However, it is also possible that clinical trials could be conducted, in which case a pre-deployment assessment could also rely on prospective data.

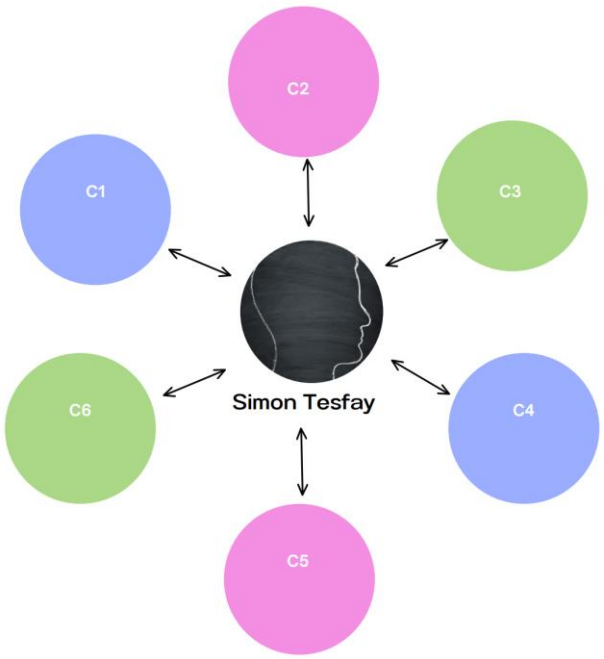


Figure 5: Illustration of comparison in a post-deployment assessment context, with reference to the fictitious case of Simon Tesfay v. Storevik University Hospital (section 5.1).

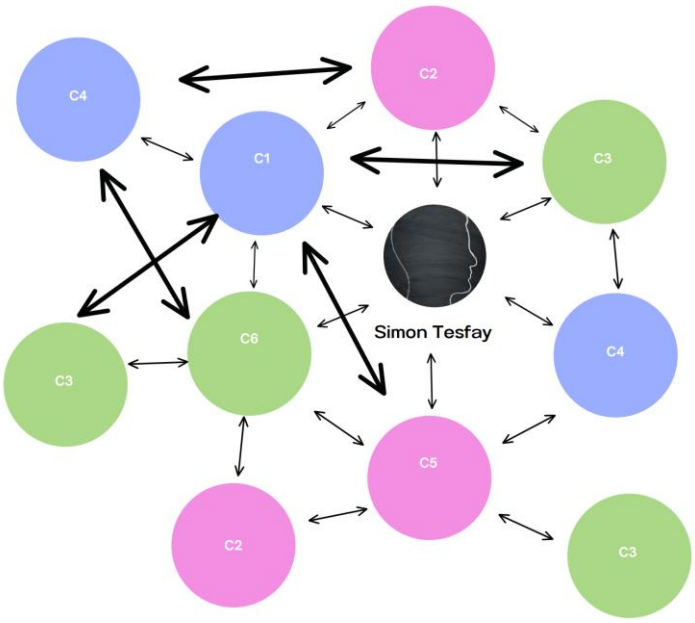


Figure 6: Illustration of comparison in a pre-deployment discrimination assessment.

Where an AI-CDS system is intended to be distributed to several different deployers, the comparator pool should arguably include patients from all relevant geographic areas, i.e., the areas where the system is intended to be deployed. Of course, this does not guarantee that the system will be suitable for all patients on which it may be used. Any given healthcare

provider may receive patients from all over the world, but it is not feasible to test the system based on data that represents all patients in the world. Given the necessity of delimiting the pool of test cases used for comparison purposes, the population in the geographic area where a system is intended to be used is an appropriate starting point. This starting point aligns well with the CJEU's emphasis on considering service recipients within the geographical reach of a service provider.[923]

This implies that if a healthcare institution develops an AI-CDS system for use only within the institution,[924] consideration of patients from the geographic area served by that institution could be sufficient. In recognition of the fact that a healthcare provider will treat patients that are not representative of the geographically defined population for which the system is intended, specific safeguards should be implemented to prevent less favourable treatment of such patients. This implies a need for defining a way to determine whether a patient is within the group that the system should be used for or not. For instance, ML research looks into various ways for models to estimate the extent to which they are confident when faced with an individual case.[925] If the case seems different from the cases the model knows from its training, this should lead to a low confidence score. These algorithmic confidence scores might be used to determine whether additional safeguards should be applied before an AI-CDS system is used on an individual patient.

Furthermore, it is worth noting that there are certain challenges of constructing test datasets based only on the geographic areas where an AI-CDS system is intended to be used. To illustrate the practical difference between systems intended for one institution and systems intended for several institutions, consider two different strategical directions contemplated in the NORspine project.[926] One option for the project could be to develop a model solely for the University Hospital of North Norway. Alternatively, the project could aim to develop a model to be integrated into an EHR system used by several healthcare institutions across Norway. If the first strategy is chosen, one may argue that data from the population in northern Norway

---

[923] CHEZ, C-83/14, para. 90.

[924] In which case the system is exempted from the certification scheme according to the MDR.

[925] e.g., Chuan Guo et al., "On Calibration of Modern Neural Networks" (paper presented at the International conference on machine learning, 2017).

[926] Section 5.2.

would suffice for comparison purposes. With the latter strategy, in contrast, if a comparison dataset is collected from that population only, this would arguably not be sufficient. For instance, a comparison based solely on this population might not reveal discrepancies in the system's performance when applied to ethnic groups primarily residing in southern parts of Norway.

Geography is not the only aspect that may be relevant to consider when defining an appropriate dataset for comparison purposes. In relation to an AI-CDS system intended for the Norwegian population, one could argue that data from other countries with similar living standards, education levels, health conditions, etc., are also representative of the target population. This argument implies that a discrimination assessment can be based on a dataset that includes data from other, comparable geographical areas. This argument is less loyal to the CJEU, because the Court emphasises the geographical reach of a service provider, as noted. However, it has also been noted that the Court's instructions on how to define the relevant pool of comparators tend to be articulated in a highly case-specific manner. In the specific context of AI-CDS systems, requiring that comparison must be based only on data from the geographic area where a system is intended to be deployed would mean that AI provides could not distribute their systems to deployers in other countries than those from which test data is collected. Arguably, a proper pre-deployment discrimination assessment could rely on data from other areas if the populations are comparable – an issue worthy of further contemplation in research and policy discussions going forward.

Methodological element

- □ Is the AI-CDS system intended to be distributed to different deployers and used without local modifications?
- ➢ Include data collected from the geographical area(s) where the system is intended to be deployed, or from comparable populations

While the abovementioned considerations are developed based on an interpretation and adaptation of the non- discrimination principle, the AIA should also be consulted when determining how to appropriately construct a dataset for comparison purposes. The comparison that must be conducted as part of a pre-deployment discrimination assessment will form part of an AI provider or deployer's testing procedures, which encompass broader aspects beyond discrimination. While the AIA does not specifically address the comparison aspect of testing, it does address the dataset used for testing purposes more generally. It requires an AI provider

to into account the intended population on which an AI system will be used when defining training, validation and testing data.[927] This shows that the AIA assumes that the target population will be specified by the provider. Moreover, the AIA requires that "characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk AI system is intended to be used" must be considered.[928] Thus, the AIA's general data governance rules support the notion that geographic reach, while important, is not the only aspect that should be considered when defining the relevant test data.

9.5.4.2 Comparability of Cases in Pre-Deployment Comparison

The second step of a pre-deployment comparison would be to determine, within the relevant dataset, which cases are *comparable* to each other. The non-discrimination principle suggests that cases representing patients in materially similar circumstances are comparable.[929] This section outlines how comparability of cases may be determined in the context of a pre-deployment discrimination assessment. As noted in section 9.5.2, the problem is to define the criteria according to which two patients should be deemed materially similar in relation to specific clinical decisions.

Methodological element

Step 2:

- □ Determine which cases within the relevant test dataset are comparable to each other.

According to the CJEU, a comparison should be carried out in a "specific and concrete manner in the light of the benefit concerned."[930] Thus, the question is whether two cases are sufficiently similar in respect of the specific clinical decision concerned. This implies the necessity of methods for determining whether patients within a given dataset differ in

[927] According to Article 10(3) AIA (EP), training, validation and testing data shall be sufficiently representative and have appropriate statistical properties as regards the persons or groups of persons on which an AI system is intended to be used.

[928] Article 10(4) AIA (EP).

[929] Section 9.5.2.

[930] Maruko, C-267/06, paras. 67-69; Römer, C-147/08, para. 42; Frédéric Hay, C-267/12, para. 33.

clinically relevant aspects. The following discussion highlights certain existing methods within ML science which may be utilised to compare patients in the context of a pre-deployment discrimination assessment.

As mentioned in section 9.5.3, comparison can be based on actual or hypothetical persons. In the context of a pre-deployment assessment of an AI-CDS system, where the model is run on retrospective patient data, each comparator case can be seen as a hybrid between an actual and a hypothetical case. It does not reflect an actual case demonstrating how a person was treated, because the model has not yet been used to make decisions in a clinical setting. However, the data are derived from real patients, thus making it more than hypothetical. Synthetic data could potentially allow for the creation of hypothetical comparator cases, but this practice is not yet widespread.[931] Whether one uses real patient data or synthetic data, the criteria for comparability remain the same.

9.5.4.3 Ground Truth Comparability in Relation to the Target Variable

When running a model on patient data for the purposes of comparison between cases, the crux of the test is to identify cases that are similar enough to warrant equal treatment irrespective of differences pertaining to protected characteristics. In non-discrimination law, the purpose of the comparison is to compare persons of different protected characteristics. When defining 'comparability,' therefore, one must look away from differences in protected characteristics. In the context of the RED, comparability must be assessed irrespective of ethnicity and, under the GSED, comparability must be assessed irrespective of biological sex.

Methodological element

Criterion for comparability:

- □ Ground truth comparability

One way to understand comparability is as a matter of similarity in ground truths of the target variable. The term 'ground truth' refers to the accurate, true value of a feature or data point. For example, two patients who have coronary artery disease should both be diagnosed with coronary artery disease. Here, the target variable is the patient's state in relation to this diagnosis (positive or negative), and the relevant ground truth is therefore the patient's actual

[931] Chen et al. (2021).

state. When this ground truth is the same for two patients, i.e., both are either negative or positive, they are in comparable situations in relation to the diagnosis of coronary artery disease. This argument builds on a rather straightforward principle for individual comparison when assessing an AI-CDS system intended for classification tasks:[932] patients who share the same ground truth are in comparable situations. For ease of reference, the following refers to this criterion as 'ground truth comparability.'

There are important limitations to the ground truth comparability principle. One issue with ground truth comparability is that the ground truth of the target variable is sometimes debatable or unreliable, because the target variable is a feature which is not easily or objectively measurable. This issue may arise in relation to classification models (e.g., if medical experts disagree on a diagnosis)[933] as well as regression models. When the intended output from an AI-CDS system is a regression, such as a risk score, there is no singular data point that can be used to determine comparability between two cases. For the purposes of a risk prediction, for instance, two patients with coronary artery disease are not necessarily comparable, because other factors setting the two patients apart may be relevant to the risk prediction. Rather, if the target variable is a risk score, comparability is ideally a matter of the 'true' risk. Similarly, if the target variable is a score representing a patient's health needs, the ideal comparability determinator is the 'true' health needs of the patients. For example, in relation to resource allocation, two patients with similar needs should have a similar chance of receiving a resource. However, 'risk' and 'need' are not easily measurable sizes.[934]

When it comes to measuring the disadvantage of disparate performance, it is arguable that the determination of comparability cannot be separated from the measurement of disadvantage. The measurement of disparate performance presupposes that the ground truth is available

---

[932] Section 1.5.5.

[933] Especially in psychiatry, there may be few biologically or physically measurable markers of a disease: Hugo Corona Hernández et al., "Natural Language Processing Markers for Psychosis and Other Psychiatric Disorders: Emerging Themes and Research Agenda from a Cross-Linguistic Workshop," *Schizophrenia Bulletin* 49, no. Supplement_2 (2023), https://doi.org/10.1093/schbul/sbac215.

[934] Paulus and Kent (2020) 1. ("… "risk" is not a property that can be objectively measured in an individual (like blood pressure or cholesterol) – but instead can only be estimated in a group of other individuals judged to be similar in a set of selected features…").

when testing a model. In relation to disparate performance, it is the relationship between the ground truth and the model's output that determines whether someone is disadvantaged. This type of disadvantage can therefore be measured without a separate consideration of which patients are in comparable situations.

9.5.4.4 Feature Comparability

Methodological element

Criterion for comparability:

- □ Feature comparability

The challenges of determining ground truth comparability based on target variables, invokes the need for alternative or additional approaches. One possible way to determine comparability is to identify cases that share certain key features. This can be referred to as 'feature comparability.' While ground truth comparability is based on the actual value of a specific data point defined as a target variable, feature comparability considers multiple attributes of a patient. Feature comparability implies a broader form of comparison that is based on a set of selected features. For instance, two patients may be considered similar not only based on a specific diagnosis (ground truth), but also on other characteristics such as age,[935] diet, or medical history. In the NORspine project, identification of similarly situated patients based on selected features is being considered because this might be a way of enhancing interpretability of algorithmic outputs. In this project, it has been proposed that comparability might be defined by reference to features such as age, sex, self-reported pain and time passed since pain debut. Because the purpose of identifying similar patients in this project is to enhance interpretability, similarity could be defined by reference to protected characteristics including sex. However, for the purposes of discrimination assessment, sex should not be included as a feature when determining the comparability of cases.

Contrasted with ground truth comparability, feature comparability gives a more holistic comparison not only based on the true value of a target variable, but also on a wider range of factors that may be predictive of the value of the target variable. Ground truth comparability is a narrower form of comparison that might not account for all relevant factors that should be considered when comparing patients as part of a discrimination assessment. Even though two

[935] Age is a protected characteristic in some circumstances, but it is not protected by the RED or the GSED. It is often a relevant factor in clinical decision-making.

patients in a dataset have the same ground truth for the target variable, they could still be different in terms of other features. This is why considering feature comparability in addition to ground truth comparability can result in a more comprehensive and nuanced comparison.

While feature comparability provides a more comprehensive basis for comparison, it is also more complex and may be more difficult to determine accurately, especially when there is a large number of features to consider or when the relationships between features are not well understood. There is often an incomplete understanding of causal relations in observational health data.[936] Frequently, it is difficult to determine whether a feature distinguishing two patients is relevant to a clinical decision. Therefore, there is a risk that the determination of comparability based on a certain features could either include a feature that should not be relied on or exclude a feature that should be relied on.

Methodological element

Caveat for disparate performance:

- Comparability mapping not necessary to assess disparate performance

9.5.4.5 Technical Methods

To conduct a comparability analysis in practice, one must identify a technical method through which the relevant criteria for comparability (e.g., ground truth comparability or feature comparability) can be applied to the dataset. Various techniques could be considered as methods of identifying comparable cases.

If comparability is defined solely with reference to the ground truth of target variables, the identification of comparable cases does not require advanced technical methods; a simple search for patients sharing the relevant ground truth is sufficient, assuming that the ground truth is a searchable data point in the dataset.

Methodological elements

Technical methods:

- Ground truth comparability
  - Search and compare
- Feature comparability
  - Cluster analysis
  - Propensity score matching

[936] Thus, in medical literature, Paulus and Kent note how difficult it can be to distinguish "legitimate causal attributes from illegitimate race-proxies.": Paulus and Kent (2020) 3.

When feature comparability is applied, technical methods of cluster analysis are promising. These methods are subject to ongoing research within ML science.[937] Cluster analysis refers to the application of an ML algorithm that automatically groups instances (in this case, patients) into clusters based on their similarities across multiple features.[938] When the groups identified through cluster analysis are held up against the outputs produced by an AI-CDS system, there should be a strong correlation between the groups and the outputs. It is the cases where patients in the same cluster receive different outputs that must be scrutinised to determine whether there the model produces different outputs for similarly situated patients.

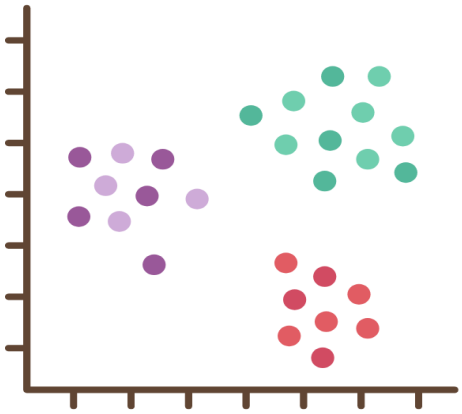


Figure 6: Illustration of how a cluster analysis may be visualised.[939]

Another category of techniques that could be utilised to identify similarly situated patients based on feature comparability is various forms of 'propensity score matching.' This approach is generally used to measure the distance between objects (in this context, patients) in terms of similarity. In medical research, propensity score matching is particularly used to

---

[937] Caroline X. Gao et al., "An Overview of Clustering Methods with Guidelines for Application in Mental Health Research," *Psychiatry Research* 327, no. 115265 (2023), https://doi.org/10.1016/j.psychres.2023.115265.

[938] e.g., Junyuan Xie, Ross Girshick, and Ali Farhadi, "Unsupervised Deep Embedding for Clustering Analysis," *Proceedings of the 33rd International Conference on Machine Learning* 48 (2016); Vranas et al. (2017).

[939] The figure is derived from Canva and included in this thesis in accordance with Canva's terms and conditions.

minimise the impact of confounding factors in observational studies.[940] For example, in an observational study aiming to assess the effect of a COVID vaccine, the most straightforward way to measure the effect would be to compare outcomes for vaccinated persons with outcomes for unvaccinated persons. However, this would fail to account for important factors that influence patient outcomes, other than whether or not a person is vaccinated (e.g., age, health issues, smoking habits, etc.). To properly measure the effect of a vaccine through an observational study, vaccinated persons should be compared to persons who are as similar as possible to themselves but who did not have the vaccine. Propensity score matching could be applied to identify similar patients in a dataset.[941]

As already indicated, feature-based methods offer several benefits. They allow for comparisons based solely on the features used as inputs in an AI-CDS system. This comparison is independent of any specific outcome and is therefore not dependent on knowing the ground truth of the target variable. Moreover, patients are not grouped together based on the prediction an AI model would make about them. Instead, they are grouped together based on similarities in certain features. These features are considered independently of the model's predicted outcome.

The purpose of cluster analysis, propensity score matching, or other techniques used to identify comparable patients, would be to take the results and compare them with model outputs during a pre-deployment discrimination assessment. This way, one could identify the cases that receive different outputs despite being deemed comparable. The point of identifying comparable cases is to facilitate a disadvantage measurement, where it is considered whether a protected group is more often represented among the patients who receive a disadvantageous output compared to similarly situated patients. This would indicate

---

[940] Confounding factors are factors that may contribute to the patients' health outcomes without being accounted for in the study: P. C. Austin, "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research* 46, no. 3 (May 2011), https://doi.org/10.1080/00273171.2011.568786; Janick Weberpals et al., "Deep Learning-Based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-Scale, Real-World Data Study," *Epidemiology* 32, no. 3 (2021), https://doi.org/10.1097/ede.0000000000001338.

[941] "Propensity Score Matching Methodology: Why and How It Is Used," (Video Journal of Biomedicine, 13 December 2022), Youtube. https://www.youtube.com/watch?v=40U_0DtNrQc.

a potential for indirect discrimination, as further discussed in section 9.6 below. In relation to pre-deployment assessment of direct discrimination, chapter 10 will argue that it is often not relevant to consider the comparability of patients in a test dataset, because direct discrimination is bound to occur if a direct causation assessment finds that protected characteristics are sufficiently influential on a model's outputs. However, there is one caveat to this argument: If determining a protected characteristic's degree of influence on model outputs is not feasible due to its opacity, the distribution of outputs among patients who are similarly situated except for a protected characteristic could be considered when assessing the likelihood of direct discrimination.

9.5.5 Conclusion

The test data used when assessing discrimination in an AI-CDS system before deployment should include a mapping of patients that are in comparable situations in relation to the clinical decision at hand. Such a mapping plays its most important role when assessing indirect discrimination in AI-CDS systems intended for the allocation of scarce resources, where the principal disadvantage to consider is resource denial. It could also play a role in assessing the likelihood of an opaque model leading to direct discrimination if Feature Identification (chapter 8) and direct causation assessment (chapter 10) is not sufficient to conclude on deployment. In relation to indirect discrimination where the principal disadvantage to consider is disparate performance, comparability is not relevant as a separate consideration. Comparability in relation to disparate performance would have to be based on ground truth comparability, and the measurement of disparate performance in itself assumes the consideration of ground truths.

The CJEU has provided limited casuistic guidance on how to define a relevant pool of comparators and how to determine whether two situations are comparable. Drawing upon the CJEU's jurisprudence, this section has developed certain principles for the construction of an appropriate dataset for comparison purposes in a pre-deployment assessment context. Ground truth comparability and feature comparability have been suggested as relevant criteria that are technically applicable while being compatible with an interpretation of EU law that is loyal to the CJEU's guidance in ex post contexts.

While this section has developed criteria for comparability based on the non-discrimination principle and relevant technical approaches to discrimination assessments based on these criteria, these methodological elements are not without limitations. One general limitation is

that criteria such as ground truth comparability and feature comparability assume that relevant features have been measured and included in the dataset used for comparison. This may not always be the case in practice. Moreover, determining comparability between cases in a dataset requires careful evaluation of the choice of which features to consider in the first place.[942]

How to define and identify similar cases in a dataset is a subject of ongoing research within ML science. This section has highlighted the potential for applying existing ML techniques to identify comparable cases in the context of a pre-deployment discrimination assessment. However, ML science tends to focus on general *similarity* rather than the *comparability* that is of interest when applying non-discrimination law. For example, identification of similar cases can be used to enhance the interpretability of AI systems. For general interpretability purposes, AI developers may include protected characteristics when identifying patients that are similar to each other. This is the case in the NorSpine case study described in section 5.2, where developers have proposed a function that identifies similarly situated patients based on features that may include sex and ethnicity (both feature variables are collected in the Norwegian Spine Registry which is used for training purposes). In contrast, to assess discrimination in an AI-CDS system, it is necessary to identify patients that are similarly situated according only to the data points that are medically relevant while disregarding protected characteristics such as sex and ethnicity.

Further methods-building in ML science is encouraged, to develop new techniques and improve existing techniques that may be used to identify patients in materially similar circumstances except for the protected characteristics under non-discrimination law. Such methods-building should be informed by clinical expertise relevant to specific clinical decisions, as well as medical ethics.

---

[942] Xie, Girshick, and Farhadi note that in the literature on the methodology of cluster analysis, "[t]he choice of feature space is customarily left as an application-specific detail for the end-user to determine": Xie, Girshick, and Farhadi (2016) 478.

9.6 Measuring Comparative Disadvantage at the Group Level

9.6.1 Introduction

The previous sections explained the notion of disadvantage in EU non-discrimination law (section 9.2), defined relevant disadvantages potentially arising from the use of AI-CDS systems (section 9.3), and elaborated the disadvantage requirements for direct and indirect discrimination (section 9.4). In relation to indirect discrimination, it was emphasized in section 9.4.2 that a disadvantage, measured at the group level, must exceed a quantitative threshold before indirect discrimination arises. In relation to direct discrimination, there is no quantitative threshold – it is sufficient that one person is treated less favourably than another (section 9.4.1). Considerations relating to the construction of an appropriate dataset and mapping comparability of cases within such a dataset were developed in section 9.5.

To assess indirect discrimination, a comparative assessment is necessary to determine the extent to which a protected group is affected by an AI-CDS system relative to other groups.[943] A significantly higher proportion or number of individuals from a protected group must be disadvantaged by an AI-CDS system in comparison to others (a "particular disadvantage"), for the system to be deemed indirectly discriminatory.[944] While there are no fixed, generally applicable thresholds derived from CJEU case law, an ex post assessment of a discrimination claim as well as a pre-deployment discrimination assessment, must somehow measure the comparative disadvantage incurred by protected groups. The following proceeds to the question of how to measure the disadvantage for one group relative to others, focussing on the pre-deployment discrimination assessment context.

---

[943] In chapter 8, concerning the distinction between direct and indirect algorithmic discrimination, it was shown that direct discrimination may arise if a group is affected to such an extent that persons from the group are exclusively disadvantaged or the entire group is excluded from the chance of being advantaged. It is therefore possible that a measurement of the disadvantage could lead to a finding of direct discrimination. However, it was argued in chapter 7 that this will be extremely rare in the context of AI-CDS systems.

[944] Section 9.4.2.

9.6.2 Quantifiable and Non-Quantifiable Disadvantages

When discussing how disadvantage may be properly measured, it is important to distinguish between the different types of disadvantages that AI-CDS systems can cause. Section 9.3 outlined three main categories of relevant disadvantages: disparate performance, resource denial (regardless of performance), and stigma and prejudice. The question is how these disadvantages can be measured, to determine whether a protected group is at a "particular disadvantage" when testing an AI-CDS system before deployment.

One immediate observation is that the disadvantages of disparate performance and resource denial are more amenable to quantification than the disadvantage of stigma and prejudice. As noted in section 9.4.2, the CJEU primarily takes a quantitative approach to the determination of whether a disadvantage is "particular," by asking whether a protected group is affected in significantly higher proportions or numbers than other persons. While CJEU case law generally recognises that stigma and prejudice are relevant disadvantages, these disadvantages are usually not the most salient disadvantages motivating a claimant's case. Stigma and prejudice tend to be treated as complementary to more tangible, allocational disadvantages.

In the context of clinical decision-making, it is likely that decision-making practices causing stigma and prejudice will do so primarily because of the allocational harms of disparate performance and resource denial. For example, if an AI-CDS system consistently

Methodological elements

Step 3: Define the relevant disadvantages produced by the AI-CDS system (disparate performance is always relevant)

Considerations:

- □ Is it intended that the system's outputs will be relied on to determine the allocation of scarce resources?

  Implication:
  - ➢ The disadvantage of resource denial should be measured, in addition to disparate performance

- □ Does the AI-CDS system display a type of bias that is likely to cause/reinforce stigma and/or prejudice?

  Implications:
  - ➢ Lower threshold for finding a "particular disadvantage"
  - ➢ Take stigma/prejudice into account in the objective justification test

yields more false positives for certain psychiatric diseases in an ethnic minority group compared to other patients, this is stigmatising in addition to being directly harmful to the health of the patients who receive false positives. In terms of developing an assessment methodology, a relevant consideration emerges at this point: The possibility that a biased AI-CDS system could cause or reinforce stigma or prejudice should be considered when measuring disadvantage. However, because the disadvantages of stigma and prejudice are difficult to quantify and can be assumed to arise in addition to more allocational harms, this consideration should not concentrate on estimating the extent of stigma or prejudice potentially caused by an AI-CDS system. Rather, the implication of finding that bias in an AI-CDS system may cause stigma or prejudice is, arguably, that the threshold for finding a "particular disadvantage" could be lowered, due to the additional disadvantage of stigma/prejudice. Furthermore, the potential stigma and prejudice resulting from an AI-CDS system are important considerations to include as part of the assessment of whether biases in the system are objectively justified.[945]

The following sections discuss, in the light of CJEU case law, methods which may be used to measure the disadvantages of disparate performance and resource denial, including how the threshold for finding a particular disadvantage can be articulated when these methods are applied. However, CJEU jurisprudence does not specify a general quantitative threshold for the assessment of a particular disadvantage. Rather than aiming to specify a threshold, the following therefore aims to outline considerations that may influence whether the threshold should be lowered or heightened when assessing discrimination in an AI-CDS system before deployment.

9.6.3 Which Methods of Disadvantage Measurement are Recognised in EU Non-Discrimination Law?

9.6.3.1 Comparison of the Probability of Receiving an Advantageous Output

There is not one universally applicable method for the measurement of disadvantages in EU non-discrimination law. Depending on the circumstances, a variety of approaches may be relevant to the determination of whether there is a "particular disadvantage." The CJEU has

[945] Nilsson notes in relation to Article 14 ECHR that the harm of negative stereotyping implies that particularly weighty reasons will be required if such policies are to be considered acceptable: Nilsson (2020) 143.

provided limited general guidance on the matter, and it has applied different approaches in its case law, depending on the circumstances of each case. As noted in section 9.4.2, the threshold of a "particular disadvantage" implies that persons from a protected group are disadvantaged in "considerably" or "significantly" greater proportions or numbers than other people.[946]

The earliest explicit statement the CJEU has made about the appropriate method of measuring a comparative disadvantage, is from the *Seymour-Smith* case, which concerned differential treatment between different groups of workers. Here, the Court states that

> the best approach to the comparison of statistics is to consider, on the one hand, the respective proportions of men in the workforce able to satisfy the requirement of two years' employment under the disputed rule and of those unable to do so, and, on the other, to compare those proportions as regards women in the workforce.[947]

The *Seymour-Smith* ruling is based on Directive 76/207/EEC, which has since been repealed by Directive 2006/54/EC on the equal treatment between men and women in matters of employment and occupation. Unlike the directive that replaces it, Directive 76/207 did not define indirect discrimination and, thus, it did not contain the explicit requirement of a particular disadvantage. Nonetheless, the instructions from the CJEU in this case have been labelled by Wachter, Mittelstadt and Russell as the CJEU's 'gold standard' for comparison.[948] The status of the comparison method suggested in *Seymour-Smith* is further strengthened by the CJEU's more recent ruling in *Villar Láiz*.[949] In this employment-related case, the Court reiterates that *the best approach* to comparison is to consider the respective proportions of persons who are who are not affected by the rule at issue among men and, on the other hand, the same proportions among women.[950] The Court further underscores that it "is not sufficient to consider the number of persons affected, since that depends on the number of working people active in the Member State as a whole as well as the percentages of men and women

---

[946] CHEZ, C-83/14, para. 107, cf. paras. 99-02; Villar Láiz, C-161/18, para. 38; Schuch-Ghannadan, C-274/18, para. 45; Bvaeb, C-405/20, para. 49; YS, C-223/19, para. 49.

[947] Seymour-Smith, C-167/97, para. 59.

[948] Wachter, Mittelstadt, and Russell (2021 B) 17. The authors also refer to other cases where the same approach has been applied.

[949] Villar Láiz, C-161/18.

[950] Villar Láiz, C-161/18, para. 39.

employed in that Member State."[951] This suggests that comparing the proportion of advantaged *and* disadvantaged women with the proportion of advantaged *and* disadvantaged men is an appropriate approach, especially where considering the number of disadvantaged persons would not be sufficient.[952]

As noted by Tobler, the *Seymour-Smith* method of comparison is not consistently applied by the Court.[953] One may indeed argue that EU non-discrimination law is rather flexible in terms of permitting the method of comparison that is most suitable given the specific circumstances of each case.[954] For instance, when the guidance derived from *Seymour-Smith* and *Villar Láiz* is adapted to the context of assessing AI-CDS systems, it appears that the appropriate method of measuring disadvantage may depend on whether different protected groups are equally represented in the test dataset. Whether there is unequal representation of protected groups in the test dataset used for comparison purposes is therefore a relevant consideration in a pre-deployment discrimination assessment of an AI-CDS system. If there is unequal representation, it would be loyal to CJEU jurisprudence to consider both the advantaged and disadvantaged group.

Specifically, what the CJEU recommends in *Seymour-Smith* is to compare the *proportion of persons* from one protected group that are advantaged/disadvantaged to the proportion of other persons that are advantaged/disadvantaged.[955] Adapting this formula to the context of assessing discrimination in AI-CDS system before its deployment, it may be more efficiently expressed as implying a comparison of the *probability* that a person from a protected group will receive an advantageous output from the system to the probability that other persons will

---

[951] Villar Láiz, C-161/18, para. 39.

[952] Similarly, the ruling in *Seymour-Smith* notes that, in that case, it was not sufficient to consider only the disadvantaged group, because the composition of this group "depends on the number of working people in the Member State as a whole as well as the percentages of men and women employed in that State": Seymour-Smith, C-167/97, para. 59.

[953] Tobler notes that this approach is not even applied in the very case where it is recommended: Tobler (2008): 41.

[954] Tobler concludes that "in practice the test does not necessarily have to be a relative one, but can be flexible and pragmatic": Ibid. See also: Timo Makkonen, *Measuring Discrimination: Data Collection and Eu Equality Law*, European Commission Directorate-General for Employment, Social Affairs and Equal Opportunities (Luxembourg: Publications Office of the European Union, 2007), 36.

[955] Seymour-Smith, C-167/97, para. 59; Seymour-Smith, C-167/97, para. 39.

receive an advantageous output.[956] This is a more suitable articulation when applying the non-discrimination principle preventively rather than retrospectively. Moreover, as noted in chapter 7, the probability of discrimination is central to the pre-deployment discrimination assessment requirements proposed in the AIA. In addition to comparing the probability of advantage/disadvantage, other measurement methods may also be relevant, as summarised in section 9.6.3.3 below.

One issue that remains elusive regardless of the method one applies when measuring disadvantage, is the threshold for a particular disadvantage. In the *Moreno* ruling, it is concluded without discussion that a measure that disadvantages part-time workers puts women at a particular disadvantage given that 80 % of part-time workers in the relevant population are women.[957] Thus, it seems clear that if 80 % of the disadvantaged group consists of persons from a protected group, there is a particular disadvantage. This applies if, as in *Moreno*, one considers only the composition of the disadvantaged group, which is an alternative method to the one derived from *Seymour-Smith.* When considering only the composition of the disadvantaged group, it is not clear what the lower threshold for a particular disadvantage would be.

In the *Brachner* case, it appears that the CJEU applies the measurement method recommended in *Seymour-Smith.* This case is based on Directive 79/7/EEC. The CJEU reiterates the percentage of men and the percentage of women that were advantaged and disadvantaged by a pension scheme. In this case, 57 % of female pensioners and 25 % of male pensioners were disadvantaged. Thus, the percentage of women in the advantaged pool was 43 %, compared to 75 % for men. The CJEU indicates that the disparity between men and women in this case is significant enough to find that the disputed scheme disadvantages a "significantly higher percentage of female pensioners than male pensioners."[958]

The *Brachner* Court does not explicitly quantify the difference between men and women in terms of the probability of being advantaged/disadvantaged based on the statistics presented in that case. However, based on the numbers mentioned in the ruling, it appears that women

---

[956] Wachter, Mittelstadt, and Russell relate the formula from *Seymour-Smith* to the statistical 'fairness metric' they refer to as 'demographic parity': Wachter, Mittelstadt, and Russell (2021 B) 22.

[957] Judgment of 22 November, 2012, Moreno, C-385/11, ECLI:EU:C:2012:746, para. 31.

[958] Brachner, C-123/10, paras. 60-63.

had a 43 % advantage rate, leaving them with 57 % of the advantage rate for men, which was 75 %. If the statistical probability of being advantaged is deemed an appropriate measure, it seems arguable based on *Brachner* that persons in a protected group are indeed at a "particular disadvantage" if their chance of receiving an advantageous outcome is less than 60 % of other persons' chances. In non-discrimination law scholarship, a threshold around 75 % has been alluded to.[959] Thus, it seems reasonable to conclude that a "particular disadvantage" is established if a protected group has a probability of being advantaged that lies below a threshold somewhere between 60 % and 75 % of the probability that other persons have of receiving an advantageous decision, as a rule of thumb. However, the threshold cannot be fixed exactly under current law.[960] Moreover, as noted in section 9.4.2, the CJEU has suggested that the threshold may be lowered in cases where there is a "persistent and relatively constant" disparity over a long period of time.[961] In the context of a pre-deployment assessment of an AI-CDS system it does not make sense to consider the persistence and duration of a disadvantage because one does not study how a practice affects persons over time. However, in this context, the consideration of persistence and duration could be understood as tantamount to considering the sample size in a test dataset. As a general guiding principle, a lower disparity should be deemed to indicate a potential for indirect discrimination in a larger test dataset. Or, rather, a lower disparity would indicate a higher *probability* of indirect discrimination in a larger test dataset.

9.6.3.2 Considering the Composition of the Disadvantaged and Advantaged Group

The wording of the Equality Directives, which requires a "particular disadvantage" for indirect discrimination be found, is explicitly oriented towards the measurement of disadvantage, without mention of the advantaged group.[962] Given this wording, it is not evident that the aforementioned principle from the *Seymour-Smith* ruling is the only appropriate measurement method under the Equality Directives. As already mentioned, an alternative measurement method is to consider the composition of the disadvantaged group.

---

[959] Hacker alludes to a 75 % limit: Hacker (2018) 1153. Wachter, Mittelstadt and Russell

[960] Wachter, Mittelstadt, and Russell (2021 B) 22.

[961] Seymour-Smith, C-167/97, para. 61.

[962] Tobler (2008): 41.

However, is it sufficient to consider only the composition of the disadvantaged group, or should the composition of the advantaged group be considered as well?

At the time of submitting this thesis, the question of whether the determination of a "particular disadvantage" must be made with reference only to the disadvantaged group or to both the disadvantaged and the advantaged group has been raised in a request for preliminary ruling.[963] The request, which has not yet been processed by the CJEU or the Advocate General, concerns Directive 2006/54/EC (Recat Directive). This is the directive that repealed the directive on which the *Seymour-Smith* ruling is based. It defines discrimination with the same language as the RED and GSED, referring to a "particular disadvantage" in relation to indirect discrimination. The referring court asks how to determine if there is indirect sex discrimination in a case where a national collective agreement entails a better treatment of full-time employees in comparison with part-time employees.[964] Specifically, the referring court asks whether it is sufficient to find that the part-time employees (the disadvantaged group) are made up of considerably more women than men or whether it is necessary to consider also whether the group of full-time employees (the advantaged group) is made up of considerably more men than women.

These referred questions point towards two potential methods of measuring disadvantage at the group level, which differ from the method recommended by the CJEU in *Seymour-Smith:* considering the *composition* of the disadvantaged group and potentially also the advantaged group. If one considers the composition of the disadvantaged group, this will mean that a particular disadvantage is found if there are considerably more persons from a protected group than there are other persons in the disadvantaged group. Given the focus on a "particular disadvantage" in the wording of the Equality Directives, and the flexible, contextual approach taken by the CJEU, this is arguably a relevant method for measuring disadvantage. Moreover, if this method is accepted, it is also necessary to accept that the

---

[963] Request for a preliminary ruling from the Bundesarbeitsgericht (Germany) of 10 March, 2022, Ik V Kfh Kuratorium Für Dialyse Und Nierentransplantation, C-184/22; Request for a preliminary ruling from the Bundesarbeitsgericht (Germany) of 10 March, 2022, Cm V Kfh Kuratorium Für Dialyse Und Nierentransplantation, C-185/22.

[964] The payment of overtime supplements was available only for hours worked in excess of the standard working time of a full-time employee, which made it a lot more difficult to achieve such payments for part-time employees.

composition of the advantaged group should be taken into account in cases where there is unequal representation of protected groups in the dataset used for comparison purposes.[965] This is based on the reasoning that the CJEU applied in *Seymour-Smith*, even though the Court in that case focused on the *proportions* of advantaged/disadvantaged persons in a protected group.

9.6.3.3 Summary of Recognised Methods

Arguably, various approaches to the estimation of a particular disadvantage are acceptable under EU non-discrimination law, depending on the circumstances. Further below is a table outlining the approaches that should be considered in a pre-deployment discrimination assessment.

In relation to the disadvantage of resource denial, the measurement of disadvantage should conducted among cases identified as comparable in accordance with the criteria for comparability discussed in section 9.5. The measurement requires consideration of the outcome when an AI-CDS system is used on comparable cases. Because a "particular disadvantage" arises when persons from a protected group are disadvantaged in significantly higher proportions or numbers than other people, the outcome of interest is whether a patient is advantaged or disadvantaged by an AI-CDS system. The appropriate criteria for comparability between cases may depend on the clinical decision for which an AI-CDS system is intended and the different types of outputs it may generate. Similarly, different decision and output types may require that the group-level disadvantage is measured in different ways.

The following table outlines measurement methods and criteria that should be considered in a pre-deployment discrimination assessment, when assessing whether disadvantages in AI-CDS systems are particular to a protected group. The measurement methods proposed here would likely have been accepted by the CJEU in an ex post enforcement context. However,

---

[965] This is arguably also supported by the CJEU's reasoning in *Cachaldora Fernández*, where the CJEU dismisses the referring court's reasoning because it had failed to consider to what extent the disputed measure advantaged part-time workers. It had only looked at the extent to which part-time workers ended up in the disadvantaged pool: Judgment (GC) of 14 April, 2015, Cachaldora Fernández, C-527/13, ECLI:EU:C:2015:215, paras. 30-32.

adaptations are made as necessary to accommodate the specific context of assessing discrimination in AI-CDS systems before deployment. Especially, the table suggests that there might be a need to prioritise false negatives or false positives in some cases.

The following articulates the applicable measurement methods and criteria in a general manner. However, because technical methods for measuring performance in AI systems differ somewhat between classification models and regression models, the application of these approaches to the measurement of disparate performance as a disadvantage in the respective model types are discussed further below, in sections 9.6.4.1-9.6.4.2. Section 9.6.4.3 reflects on the specific considerations pertaining to the measurement of resource denial, which has been defined as a separate type of disadvantage, cf. section 9.3.2.

Method	Consideration	Criterion determining whether there is a "particular disadvantage"
Compare the number of persons from a protected group within the disadvantaged group, to the number of other persons in the disadvantaged group. **('Composition of the Disadvantaged Group' – CDG)**[966]	CDG is a generally applicable method when different groups are equally represented in the dataset used for comparison purposes.	Within the disadvantaged group, the number of protected group members is significantly higher than the number of members from other groups
Consider the CDG, in addition to comparing the number of persons from a protected group within the advantaged group to the	CDAG is appropriate when there is unequal representation of groups in the dataset used for comparison purposes.	If the same group also makes up a considerable proportion of the advantaged group, the particular disadvantage based on CDG is rebutted.

[966] This method is arguably supported by the "particular disadvantage" phrase in the Equality Directives' definition of indirect discrimination.

number of other persons in the advantaged group. **('Composition of the Disadvantaged and the Advantaged Group' – CDAG)**[967]		
Measure the probability of receiving an advantageous output for persons from a protected group and compare it to that of other persons. In practice, this means comparing the proportion of advantaged persons in one group with the proportion of advantaged persons in another group.[968] **('Probability Comparison')**	This method is based on what has been called the 'gold standard' for comparison. It is a generally applicable method that should be considered by default.	The probability of receiving an advantageous output is significantly lower for persons from a protected group compared to other persons. If the probability for persons from a protected group is < 60 % of other persons, there is clearly a particular disadvantage. If it is closer to 75 %, the assessment should be combined with qualitative considerations, such as whether the disadvantageous outputs are associated with stigma or prejudice.
Prioritising the false negative rate or the false positive rate.[969]	Supplementary method of measuring disadvantages of disparate performance in classification models, in cases where one error type is	A particular disadvantage is determined by the application of one of the abovementioned methods, adjusted for the additional weight given to false negatives or false positives in the assessment.

[967] The question has been raised in the preliminary ruling in C-184/22 and C-185/22 whether the composition of the advantaged group must be considered in addition to the composition of the disadvantaged group. In my opinion, this question should be answered affirmatively if there is unequal representation in the underlying population.

[968] This is, in my view, the most efficient way of expressing the formula that the CJEU recommended in *Seymour-Smith,* which was also applied, e.g., in *Brachner,* cf. section 9.3.1; Seymour-Smith, C-167/97, para. 59; Brachner, C-123/10, paras. 60-63.

[969] This aspect has not been highlighted in CJEU case law, but is a necessary adaptation to the context of AI-CDS systems.

	considerably more important than the other.	

Table 2: Methods of measuring disadvantages in AI-CDS systems and criteria for determining whether there is a "particular disadvantage".

In summary, the following considerations should be included in a pre-deployment discrimination assessment of an AI-CDS system based on the findings in this section regarding the identification of relevant measurement methods:

Methodological elements

Step 4: Identify relevant methods of measuring disadvantage

Unequal representation:

- □ Determine whether both the disadvantaged and advantaged groups need to be considered.
  - ➢ Is there unequal representation in the comparison dataset?

If unequal representation:

- □ Both the disadvantaged and advantaged groups need to be considered.
  - ➢ Relevant measurement methods include CDAG and Probability Comparison

If not:

- □ CDG is also a relevant measurement method

False positives vs. false negatives:

- □ Are false positives or false negatives more important, or are they equally important?

Implication:

- ➢ If one is more important than the other, adjust the relevant measurement methods by prioritising false negatives or false positives.

9.6.4 Applying the Relevant Methods in Practice

9.6.4.1 Measuring Disparate Performance in Classification Models Intended for Diagnostic Purposes

The following tables specifically illustrate how each of the abovementioned methods would play out in a measurement of the disparate performance in a classification model intended for diagnostic purposes. In classification models, model performance is usually provided in terms of accuracy, which is measured in terms of error rates. The disadvantage of disparate performance in classification models can be measured by considering the distribution of errors across groups. For the sake of illustration, the following tables consistently refer to examples where the question is whether men are at a particular disadvantage compared to women.[970]

Method of measurement	Criterion determining whether there is a "particular disadvantage"	Example of measurement
Composition of the Disadvantaged Group (CDG)	Within the disadvantaged group, the number of men is significantly higher than the number of women.	1000 patients receive either a false negative or a false positive (the disadvantaged group). Out of these, 800 patients (80 %) are men. If the threshold is set below 80 %, men are particularly disadvantaged. It may be necessary to consider also the composition of the advantaged group in cases where men and women are unevenly represented in the dataset used for comparison.

Table 3: Composition of the disadvantaged group – diagnostic classification model

Method of measurement	Criterion determining whether there is a "particular disadvantage"	Example of measurement

[970] When measuring a disadvantage incurred by an ethnic group, the main principles are the same. However, the measurement of disadvantage for ethnic groups is a more complex matter where certain nuances should be introduced, see section 9.7.

Composition of the Disadvantaged and the Advantaged Group (CDAG)	If men also make up a considerable proportion of the advantaged group, the particular disadvantage based on CDG is rebutted.	Consider a case where there are more men in the disadvantaged group, but there are also more men in the advantaged group. As before, 800 out of 1000 patients receiving an erroneous diagnosis are male. A particular disadvantage therefore appears when one only considers CDG (see above). However, assume that the total pool of patients within the dataset used for comparison consists of 100 000 patients. Out of these patients, 70 000 are men (70%) and 30 000 are women (30%). This unequal representation means that men are far more represented in the dataset than women. If 800 men and 200 women receive an erroneous diagnosis, this means that out of the 99 000 patients who receive a correct diagnosis (the advantaged group), there are 69 200 (~70%) and 29 800 women (~30%). In this scenario, men constitute 80 % of the disadvantaged group, but they also constitute 70 % of the advantaged group. Although the representation of men in the disadvantaged group is somewhat higher than in the advantaged group, the disadvantage that is established when looking only at CDG is arguably balanced out by the representation of men in the advantaged group.

Table 4: Composition of the disadvantaged and advantaged groups – diagnostic classification model

Method of measurement	Criterion determining whether there is a "particular disadvantage"	Example of measurement
Probability Comparison	The probability of being disadvantaged is significantly higher for	In a dataset of 200 000 patients (100 000 men and 100 000 women), a diagnostic decision is produced as an output for each individual. In the male group, there are

	men compared to women (which means that the percentage of men that are disadvantaged is significantly higher than the percentage of women that are disadvantaged).	5 000 diagnostic errors, which gives an error rate of 5%. In the female group, there are only 800 errors, which equates to an error rate of 0,8%. Consequently, the success rate for men is 95 %, while it is 99,2 % for women. The probability of a man receiving an advantageous (correct) output in this scenario is 95,77% of the probability of a woman being correctly diagnosed (95% / 99,2%). Given the threshold for a particular disadvantage, which probably lies between 60 % and 75 %,[971] the error rate for men would have to be extremely high for there to be a particular disadvantage based on this criterion. Interestingly, if one turns the question around and asks what the probability is of receiving a *disadvantageous* output, the numbers have a different appeal. Men's error rate of 5 % is actually 6,25 times the error rate of women (!).

Table 5: Probability of advantage/disadvantage – diagnostic classification model

Method of measurement	Criterion determining whether there is a "particular disadvantage"	Example of measurement
Prioritising the false negative rate or false positive rate[972]	A particular disadvantage is determined by the application of one of the abovementioned methods, adjusted for the additional weight given to false	An AI-CDS system is intended to diagnose a life-threatening disease where missing a truly positive case by producing a false negative output is a much more serious error than falsely identifying a healthy person as having the disease (false positive). This is because, in this example, a false negative might delay critical treatment, whereas a false positive would mainly cause temporary distress and require further confirmatory tests.

[971] As per the analysis in section 9.6.3.1.

[972] It should be noted, however, that the severity of the disadvantage is not relevant to the determination of a "particular disadvantage" according to CJEU case law.

	negatives or false positives in the assessment.	Consider a dataset consisting of 100 000 men and 100 000 women. Among the men, there are 6000 errors (a 6 % error rate). Among the women, there are 4000 errors (a 4 % error rate). Looking at the overall error rates, men have **1,5 times** the probability of women of being disadvantaged. Now, consider that the errors among the male patients consist of 2000 false positives and 4000 false negatives. In the female group, there are 3000 false positives and 1000 false negatives. In this scenario, the probability of a man receiving a false negative is 4 %. The probability of a woman receiving a false negative is 1 %. Consequently, men have **4 times** the probability of women of receiving a false negative. This shows that when specifically looking at false negatives, men are more likely to be disadvantaged, even though the overall error rate is less drastic. If the threshold for a "particular disadvantage" is set below 4,0 here, in relation to false negatives, men are particularly disadvantaged in this scenario.

Table 6: Prioritising the false negative rate or false positive rate – diagnostic classification model

9.6.4.2 Measuring Disparate Performance in Predictive Regression Models

In ML science and practice, the performance of regression and classification models, respectively, tend to be measured using different sets of evaluation metrics.[973] Regression models are particularly used in systems intended for treatment recommendations, predictive

[973] Techniques for measuring performance in ML-based models are typically categorised in terms of 'calibration' and 'discrimination.' In this context, the word 'discrimination' refers to a model's ability to distinguish between different classes. 'Calibration', on the other hand, refers to a category of techniques aimed at measuring "how well the predicted probabilities match the actual probabilities": Chen, Liu, and Peng (2019) 413.

interventions or resource allocations. In regression models, unlike classification models, error rates cannot be defined in terms of false positives and/or false negatives. This is because regression models predict continuous variables, not binary outcomes. As illustration, consider that a patient named Rebecca is considered for spine surgery using the NORspine AI-CDS system.[974] If the NORspine system gives Rebecca a 75 % probability of improving her condition through surgery, it is difficult to claim that the prediction is 'false,' regardless of how Rebecca's surgery turns out. After all, it must be expected that 25 % of the patients with that score will not improve.

Another important difference between the use of AI-CDS systems employing regression models versus those employing classification models, is that the *decision* will in most cases not correspond exactly to the AI-CDS system's output. For instance, an AI-CDS system for the diagnosis of coronary artery disease will indicate whether a patient has coronary artery disease (classification). This output corresponds directly to the clinical decision, which is to diagnose the patient with coronary artery disease or not.[975] However, a regression model such as in the NORspine example will not generate an output stating whether Rebecca should have surgery. While Rebecca's decision is a binary choice (to have or not to have surgery), the decision is based on non-binary outputs from the NORspine AI-CDS system, indicating a probability score or risk estimate, or perhaps a predicted score on the Oswestry Disability Index.[976] In the NORspine system, a probability score might suggest Rebecca's probability of improving from surgery, whereas a risk estimate might indicate the probability of Rebecca's condition worsening after surgery. Based on these outputs, Rebecca and the clinician can decide whether Rebecca should have surgery. Whether the decision is good depends on a scenario,[977] that is not known at the time the decision is made. This scenario serves a function similar to the 'ground truth' in relation to classification models.

---

[974] Section 5.1.

[975] Such a system might output a probability of the patient having coronary artery disease, but the point remains the same.

[976] Section 5.2.

[977] A 'scenario' here corresponds to what Hansson calls a 'state of nature': Sven Ove Hansson, "Decision Theory," *A brief introduction. Department of Philosophy and the History of technology. Royal Institute of Technology. Stockholm* (1994): 25.

In relation to the decision concerning Rebecca's spine surgery, there are two relevant scenarios at the time of the decision: the scenario where Rebecca would have improved from surgery and the scenario where she would not have improved from surgery. The situation is summarised by the following decision matrix:[978]

Scenario	**Surgery**	**No surgery**
Rebecca would have improved from surgery	Good decision	Bad decision
Rebecca would <u>not</u> have improved from surgery	Bad decision	Good decision

Table 7: Decision matrix based on the NORspine case study.

Which decision Rebecca and her clinician makes, may depend on Rebecca's personal preferences (e.g., how does she weigh the risk of having surgery against the probability of improving her condition?). Therefore, when testing the system based on retrospective data, the performance of the AI-CDS system cannot be evaluated based on the decision that is made, even if such data is available. Because the clinical decision is influenced by subjective factors beyond the model's control, like Rebecca's risk tolerance, the disadvantages of the system should be measured based on the outputs generated by the system, rather than the clinical decisions that are made.

This reasoning applies to post- as well as pre-deployment assessment contexts. In a pre-deployment setting there is no information about Rebecca's choice or outcome. In a post-deployment setting, there might be information about how Rebecca's condition actually developed, but it is difficult to tell how her condition would have developed if the decision had been the opposite of what it was. In other words, the scenario is difficult to measure, because it is hypothetical. Ideally, the dataset used for testing and assessment of the NORspine AI-CDS system would include information about what would have happened to

[978] The matrix is inspired by Hansson: Ibid.

each patient if an alternative choice had been made (counterfactual information).[979] However, when patients do not receive a certain clinical intervention, there is no counterfactual information about the outcomes had they received the intervention. Counterfactual information can be simulated, but it is questionable whether simulated counterfactual information for clinical interventions such as surgery are reliable.[980] This is an ongoing area of research within ML science.

Overall, performance measurement is a more complex issue in relation to regression models compared to classification models. While technical literature explains various ways of measuring the performance of regression models,[981] no consensus or gold standard has been established.[982] Several performance metrics may be used at the same time due to the complexity of measuring the performance of such models.[983] For example, during the training or model development phase, the accuracy of regression models is often measured based variations of the squared error between the predicted target and the ground truth for each patient.[984] When the ground truth is available, it is therefore feasible that the squared error can also be used to measure disparate performance between groups.

Finally, in relation to regression models, it should be noted that models that are trained and tested as regression models sometimes end up as classification models when they are deployed. For example, in breast cancer screening, tissue density is predicted on a continuous scale. During training and testing, the performance of a model may be assessed by means of the squared error between the predicted target and the ground truth. However, when designing an AI-CDS system, the continuous scale may be converted into, for example, four intervals

---

[979] Mattia Prosperi et al., "Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare," *Nature Machine Intelligence* 2, no. 7 (2020): 369, https://doi.org/10.1038/s42256-020-0197-y.

[980] Ibid.

[981] Max Kuhn and Kjell Johnson, "Measuring Performance in Regression Models," in *Applied Predictive Modeling* (New York: Springer, 2013).

[982] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman, "The Coefficient of Determination R-Squared Is More Informative Than Smape, Mae, Mape, Mse and Rmse in Regression Analysis Evaluation," *PeerJ Computer Science* 7, no. e623 (2021): 1, https://doi.org/10.7717/peerj-cs.623.

[983] This is recommended by Kuhn and Johnson: Kuhn and Johnson (2013) 95.

[984] Ibid; Chicco, Warrens, and Jurman (2021) 5.

(corresponding to four class labels), by radiologists.[985] Such conversion could make it easier for radiologists to apply the model's outputs in practice. In these cases, the performance of the AI-CDS system could be measured based on the error rates for the classifications outputted by the system or based on the accuracy of the regressions predicted by the underlying model (or based on a combination of the two).

9.6.4.3 Measuring Resource Denial (Regardless of Performance)

The performance of an AI-CDS system is important regardless of what type of clinical decision the system is intended for. Disparate performance is therefore a relevant disadvantage in relation to all clinical decisions. In addition to disparate performance, an important disadvantage to consider in relation to AI-CDS systems intended for the allocation of scarce resources is resource denial, cf. section 9.3.2.

As an illustration of this type of disadvantage, consider the difference between two case studies referred to in this thesis, i.e., an AI-CDS system intended to support decisions on spine surgery and an AI-CDS system intended to identify patients eligible for the Preventive Care Program at University Hospital of Storevik. Both systems employ regression models. However, the system for spine surgery is intended to be used in a context where *disparate performance* is the relevant measurement of disadvantage in a non-discrimination perspective; the disadvantage incurred by a patient directly corresponds to how good or correct the decision is. The important aspect of an AI-CDS system involved in making such a decision is therefore the performance of the system and whether the performance varies between groups.

In contrast, the AI-CDS system disputed in the case of Simon Tesafy v. UHS is used in a context where the output from the system dictates the allocation of a scarce resource which is desirable for Simon. If Simon is denied this resource, he is disadvantaged regardless of the accuracy of the algorithmic output and the correctness of the clinical decision relying on it.

---

[985] This is a type of feature engineering that is sometimes called 'binning' or 'bucketing': Theobald (2020) 43.

When measuring disadvantage in this AI-CDS system, the extent of resource denial at the group level should therefore be examined.

The general principles for the measurement of disadvantage are the same for resource denial as for disparate performance, with the caveat that a separate consideration of comparability is not necessary in relation to disparate performance. In both cases, a comparison should be conducted between protected groups by looking at them in terms of advantaged and disadvantaged individuals, and applying the measurement methods outlined in section 9.6.3 (see Table 2). For example, if one assesses the AI-CDS system used in UHS's Preventive Care Program, patients receiving a score rendering them eligible for the program constitute the advantaged group in relation to resource denial as a disadvantage. Patients receiving a score rendering them ineligible constitute the disadvantaged group in relation to this disadvantage.

Methodological elements

Step 5: Measure disadvantage in a test dataset

- □ Disparate performance
  - ? Is it a classification or regression model?
    - ➢ Apply state-of-the-art performance measurement techniques for the relevant model type
- □ Resource denial
  - ➢ Apply the method(s) identified in Step 2.4.
- □ Stigma/prejudice
  - ➢ If stigma/prejudice is likely to be caused or reinforced, the threshold for a particular disadvantage based on disparate performance or resource denial is lowered

9.6.5 Conclusion

When assessing discrimination in an AI-CDS system, the relevant methods of measuring the extent of disadvantage at the group level should be determined with regard to the specific clinical decision that an AI-CDS system is intended for. The ability to measure disparate performance is important in a pre-deployment discrimination assessment of all AI-CDS systems. Systems intended for the allocation of scarce resources should additionally be assessed for disadvantages pertaining to resource denial. Furthermore, it must be considered whether the system employs classification models or regression models, as there are different technical methods for measuring disparate performance in the respective model types.

This section has identified several methods for measuring disadvantage that appear to be compatible with EU non-discrimination law. Depending on the applied method, there are slightly different ways of articulating the threshold for a particular disadvantage. The relevant methods and their implied thresholds were articulated in section 9.6.3 in a language that was adapted to the specific context of AI-CDS systems. There are no fixed thresholds for when a "particular disadvantage" occurs in EU non-discrimination law. However, when applying the measurement method coined Probability Comparison in section 9.6.3, it was suggested that there certainly is a particular disadvantage if a protected group has less than 60 % of other persons' chances of receiving an advantageous output. It was argued that qualitative considerations might influence the threshold, particularly if the chance of an advantageous output is closer to 75 % of other persons' chances. A lower threshold could be applied in cases of scarce resource allocation, and in cases where a disadvantageous output might be associated with stigma or prejudice.

9.7 Comparison and Ethnic Minorities

9.7.1 Must the Disadvantage Pertain to a *Particular* Ethnic Group?

The general principles and methods for comparison and measurement of disadvantage developed in this chapter, are applicable to discrimination based on any protected characteristic,[986] including ethnicity and sex. However, comparison between patient groups defined by ethnicity is more complex than comparison between men and women, because there are more than two groups that may be considered. The GSED, which only prohibits sex discrimination, defines indirect discrimination as occurring when a practice exists that would put "persons of one sex at a particular disadvantage compared with persons of the other sex."[987] In contrast, the RED, which is solely concerned with ethnic discrimination, defines it as indirect discrimination when "persons of a racial or ethnic origin" are put at a particular

[986] Nuances in the methodology of comparison might exist due to the different natures of discrimination grounds. For instance, the identification of comparable persons in relation to age discrimination requires that one determines the age groups to compare. Moreover, in relation to discrimination based on disabilities, CJEU jurisprudence indicates that comparison can be made between persons with disabilities and persons without disabilities, but also between persons with a specific disability and persons with other disabilities, or even between persons with disabilities regardless of the exact types of disabilities they have.

[987] Article 2(b) GSED.

disadvantage "compared with other persons."[988] An important question is whether this means that a particular disadvantage should be found by

A) comparing one distinct ethnic group with other ethnic groups, to see if a *particular* ethnic group is put at a particular disadvantage; or
B) considering the ethnic composition of the disadvantaged and the advantaged group, to see if some ethnicities (e.g., ethnic minorities) are at a particular disadvantage compared to one or more other ethnicities (typically the majority group or a larger minority).

An isolated consideration of the wording of the definition of indirect discrimination in the RED does not yield a definitive solution to this issue. When the RED's definition of indirect discrimination refers to measures that would put "persons of a racial or ethnic origin" at a particular disadvantage, this may be interpreted as meaning *one particular* ethnic origin (Interpretation A).[989] A strictly isolated reading of the wording, uninformed by the objectives of the directive and the consequences of Interpretation could arguably support this interpretation, although legal scholars tend to deny such an interpretation of the wording.[990] For example, Atrey argues that that the phrase "racial or ethnic origin" is preceded by the indefinite article "a" which means "non-specific" or "any racial or ethnic origin."[991] Moreover, as I shall return to below, a more teleologically oriented interpretation also offers compelling arguments against Interpretation A.

9.7.2 CJEU Case Law

There exists case law where the CJEU has held that a disadvantage must pertain to persons of a *particular* ethnic origin, thus arguing in the direction of Interpretation A. In *Jyske finans,* a

[988] Article 2(2)(b) RED.

[989] In the Framework Directive (Directive 2000/78/EC), which does not cover ethnic discrimination, it is made clear in the definition of indirect discrimination that it pertains to particular disadvantages for "persons having a *particular* religion or belief, a *particular* disability, a *particular* age, or a *particular* sexual orientation" (my italicisation): Article 2(2)(b) Framework Directive.

[990] Shreya Atrey, "Race Discrimination in Eu Law after Jyske Finans: Case C-668/15, Jyske Finans a/S V. Ligebehandlingsnævnet, Acting on Behalf of Ismar Huskic, Judgment of the Court (First Chamber) of 6 April 2017 Eu: C: 2017: 278," *Common Market Law Review* 55, no. 2 (2018); Ward also denies that an isolated reading of the wording supports Interpretation A: Ward (2018) 49.

[991] Atrey (2018) 634.

financial institution required an additional proof of identification from persons whose driving license indicated a different country of birth that was not a member state of the EU or EFTA. Such persons were required to produce additional identification that was not required from persons born in an EU or EFTA state. The CJEU holds, first, that there was no direct discrimination because a person's country of birth could not be equated with ethnicity for the purposes of the RED.

In relation to the question of indirect discrimination, the CJEU holds that indirect discrimination under the RED can arise "only if the allegedly discriminatory measure has the effect of placing a person of a particular ethnic origin at a disadvantage."[992] On this point, the CJEU follows the opinion of Advocate General Nils Wahl.[993] The AG argues in his opinion that the provision on indirect discrimination requires identifying that persons of one or more *particular* ethnic origins are put at a disadvantage.[994] The AG denies the possibility of finding a particular disadvantage based on a comparison between one advantaged ethnic group and other persons (i.e., persons who are not of the advantaged ethnicity).[995] The implication of the AG's position in the specific circumstances of the *Jyske Finans* case, was that a particular disadvantage could not be established because the disputed PCP was equally disadvantageous towards all non-Danish persons.

The CJEU follows AG Wahl's opinion,[996] not only in *Jyske Finans*, but also in the subsequent ruling in *Maniero.*[997] That case revolved around a law student in Germany who had a degree from Armenia and who was excluded from applying for a scholarship because of a criterion requiring that applicants had completed a specific exam in the German legal education system. Referring to *CHEZ* and *Jyske finans*, the CJEU once again holds that "[t]he concept of 'particular disadvantage' […] must be understood as meaning that it is particularly persons of a particular racial or ethnic origin, because of the provision, criterion or practice in

---

[992] Jyske Finans, C-668/15, para. 31.

[993] Opinion by AG Wahl of 1 December, 2016, Jyske Finans, C-668/15, ECLI:EU:C:2016:914, para. 60.

[994] Ibid.

[995] Ibid.

[996] Liu and O'Cinneide (2019): 9.

[997] Maniero, C-457/17.

question, who are disadvantaged."[998] Like in *Jyske finans*, the CJEU therefore does not consider the impact of the disputed practice on ethnic minorities as compared to the ethnic majority group.

9.7.3 Criticism

The CJEU's approach to the search for a particular disadvantage in *Jyske finans* and *Maniero* is heavily criticised in non-discrimination law scholarship.[999] For instance, Atrey argues that the Court in *Jyske finans* court misses an opportunity to apply the RED in a way that would have furthered the objectives of combatting stereotyping and xenophobia.[1000] Moreover, Ward notes that the AG's statements in *Jyske Finans* (requiring a particular disadvantage pertaining to persons of a 'certain' ethnic origin), are made with reference to the *CHEZ* ruling, where "[p]recise identification of a group effect by a neutral practice was relevant […] due to its special facts".[1001] Ward correctly notes that the *CHEZ* ruling does not suggest that a disadvantage must pertain to a specific ethnic group.[1002]

In *Jyske finans*, the CJEU arguably fails to consider important differences between that case and the *CHEZ* case, which it relies heavily on. The CJEU cites its statement in *CHEZ* according to which "it is particularly persons of a given ethnic origin" who must be at a disadvantage and the measure must work "to the disadvantage of far more persons possessing the protected characteristic than persons not possessing it."[1003] However, in *CHEZ*, the disputed measure clearly disadvantaged one particular ethnicity (it seemed possible that the measure was targeted exactly against one ethnicity, but that was a matter for the referring court to consider). Therefore, there was no reason for the CJEU to consider other ways of identifying a disadvantaged, ethnic group. The CJEU could concentrate on determining

---

[998] Maniero, C-457/17, para. 47.

[999] Liu and O'Cinneide (2019): 9; Atrey (2018). In relation to the Maniero ruling, the CJEU's unwillingness to assess the ethnic composition of the advantaged and disadvantaged groups has been criticised by Wachter, Mittelstadt and Russell: Wachter, Mittelstadt, and Russell (2021 B).

[1000] Atrey (2018) 538-639.

[1001] Ward (2018) 49.

[1002] Ibid.

[1003] Jyske Finans, C-668/15, 27.

whether the disadvantage incurred by persons of Roma origin was 'particular'.[1004] In my view, the CJEU's statements in *CHEZ* do not indicate that a particular disadvantage must pertain to one specific ethnicity.

In the context of AI-CDS systems, the most alarming aspect of the approach taken by the CJEU in *Jyske finans* and *Maniero* is the implications for the RED's protection of ethnic minorities. For instance, disparate performance issues are likely to disadvantage ethnic minorities, especially because the performance of AI-CDS systems can vary according to how well different groups are represented in training data. If it is not permissible under EU non-discrimination law to compare ethnic minorities with the ethnic majority population, EU non-discrimination law is not well equipped to protect minorities against the disadvantages of disparate performance. If there are several ethnic minorities in a patient population, and an AI-CDS system performs poorly for all or many of them, there is no 'particular' disadvantage according to the reasoning from *Jyske finans.*[1005] This approach to disadvantage measurement is difficult to align with the minority protection ambition of EU non-discrimination law and the RED, specifically.[1006]

9.7.4 Conclusion

The CJEU's reasoning in *Jyske Finans,* and *Maniero* indicates that indirect discrimination can arise only if there is a particular ethnic group being particularly disadvantaged, in comparison to other ethnic groups. As a matter of *lex lata*, this therefore appears to be the current starting point for an assessment of potential indirect discrimination. However, considering the importance of teleological interpretation in EU law, it is not certain that the CJEU would, in all circumstances, deny the possibility of finding indirect discrimination based on consideration of the ethnic composition of the disadvantaged and advantaged groups. The CJEU's practice on this matter has been subject to strong criticism in the scholarly literature. To strengthen the protection against ethnic discrimination in the EU, especially in the context

---

[1004] Atreu emphasises how the identification of the disadvantaged persons was not an issue in *CHEZ: Atrey (2018) 633.*

[1005] Liu and O'Cinneide note that the approach is out of step with the approach adopted by the Grand Chamber of the ECtHR in Biao v Denmark: Liu and O'Cinneide (2019): 58; Judgment of the European Court of Human Rights (GC) of 24 May, 2016, Biao V. Denmark (Application No. 38590/10).

[1006] Recital 8 RED explicitly refers to "ethnic minorities," which can be taken as a sign that the Directive is intended to protect minority groups: Ellis and Watson (2012) 33.

of algorithmic discrimination in AI-CDS systems, it should be possible to find a "particular disadvantage" based on consideration of the ethnic composition of the advantaged and disadvantaged groups, such that systems having significantly higher error rates for ethnic minorities, compared to the majority group, can be assessed as potential indirect discrimination. To embrace such an approach, the CJEU would have to change course from its current jurisprudence.

It is arguable that a discrimination assessment is conducted in line with the approach taken in the CJEU's jurisprudence so far, if it involves measuring the disadvantage for particular ethnic groups rather than for ethnic minority patients as a whole. However, this choice has severe ramifications, making it unlikely that a particular disadvantage will be established during pre-deployment testing. Disadvantage would have to be measured based on comparable patients across all different ethnicities in the dataset. Consequently, there would be several small groups, which makes disadvantage measurement especially uncertain due to the small sample sizes. Additionally, it is not obvious how to associate patients with specific ethnicities, where as ethnic minority status can be more easily defined. If all these ramifications are taken into account, considering also the heavy criticism of the *Jyske finans* approach in academic literature, it seems likely that the CJEU would require that disadvantages pertaining to ethnic minority patients, as a group, must be taken into account when applying the non-discrimination principle in the specific context of AI-CDS systems. Thus, despite the uncertainty, it is recommendable that a discrimination assessment methodology based on the non-discrimination principle in EU law considers discrimination against ethnic minority patients as a group. Otherwise, there would for all practical purposes be no protection against ethnic discrimination in this context, and the AIA would have no chance of achieving its aim of ensuring effective protection of fundamental rights through pre-deployment discrimination assessments. This interpretation may not maximise loyalty to the existing jurisprudence of the CJEU, but it is loyal to the objectives of the Equality Directivs as well as the AIA.

9.8 Conclusion

EU non-discrimination law encompasses a broad range of disadvantages. Disparate performance is a type of disadvantage caused by AI-CDS systems that is universally important to consider as part of a pre-deployment discrimination assessment. In relation to AI-CDS systems intended for scarce resource allocation, the disadvantage of resource denial

must be considered, additionally. As a matter of discrimination assessment methodology, it is submitted that other types of disadvantages might influence the threshold for finding a particular disadvantage. Additionally, other types of disadvantages should be considered as part of the objective justification test.[1007] For example, when stigma and prejudice are caused or reinforced by AI-CDS systems, these are relevant disadvantages to consider when assessing whether a bias is objectively justified. While these disadvantages can, in principle, constitute a 'particular disadvantage' according to EU non-discrimination law, they do not lend themselves well to the quantitatively oriented measurement of disadvantage primarily applied in CJEU jurisprudence.

To prepare the ground for a pre-deployment discrimination assessment, the target variables of an AI-CDS system should be considered specifically, so that disadvantageous and advantageous outputs in relation to the target variables can be defined. For example, consider the AI-CDS system at issue in the fictional case of Simon Tesfay v UHS. This system identifies patients as eligible for the Preventive Care Program if the system predicts that a patient's health needs will exceed a certain threshold. In this case, because the Preventive Care Program can generally be deemed desirable for patients, the patients with predictions above the threshold can be defined as advantaged, whereas patients predicted below the threshold are disadvantaged. The latter group of patients represent those who would have been denied the scarce resource if the AI-CDS system had been deployed and its outputs relied on by clinicians.

The threshold for determining a particular disadvantage might arguably vary depending on the type of clinical decision and disadvantage in question. In relation to scarce resource allocation, it is arguable that a discrimination assessment should apply a low threshold, because of the serious consequences denial of resources can have. This would particularly be true in relation to life-critical matters such as intensive care triage or ventilator allocation. It was also suggested that the size of the test dataset, in which disadvantage is measured, should be considered. A disparity observed in a larger dataset arguably indicates higher probability of indirect discrimination, compared to when the same degree of disparity is observed in a smaller dataset.

[1007] Chapter 11.

The threshold for finding a particular disadvantage is an important issue for future legal and policy discussions. It is perhaps an issue that could be considered as part of efforts to standardise requirements for AI-CDS systems, in which case wide stakeholder participation would be crucial to enhance the legitimacy of such standardisation efforts. However, it is doubtful whether clear-cut thresholds should be standardised. This chapter's analysis of the thresholds and methods applied by the CJEU suggests that it is more important to highlight the quantitative and qualitative considerations and the relevant measurement methods that should be included in a pre-deployment discrimination assessment, so that assessors can make a reasoned decision on whether to deploy an AI-CDS system. It is also important to keep in mind that the discrimination assessment requirements discussed in chapter 7 emphasise probability and consequence. Defining a clear-cut threshold for when a disadvantage becomes 'particular' does not necessarily aid the assessment of how probable the occurrence of indirect discrimination in an AI-CDS system is.

In terms of the methodological elements of assessing discrimination in an AI-CDS system before deployment, this chapter has outlined certain ways of measuring disadvantages based on the non-discrimination principle in EU law. In summary, this chapter has developed the following methodological elements:

Comparison and Disadvantage Measurement

Objective: Determine whether the AI-CDS system would put a protected group at a particular disadvantage

Step 1: Determine how to construct an appropriate dataset for comparison purposes.

Intended distribution of system:

? Is the AI-CDS system intended to be distributed to different deployers and used without local modifications?

Implication:

- ➢ Include data collected from the geographical area(s) where the system is intended to be deployed.

Step 2: Determine which cases within the relevant test dataset are comparable to each other.

Criteria for comparability:

- □ Ground truth comparability
  - ➢ Technical method:
    - □ Search and compare
- □ Feature comparability
  - ➢ Technical methods:
    - □ Cluster analysis
    - □ Propensity score matching

Caveat for disparate performance:

- □ Comparability mapping not necessary to assess disparate performance

Step 3: Define the relevant disadvantages produced by the AI-CDS system:

- □ Disparate performance (always relevant)
- □ Resource denial
- □ Stigma and/or prejudice

Intended purpose:

? Is it intended that the system's outputs will be relied on to determine the allocation of scarce resources?

Implication:

- ➢ The disadvantage of resource denial should be measured, in addition to disparate performance.

Stigma/prejudice as consequence:

- □ Does the AI-CDS system display a type of bias that is likely to cause/reinforce stigma and/or prejudice?

Implications:

- ➢ Lower threshold for finding a "particular disadvantage"
- ➢ Take stigma/prejudice into account in the objective justification test

Step 4: Identify relevant methods of measuring disadvantage

Unequal representation in test dataset:

- □ Determine whether both the disadvantaged and advantaged groups need to be considered.
  - ➢ Is there unequal representation in the comparison dataset?

If unequal representation:

- □ Both the disadvantaged and advantaged groups need to be considered.
  - ➢ Relevant measurement methods include CDAG and Probability Comparison

If not:

- □ CDG is also a relevant measurement method

False positives vs. false negatives:

- □ Are false positives or false negatives more important, or are they equally important? (relevant to the measurement of disparate performance)

Implication:

- ➢ If one is more important than the other, adjust the relevant measurement methods by prioritising false negatives or false positives.

Step 5: Measure disadvantage in a test dataset

- □ Disparate performance
  - ? Is it a classification or regression model?
    - ▪ Apply state-of-the-art performance measurement techniques for the relevant model type
- □ Resource denial
  - ▪ Apply the method(s) identified in Step 2.4.
- □ Stigma/prejudice
  - ▪ If stigma/prejudice is likely to be caused or reinforced, the threshold for a particular disadvantage based on disparate performance or resource denial is lowered

10 Causation

10.1 Introduction

10.1.1 The Role of Causation in EU Non-Discrimination Law

There are two different causation requirements in EU non-discrimination law: one for direct discrimination, and one for indirect discrimination. For direct discrimination, the requirement is that a person must be treated less favourably than another "on grounds of" a protected characteristic.[1008] Thus, causation in relation to direct discrimination refers to the connection *between a protected characteristic and a disadvantageous treatment*.

In contrast, the prohibition on indirect discrimination is not oriented towards the characteristics relied on in a decision-making process, but rather the effects of that process. Indirect discrimination occurs where the disputed practice "would put" a protected group at a particular disadvantage. Thus, the causation requirement for indirect discrimination relates to the connection *between the disputed practice and a disadvantage incurred by a group*. Due to the distinct orientations of the causation requirements for direct and indirect discrimination, they are treated separately in this chapter.

10.1.2 Purpose and Structure of the Chapter

The purpose of this chapter is to develop methodological elements pertaining to the application of causation requirements for both direct and indirect discrimination in the context of assessing discrimination in an AI-CDS system prior to its deployment. With regards to the assessment of potential direct discrimination, this chapter aims to develop methods and principles for determining whether an AI-CDS system uses a Protected characteristic or Inextricably Linked Factor (PILF) in such a manner that it amounts to direct causation under EU non-discrimination law.

According to the methodological elements of discrimination assessment outlined in this this thesis, a direct discrimination assessment is relevant in cases where a Feature Identification process (as described in chapter 8) indicates that a model employed within an AI-CDS system likely incorporates one or more PILFs among its feature variables. If this is found to be the case, chapter 8 concluded that one should proceed with assessing whether there exists a causal

[1008] Article 2(2)(a) RED; Article 2(a) GSED.

connection between any PILFs and a model's outputs which is sufficient to establish direct causation under the non-discrimination principle. If a sufficient causal connection is established between a PILF and a model's outputs, this indicates that patients who receive disadvantageous outputs (see section 9.3) from the system are being directly disadvantaged "on grounds of" a PILF, in comparison to patients who receive advantageous outputs.[1009]

In relation to indirect discrimination, this chapter seeks to develop the methodological elements of determining whether a group-level disadvantage is sufficiently linked to a biased AI-CDS system so that indirect discrimination must be attributed to the system. These disadvantages must be distinguished from other disadvantages which may be observed in a pre-deployment assessment context without being taken as indications of indirect discrimination attributable to the system being assessed. However, this chapter finds that the causation assessment in relation to indirect discrimination overlaps with the objective justification test. It is concluded that the considerations that arise as part of causation assessment in relation to indirect discrimination are best handled as part of the objective justification test, which is explored in chapter 11. The largest part of this chapter is devoted to the causation assessment in relation to *direct* discrimination.

This chapter proceeds as follows. Section 10.2 establishes certain fundamental differences between causation assessment in an ex post enforcement context and the pre-deployment assessment context that is primarily contemplated in this thesis. Section 10.3 provides relevant context to the subsequent legal analysis of the causation requirements by briefly illuminating the role, as well as the controversy, of causal reasoning in clinical decision-making and machine learning. It is important to recognise that non-discrimination law requires certain considerations which are framed as a matter of 'causation' even though ML researchers sometimes warn that asking about causal relationships could lead to flawed reasoning where correlation is mistaken for causation. Sections 10.4 and 10.5 analyse the causation requirements for direct and indirect discrimination, respectively, and develop methodological elements of pre-deployment discrimination assessment based on the analyses.

[1009] According to Henrard, it is arguable that a differentiation based on a characteristic inextricably linked with a protected ground should automatically be categorised as direct discrimination: Henrard (2019) 106-07. See also Ellis and Watson (2012) 163-69.

10.2 Causation Assessment in Ex Post Enforcement Contexts vs. Before Deployment

In an ex post enforcement context, if a person is disadvantaged compared to another person "on grounds" of a PILF, this means that direct discrimination has occurred. However, the ex post assessment of discrimination only considers how one person – the alleged victim of discrimination – is treated in comparison with others. The decisive question, ex post, is whether a PILF has been sufficiently influential on the treatment that this one person has received. In contrast, in a pre-deployment discrimination assessment, it is necessary to examine the influence of PILFs at the system level. Given that the probability of discrimination is a central consideration according to the pre-deployment discrimination assessment requirements identified in chapter 7, the causation assessment should consider whether an AI-CDS system is influenced by PILFs to such a degree that direct discrimination is likely to occur if the system is deployed.

Methodological element

Objective (direct discrimination assessment):

The objective of *direct* causation assessment in a pre-deployment discrimination assessment is to determine whether an AI-CDS system is influenced by PILFs to such a degree that patients disadvantaged by the system would be directly discriminated against.

10.3 Causal Reasoning in Clinical Decision-Making and Machine Learning

There is an ongoing scientific discourse about the epistemological ramifications of introducing AI systems into medical practice and research. It has been suggested that AI systems based on ML introduce a shift from causally oriented reasoning to a more correlational type of reasoning.[1010] ML-based models are generally designed to identify correlations rather than causal relationships.[1011] Correlation means that events occur together

[1010] Tal Z Zarsky, "Correlation Versus Causation in Health-Related Big Data Analysis," *Big Data, Health Law, and Bioethics* (2018): 43 and 51.

[1011] Coglianese and Lehr (2016) 1157; Zarsky (2018); Richens, Lee, and Johri (2020).

even if this happens by co-incidence.[1012] Thus, correlation does not inherently imply causation.

In contrast, traditional medical reasoning often focuses on understanding the causal relationships between risk factors, diseases, and treatments. Clinicians seek to understand the underlying mechanisms through which diseases progress and treatments work, using this knowledge to inform their decisions. Some clinical tasks are directly aimed at identifying the cause of health-related events or diseases. For instance, a diagnosis can be defined as a statement about the *cause* of the symptoms that a patient experiences.[1013] The cause is inferred from the observed symptoms (the effects of the disease).[1014] In relation to treatment recommendations and preventive interventions,[1015] the causal reasoning takes a slightly different shape. The premise for those decisions is that a treatment or intervention is *expected to cause an improvement* in a patient's health, compared to alternative scenarios where the treatment or intervention is not provided.[1016]

One risk of substituting causal medical reasoning with correlation-oriented AI systems is that correlations can potentially be mistaken for causal relationships.[1017] ML algorithms are primarily pattern-finders. By default, unlike a human medical expert, algorithms do not use contextual information to determine whether a relationship between two pieces of information is causal or merely correlational. Caruana et al. illustrate this challenge by reference to a model intended to predict the mortality risk of pneumonia patients.[1018] This model indicated, based on the observed data, that pneumonia patients with asthma have a lower risk of dying

---

[1012] Cristian S Calude and Giuseppe Longo, "The Deluge of Spurious Correlations in Big Data," *Foundations of science* 22, no. 3 (2017): 597, https://doi.org/10.1007/s10699-016-9489-4.

[1013] Richens, Lee, and Johri (2020) 2.

[1014] Ibid.

[1015] Section 1.7.

[1016] On causal reasoning in relation to interventions, see: Prosperi et al. (2020) 369.

[1017] Alsaleh et al. note the potential for incorrect or misleading predictions if ML algorithms are not sufficiently validated and transparently communicated to healthcare providers: Mohanad M. Alsaleh et al., "Prediction of Disease Comorbidity Using Explainable Artificial Intelligence and Machine Learning Techniques: A Systematic Review," *International Journal of Medical Informatics* 175, no. 105088 (2023): 7, https://doi.org/10.1016/j.ijmedinf.2023.105088.

[1018] Caruana et al. (2015) 1721.

than other patients with pneumonia. This was based on the finding of a correlation between having asthma and actual patient outcomes. However, what the model did not account for, was the important contextual information that patients with asthma in this dataset were routinely admitted to the intensive care unit and, thus, followed up very closely once they were hospitalised.[1019] If the correlation between asthma and survival is mistaken for causation, one would assume that asthma causes lower risk for pneumonia patients. However, this would be a flawed and dangerous assumption. When better outcomes are observed for pneumonia patients with asthma, this is because health systems normally assume that these patients are actually at higher risk than others and they therefore receive additional attention. This contextual information may not be captured by the data used to train an AI system.

In ML science, it has been suggested that ML-based models should not be generalised beyond the specific context and population reflected in the training data, if the correlational patterns that the models rely on are not understood.[1020] The challenge of understanding the correlational patterns is being approached through attempts of developing 'causal machine learning' techniques, the purpose of which is to increase the understanding of causal relationships between variables in a model.[1021]

Although causal machine learning is a distinct subcategory of ML science, it is related to the broader category of explainable AI (XAI). The latter aims to develop techniques to make AI models interpretable and to provide understandable explanations of how a model works or why it produces a certain output.[1022] The combination of causal machine learning and XAI opens up the potential for causal explanations, i.e. explanations of causal relationships between variables and between variables and outputs. The present chapter's analysis of causation requirements in EU non-discrimination law can inform future technical efforts in

---

[1019] Ibid.

[1020] Zarsky (2018) 43 and 51.

[1021] Causal machine learning refers to a subfield of machine learning that focuses on understanding and modelling causal relationships between variables, rather than merely identifying correlations or associations: e.g., Richens, Lee, and Johri (2020); Pedro Sanchez et al., "Causal Machine Learning for Healthcare and Precision Medicine," *Royal Society Open Science* 9, no. 220638 (2022), https://doi.org/10.1098/rsos.220638.

[1022] e.g., Alsaleh et al. (2023).

causal explanations, by indicating which causal connections one needs to understand in order to assess discrimination in an AI-CDS system.

10.4 Direct Algorithmic Discrimination "On Grounds Of" a Protected Characteristic

10.4.1 Introduction

A finding of direct discrimination requires that a person must be treated less favourably than another "on grounds of" a protected characteristic.[1023] It is not required that the decision-maker intends to treat someone less favourably on grounds of a protected characteristic, or that the decision-maker is otherwise motivated by a protected characteristic,[1024] or aware of the circumstance that someone is treated less favourably due to a protected characteristic.[1025] However, if an intention or motivation is found, this would remove any doubts with regards to whether direct causation is present.[1026] Hence, the "on grounds of" requirement refers to an objective test that does not require a conclusion about the mental process of a decision-maker.

However, the wording in the Equality Directives' definition of direct discrimination leaves important questions about the substantiation of the required connection open to interpretation. In relation to a pre-deployment discrimination assessment of an AI-CDS system, the question arises how influential a feature variable in a model must be, for the model's outputs to be produced "on grounds" of it. Moreover, how can an assessor establish the degree of influence of a given feature variable? The following sections address these questions in order to develop considerations, methods and principles which may be applied in a pre-deployment discrimination assessment of an AI-CDS system.

10.4.2 Clarification Regarding Direct Algorithmic Discrimination Based on Effects

Direct discrimination occurs where a person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of a protected characteristic.[1027] The

---

[1023] Article 2(2)(a) RED; Article 2(a) GSED.

[1024] Ellis and Watson (2012) 163-64.

[1025] Hepple uses the term 'unconscious discrimination' to denote discrimination that the perpetrator is not aware of: Hepple (2011) 56.

[1026] Ellis and Watson (2012) 163-67.

[1027] Article 2(2)(a) RED; Article 2(1) GSED.

clearest cases of direct algorithmic discrimination occur where AI systems rely directly on protected characteristics as feature variables.[1028] However, as noted in chapter 8, the prohibition on direct discrimination has been applied expansively by the CJEU. Chapter 8 concluded that it is possible that algorithmic discrimination can be found based on the effects of an AI-CDS system on a protected group, even if the system does not include any protected characteristics as feature variables.

In the light of this finding, it is worth clarifying here that the following discussion of the causation requirement for direct discrimination does not concern cases where direct discrimination is found based on the effects of an AI-CDS system. If a pre-deployment discrimination assessment finds that an AI-CDS system has the effects discussed in chapter 8 (either exclusively disadvantaging persons from a protected group or entirely excluding a protected group from receiving an advantage), this would indicate a potential for direct discrimination regardless of which feature variables the system relies on. Consequently, there would be no need for a separate causation assessment. Regardless of the exact type of pre-deployment assessment one relies on, such a system would arguably not be deployable due to the high likelihood of direct discrimination.

10.4.3 Relationship Between Causation and the Notion of an 'Inextricable Link'; 'Mutually Exclusive Features' as the Typical Situation in CJEU Case Law

In some cases that have appeared before the CJEU, the causal relationship between reliance on a certain factor and the outcome of a decision is clear.[1029] In such cases, a discrimination claim hinges not on the question of causation between a certain feature and a disadvantageous treatment, but rather on whether the feature that 'caused' the disadvantageous outcome is so closely related to or correlated with a protected characteristic that it must be seen as tantamount

[1028] See chapter 8.

[1029] For instance, in the *Ingeniørforeningen i Danmark* case, a Danish law provided that some workers were, upon the termination of employment, entitled to receive an old-age pension from their employer. The decisive criterion was whether a worker was entitled to severance payment according to a pension scheme which they joined before the age of 50. In that case, it was "apparent from the documents before the court" that entitlement to an old-age pension was subject to a minimum age requirement and that there was a "difference of treatment based directly on grounds of age": Ingeniørforeningen I Danmark, C-499/08, para 23.

to relying on a protected characteristic. This question was dealt with in chapter 8, where it was framed as a question of whether a feature variable in an AI-CDS system is 'inextricably linked' to a protected characteristic. In relation to a pre-deployment discrimination assessment, chapter 8 outlined the criteria determining whether there is an inextricable link and how this question can be dealt with through the process that chapter 8 defined as Feature Identification. Feature Identification is the first step of a pre-deployment discrimination assessment, according to the methodological elements developed in this thesis.

The CJEU does not clearly distinguish between the question of whether there is a sufficient influence from a given factor and the question of whether a given factor is inextricably linked to a protected characteristic. In the *Hertz* ruling there is no discussion about whether there has been a sufficient influence from a certain factor.[1030] Instead, the ruling revolves around whether a factor that was undoubtedly decisive for how the alleged victim in that case was treated, should be seen as sufficiently closely linked to sex. The claimant had been dismissed by her employer due to absence from work. This absence was due to illness owing to maternity complications. The CJEU had the option to consider this as a dismissal caused by *illness* or as a dismissal caused by *pregnancy*. Pregnancy is inextricably linked with biological sex according to established case law.[1031] Therefore, if there was causation between pregnancy and the decision to dismiss an employee, there would have been direct discrimination. In contrast, the CJEU does not find discrimination in *Hertz,* because it deems the dismissal in the case to be 'caused' by illness, and it does not deem illness to be inextricably linked to sex.

Several examples of the same type of situation are found in age discrimination cases, where the question may be posed as whether causation should be attributed to the age of an employee or to another criteria, typically one referring to the length of experience or other time-related factors.[1032] For example, the *Tyrolean Airways* ruling concerns age discrimination under the Employment Equality Directive.[1033] Here, the CJEU finds that causation is missing because it

---

[1030] Judgment of 8 November, 1990, Hertz, C-179/88, ECLI:EU:C:1990:384.

[1031] Dekker, C-177/88; Dekker, C-177/88.

[1032] Tyrolean Airways, C-131/11; Horgan and Keegan, C-154/18; Ingeniørforeningen I Danmark, C-499/08.

[1033] Tyrolean Airways, C-131/11.

accepts the argument that the contested practice was based on a criterion different from age.[1034] Specifically, this ruling concerns a provision that categorised workers into different groups, determining their level of pay accordingly. This provision considered only the professional experience gained as a cabin crew member of a specific airline and disregarded experience obtained from working with other airlines. Although the CJEU finds that the provision in question was "likely to entail a difference in treatment according to the date of recruitment," that difference is not deemed to be based on age or on an "event linked to age."[1035] Thus, the CJEU accepts that the disadvantage was caused by an alternative factor not inextricably linked to a protected characteristic. Similarly, in *Horgan and Keegan*, another age discrimination case under the Employment Equality Directive, the CJEU finds that the only criterion in play in the disputed practice was the date of recruitment, "an objective and neutral factor […] manifestly unconnected to any taking into account of the age of the persons recruited."[1036]

The issue presented before the CJEU in the abovementioned case law can be categorised as one involving 'mutually exclusive features.' This refers to a situation in which an unfavourable treatment can only be attributed to one of two or more features, as these features are interdependent to such an extent that a decisional outcome cannot be attributed to both (or all) of them. In the *Hertz* case, the unfavourable treatment could be attributed to either illness or pregnancy. However, it was evident that the illness was caused by the pregnancy, making it illogical to attribute the unfavourable treatment to both pregnancy and illness.[1037] Consequently, the CJEU was tasked with determining to which feature the treatment should

---

[1034] Although the questions referred to the CJEU assumed that there was a case of potential indirect discrimination, the CJEU's reasoning around causation is more relevant under the rule against direct discrimination. The ruling arguably displays an ambiguous approach to the distinction between direct and indirect discrimination.

[1035] Tyrolean Airways, C-131/11, para. 29.

[1036] Horgan and Keegan, C-154/18, para. 25. As in *Tyrolean Airways*, the CJEU received questions based on the assumption that the cases would be assessed as potential indirect discrimination. Yet, the Court's reasoning around causation relies on a way of thinking that only makes sense as a matter of potential direct discrimination. In both cases, the decisive point for the CJEU is that the practices in question relied on other criteria than the protected characteristic on which the discrimination claim was based.

[1037] Similarly, features related to religion can sometimes be mutually exclusive with ethnicity. Hepple mentions a case from UK case law (the *JFS* case), where the criterion relied on was that of a matrilineal connection to Orthodox Judaism: Hepple (2011) 58.

be attributed.[1038] When one feature (maternity) is inextricably linked to a protected characteristic (sex), and the other feature (illness) is not, the choice of attributing a differential treatment to one of those features becomes decisive for the question of discrimination.

The central issue in the cases of *Hertz, Tyrolean Airways,* and *Horgan and Keegan* is essentially whether a factor, which plays a decisive role in the treatment of an individual, is adequately linked to a protected characteristic to give rise to a claim of direct discrimination. This question falls within the purview of the Feature Identification phase of a pre-deployment discrimination assessment, conducted in accordance with the methodological elements outlined in chapter 8. Therefore, a pre-deployment assessment of direct discrimination in an AI-CDS system should initially concentrate on identifying the features that influence a model's outputs (Feature Identification).[1039] Once it has been established through Feature Identification whether a model includes any PILFs, the next step is to assess the causal connection between these PILFs and the outputs of the AI-CDS system. To develop the principles or criteria according to which direct causation should be established, the following sections further interpret the causation requirement under EU law's prohibition on direct discrimination.

10.4.4 How Much Must a Feature Influence a Model's Outputs?

It is conceivable that models in AI-CDS systems may incorporate mutually exclusive features. However, algorithmic decision-making often involves more intricate data processing than the cases previously encountered by the CJEU. In certain instances, a learning algorithm may discover that Feature A, which triggers the occurrence of Feature B, also possesses independent predictive value alongside Feature B.[1040] Consequently, both features would exert an influence on the output, thereby making an unfavourable treatment partially attributable to both feature variables.

The question of whether a disadvantage is caused by a PILF or an alternative factor, is central in the context of decisions where one particular factor is decisive for how a person is treated. However, compared to the practices considered by the CJEU in the case law cited in the

---

[1038] Hellborg emphasises the importance of assessing whether it was the protected characteristic or something else that caused the disputed treatment: Hellborg (2018) 298-99.

[1039] Chapter 8.

[1040] In machine learning science, there are various methods for investigating the relationship between feature variables: e.g., Theobald (2020) 61.

previous sections, AI-CDS systems regularly rely on larger numbers of feature variables, none of which can be disregarded as non-influential on the outputs produced by the system.

Given that AI-CDS systems typically involve multiple feature variables, two crucial questions for the development of a discrimination assessment methodology aimed at these systems are the following:

(1) is it sufficient that a PILF is relied on or is it required that its influence exceeds a minimum threshold?

(2) if the influence must exceed a certain threshold, how significant must the influence be for direct causation to be established?

The first question is whether an output is generated "on grounds of" a feature when that feature is merely taken into account, or whether it is required that the feature must be given at least a modicum of weight. The wording of the definition of direct discrimination in the Equality Directives does not resolve this matter and the question does not tend to come up in CJEU rulings, because the typical cases that have emerged concern competition between mutually exclusive features. On the one hand, the highest protection against discrimination would be achieved through a rigid interpretation according to which direct causation is deemed to be present for any factors that influence the outcome of a decision, no matter how modestly.[1041] On the other hand, due to the many feature variables potentially present in an AI-CDS system, minor influences from PILFs might be a regularly occurring incident in these assessments. If any presence of PILFs should be taken as an indication that a model may lead to direct discrimination, this could make it unreasonably difficult for assessors to decide on deploying an AI-CDS system.

There is no clear answer to the question of whether EU non-discrimination law requires a modicum of weight to be given to a PILF, for causation to be established. Notably, this issue has been discussed in literature concerning UK anti-discrimination law, in which the question of whether a protected characteristic has a 'significant influence' on the outcome is

[1041] According to Henrard, it is arguable that a differentiation based on a characteristic inextricably linked with a protected ground should automatically be categorised as direct discrimination: Henrard (2019) 106-07.

relevant.[1042] In the UK context, the existence of such a significant influence has sometimes been established by asking whether changing a person's protected characteristic would change the decisional outcome. This approach is often referred to as the 'but for' test.[1043]

While CJEU has not explicitly given the 'but for' test the status as the gold standard for direct causation assessment in EU law, academic literature often assumes that it is.[1044] Indeed, there are certain traces of such a test in the CJEU's jurisprudence. One ruling that appears to support the application of a 'but for' test is the *Sarah Margaret Richards* case.[1045] This ruling, provided on the basis of Directive 79/7EEC, concerns a national law that would not recognise, in the context of retirement pension criteria, the new gender of a person who had undergone gender reassignment surgery (from man to woman). According to the national law at issue, women born before a certain year were eligible for a retirement pension at the age of 60, whereas the retirement age for men was 65.[1046] The CJEU states in paragraph 38 of the ruling:

> It is clear from the foregoing that Article 4(1) of Directive 79/7 must be interpreted as precluding legislation which denies a person who, in accordance with the conditions laid down by national law, has undergone male-to-female gender reassignment entitlement to a retirement pension on the ground that she has not reached the age of 65, *when she would have been entitled to such a pension at the age of 60 had she been held to be a woman* as a matter of national law." (my italicisation)

[1042] Adams-Prassl, Binns, and Kelly-Lyth (2023) 169 . Connolly also notes that UK courts have sometimes applied the requirement that a feature must be a substantial factor in the decision: Connolly (2011) 89.

[1043] Ellis and Watson (2012) 167.

[1044] Ruth Nielsen, *Civilretlige Diskriminationsforbud* (København: Jurist- og Økonomforbundets Forlag, 2010), 284-85. ("… det forbudte kriterium skal have haft så stor betydning, at det må antages, at den diskriminerede ville være blevet behandlet anderledes, hvis hun eller han ikke havde tilhørt den gruppe, som diskriminationskriteriet utpeger."); Ellis and Watson also assume that this is the correct approach under the Equality Directives: Ellis and Watson (2012) 167. (… "whether the alleged discrimination is direct or indirect, all that is required to establish causation is that 'but for' belonging to the protected category, the victim would not have sustained the disadvantage alleged."); Liddell and O'Flaherty (2018) 49-50; Hellum and Strand (2022) 251.

[1045] Judgment of 27 April, 2006, Sarah Margaret Richards, C-423/04, ECLI:EU:C:2006:256.

[1046] Sarah Margaret Richards, C-423/04, paras. 15-16.

According to this statement, the decisive point is that, had the claimant not been taken for a man for the purposes of the retirement pension, she would have received the pension.[1047] This is a 'but for' test, because it is asked what the result would have been 'but for' a specific fact.

There is no consistent line of CJEU case law manifesting the 'but for' test as the only relevant causation test in EU non-discrimination law (nor is it necessarily seen as the gold standard for direct causation in UK law).[1048] However, whenever a person would have been treated more favourably (as favourably as others), had it not been for the protected characteristic at issue, the CJEU seems likely conclude that there is a less favourable treatment "on grounds of" the protected characteristic.[1049]

Arguably, the 'but for' test implies that a modicum of influence is required before direct causation can be found between a given characteristic and the treatment provided to a person. It is not sufficient to determine that a PILF is somehow involved in the making of a decision, without considering the degree of the PILF's influence on the decision. However, the 'but for' test, as applied by the CJEU, is a highly case-specific test. It is a test that has been articulated with the individual ex post enforcement context in mind. In this context, the question is whether changing a protected characteristic of an individual would have changed the treatment of this individual, given all the circumstances of the specific case. Hence, it does not yield directions which can easily be generalised and applied in a pre-deployment assessment context. Particularly, it does not convey how significant the influence from a PILF must be *at the level of a system or model,* for direct causation to arise.

Before discussing how a 'but for' test may be adapted to a pre-deployment assessment context, it is worth emphasising that a finding of direct causation between a PILF and the outputs of a model may have different implications depending on the type of disadvantage one is considering. The reason for this is that, in relation to the disadvantage of disparate performance, there is arguably no disadvantage caused by a PILF if the use of the PILF is

---

[1047] The ruling is also referred to by Liddell and O'Flaherty in support of the 'but for' test that the authors find to be applicable in EU non-discrimination law: Liddell and O'Flaherty (2018) 50-51.

[1048] Connolly (2011) 95.

[1049] This has implicit support also in, e.g., the *Roca Álvarez* ruling. Here, the CJEU considers the counterfactual situation if the claimant had been a woman rather than a man: Judgment of 30 September, 2010, Roca Álvarez, C-104/09, ECLI:EU:C:2010:561, para. 23.

Methodological elements

Consideration:

? Which type of clinical decision is the system intended for?

□ Diagnosis, treatment recommendation, or preventive intervention;
  - ➢ The degree of influence from PILFs should be justified by the medical justification for relying on PILFs in the specific clinical decision

□ Allocation of scarce resources
  - ➢ The non-discrimination principle does not permit direct causation between PILFs and disadvantageous outputs

backed by an adequate medical justification. The finding of direct causation between a PILF and a model's outputs should therefore trigger an assessment of the medical justification for including the PILF in the model. In contrast, in relation to the disadvantage of resource denial, patients who are denied a resource are disadvantaged regardless of whether there are perfectly good medical reasons for the decision. Therefore, the implication of finding direct causation in a system intended for scarce resource allocation is that the system is highly likely to lead to direct discrimination. Given the severity of direct discrimination, it may generally be assumed that a finding of direct causation in such an AI-CDS system will lead to the conclusion that the system should not be deployed.

10.4.5 Adapting the 'But For' Test to the Context of Pre-Deployment Assessments

When the 'but for' test is applied in an ex post enforcement context, the question is essentially whether an individual would have been treated more favourably if their protected characteristics had been different. In a pre-deployment assessment context, the purpose of assessing causation is to decide on deployment of an AI-CDS system. This is a decision that must be made based on an assessment of the system, as such. It is not based on the claim of an individual patient and, thus, it is not limited to the scrutinization of one case. It is not sufficient to run a model on data representing one patient in a test dataset and consider whether changing this patient's characteristics leads to a different output. However, with each patient in a dataset whose output changes when a protected characteristic is changed, one could say that the probability of direct discrimination increases.

In practice, if a decision hinges only on a few factors and a PILF is one of them, it is more likely that the outcome of a decision would have changed if the PILF was changed, compared to the situation where there are hundreds of feature variables in an AI-based model. The 'but for' test suggests that the relative importance of a PILF should be considered but does not provide a generalisable threshold to determine if direct causation is present.

Methodological element

Consideration:

- ? What is the relative importance of PILFs within a model?
  - □ Technical method: counterfactual testing and explanations
    - ➢ E.g., Individual Conditional Expectation, Partial Dependence Plots, or CaCE

However, the CJEU's case law arguably points out a general direction for the assessment: The 'but for' test fundamentally relies on counterfactual reasoning, by positing that an event A can be said to cause (or at least influence) another event B, if and only if event B would not have occurred in the absence of event A.[1050] Hence, a pre-deployment assessment should be oriented towards determining the extent to which alterations of features that constitute PILFs lead to changes in the outputs produced by a model. This could be attempted through counterfactual testing, which would involve the simulation of many test cases where PILFs are altered for each case.[1051] It is recognised in ML literature that even small changes can alter the outputs from a model considerably.[1052]

[1050] The counterfactual theory, often attributed to David Hume's "An Enquiry Concerning Human Understanding," has been debated over the years and its popularity has varied: David Lewis, "Causation," *The journal of philosophy* 70, no. 17 (1974): 557; Mario Bunge, *Causality and Modern Science*, 4th ed. (New York: Routledge, 2017), xxi.

[1051] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the Gdpr," *Harvard Journal of Law & Technology* 31, no. 2 (Spring 2018): 851-53.

[1052] Bambauer, Zarsky, and Mayer (2021) 2340.

In practice, the model being assessed could be run on a large number of patients' data as well as counterfactual versions of the same data.[1053] Each test case could be based on data representing an actual patient. Counterfactual versions of this case could then be created as modified duplications of the data from the actual patients. In the modified duplications, PILFs pertaining to a patient, for example ethnicity, should be altered. If the model generates different outputs for the original test case and one of the altered cases, this will indicate direct causation between ethnicity and the outputs, based on counterfactual reasoning. However, given that the pre-deployment assessment types that are required in the relevant legal framework are oriented towards the probability of discrimination,[1054] the fact that one of the test cases indicate direct causation does not necessarily mean the applicable threshold for direct causation has been established at the system level. To determine direct causation at the system level, there is a need to somehow quantify and generalise the influence of PILFs on model outputs.

If the model is run on a large number of simulated cases, it might be possible to isolate and quantify the relative influence of sex or ethnicity on the model's outputs. ML literature suggests certain techniques for explaining how a feature influences a model's outputs. For instance, Bambauer et al. note that a technique called 'individual conditional expectation' maps how a model responds to changes in an input variable, and the related method called 'partial dependence plots' asks "how would the model respond, averaged across the dataset, to changes in the value of a specific input variable?"[1055] Another potentially relevant technical method is measuring the Causal Concept Effect (CaCE), as proposed by Goyal et al.[1056] This method is proposed for the purpose of "explaining classifiers (…) for high-level concepts whose presence or absence (everything else being equal) affect the model's prediction, as opposed to merely being correlated with the model's prediction."[1057] Thus, CaCE addresses

---

[1053] Kleinberg et al. argue that one cannot determine what an algorithm will do by reading the underlying code: Kleinberg et al. (2018) 114.

[1054] Chapter 7.

[1055] Bambauer, Zarsky, and Mayer (2021) 2415-16.

[1056] Yash Goyal et al., "Explaining Classifiers with Causal Concept Effect (Cace)," *arXiv preprint arXiv:1907.07165* (2019): 1, https://doi.org/10.48550/arXiv.1907.07165.

[1057] Ibid.

the importance of distinguishing between causation and correlation when trying to understand the influence of certain features on a model's outputs.

Further research should explore the application of techniques such as individual conditional expectation, partial dependence plots, and CaCE, specifically for the purpose of estimating the influence of PILFs on model outputs as part of a pre-deployment discrimination assessment.

In a pre-deployment discrimination assessment, such techniques could enable the assessment of whether the influence of a protected characteristic is adequately justified by the medical justification given by an AI provider, which this thesis has argued is crucial in discrimination assessments of systems intended for decisions on diagnosis, treatment recommendation or preventive intervention. However, with regards to AI-CDS systems intended to support the allocation of scarce resources, one should attempt to determine whether the influence of a PILF exceeds a certain threshold, in which case this would indicate a high probability of direct discrimination. Such a threshold cannot be derived from the CJEU's case law, which is case-specific and ex post-oriented.

10.4.6 Limitations of Counterfactual Reasoning in Non-Discrimination Law

The idea of defining and, thus, identifying discrimination through counterfactual reasoning has been criticised in literature addressing the conceptualisation of racial discrimination in the context of US law. While counterfactual reasoning plays an important role in the jurisprudence from US courts, scholars have argued that there is a need for a radically different conceptualisation of racial discrimination in relation to counterfactual reasoning.[1058] Kohler Haussman notes that causal definitions of discrimination, which rely on counterfactual reasoning, cannot properly capture the effects produced by the social construct of 'race'.[1059] She argues that counterfactual reasoning is not compatible with the notion that 'race' is understood as a social construct rather than a biological fact.[1060] The argument is transferrable

[1058] Issa Kohler-Hausmann, "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination," *Northwestern University Law Review* 113, no. 5 (2018); Selbst (2021).

[1059] Kohler-Hausmann (2018) 1168.

[1060] Ibid.

to an EU law context, because 'ethnicity' in EU non-discrimination law is also a social construct. While Kohler Haussman's critique does not relate directly to algorithmic discrimination, Selbst subscribes to her critique and sets forward similar arguments in the specific context of algorithmic discrimination.[1061]

Selbst argues that "[t]he only way to properly generate a counterfactual that could be used to measure the specific impact of race would be to remove the racial component from all variables that have it."[1062] The point is that the social construct of race/ethnicity inherits all the disadvantages experienced by people of a certain racial or ethnic origin:

> If what it means to be Black in the US is not based on skin color, but is instead a condition of Blackness that comprises all the historical disadvantages that that entails […] then simply changing the race variable in a machine learning model to encode a different skin color does not accurately account for the counterfactual.[1063]

The implication of the critique from Kohler Haussman and Selbst is that truly removing a person's race/ethnicity, as a social construct, would require the removal of all features that may have been influenced by that person's race/ethnicity. It is not sufficient, according to those authors, to simply compute an alternative decision procedure where a person's ethnicity or sex is changed.[1064] To conduct counterfactual testing of an AI-CDS system that accommodates this criticism, one would have to create a simulation where, for example, Simon Tesfay from Eritrea lived his entire life as an ethnic Norwegian. This is arguably not feasible in practice, due to the difficulty of determining with any certainty how every aspect of his life would have been, in such a hypothetical scenario. While the arguments made by

---

[1061] Selbst (2021) 131-32.

[1062] Selbst (2021) 132.

[1063] Ibid.

[1064] A similar point is made by Niklas, who argues that harms "co-created" by AI systems "cannot be addressed in isolation from everyday struggles related to exploitation, domination, and oppression": Jędrzej Niklas, "Human Rights-Based Approach to AI and Algorithms: Concerning Welfare Technologies," in *The Cambridge Handbook of the Law of Algorithms*, ed. Woodrow Barfield, Cambridge Law Handbooks (Cambridge: Cambridge University Press, 2020), 535; Seeta Peña Gangadharan and Jędrzej Niklas, "Decentering Technology in Discourse on Discrimination," *Information, Communication & Society* 22, no. 7 (2019), https://doi.org/10.1080/1369118X.2019.1593484.

Kohler Haussman and Selbst are thought-provoking and worth reflecting over, it remains the case that counterfactual computations would largely be accepted by the CJEU according to the above analysis of EU non-discrimination law. Counterfactual reasoning is therefore legitimate also when applying the non-discrimination principle in the context of a pre-deployment discrimination assessment.

10.4.7 Stereotyping or Prejudice as Direct Discrimination

In the *CHEZ* ruling, the CJEU suggests that if a practice is based on ethnic stereotyping or prejudice towards an ethnic group, this is something that would indicate that the disputed practice may qualify as direct discrimination.[1065] Academic literature largely endorses this view.[1066] It is based on the notion that when stereotyping or prejudice is involved, there remains a strong causal link between the disadvantageous treatment and a protected characteristic.[1067] Consequently, if a pre-deployment discrimination assessment finds bias stemming from stereotyping or prejudice,[1068] it can reasonably be considered an indication of potential direct discrimination, even if the model in question does not use any PILFs as feature variables.

Chapter 4 explained several ways in which stereotyping and prejudice can influence a model employed in an AI-CDS system. Compared to the situation where a decision is made directly by a prejudiced decision-maker, prejudice in AI systems can be several steps removed and, thus, more indirect. Consequently, it is questionable whether the causal connection is still strong enough for the system to be deemed directly discriminatory. The answer arguably depends on exactly where the prejudice comes from.

[1065] CHEZ, C-83/14, para. 82.

[1066] Henrard (2019) 108.; Moreover, Hacker submits the view that "if the decision maker himself makes labelling decisions affected by implicit bias, it seems more convincing to view this as a case of direct discrimination as the less favorable result is a direct consequence of the biased labelling": Hacker (2018) 1152; Adams-Prassl, Binns, and Kelly-Lyth (2023) 164. See also Connolly (2011) 107.

[1067] Henrard argues that when prejudice against an ethnic group is the root cause of how persons from that group is treated, it goes beyond mere "effect" and lies "closer to intent": Henrard (2019) 108.

[1068] Section 4.4.2.6.

Methodological element

Consideration: Are biases likely to be caused by prejudice/stereotyping, and from which source?

- Historical cognitive biases among past decision-makers, reflected in training data
  - No direct causation
- Stereotyping/prejudice influencing choices made during development of the AI-CDS system
  - Consider potential direct discrimination

One issue with AI models that was discussed in chapter 4 is the potential for training data to carry historical prejudices held by past decision-makers.[1069] An AI-CDS system may be influenced by stereotypes or prejudices involved in clinical assessments, which are reflected in the training data. In this scenario, the prejudice is arguably too detached from the disadvantageous outputs received by patients on whom the AI-CDS system is used. It therefore should not be taken as an indication of direct discrimination. When prejudice/stereotyping is seen as something that might indicate direct discrimination, this is because a prejudiced decision-maker holds an attitude that resembles a discriminatory motivation of sorts, even if it is an attitude that the decision-maker may be unaware of. When prejudice among past decision-makers creeps into a model through the training data, this is arguably one step too far removed to be equated with a prejudiced decision-maker.

Another way in which prejudices can cause bias in an AI-CDS system is through the many different decisions that are made by the development team, including labelling of training data for supervised learning techniques.[1070] In this scenario, the connection is arguably stronger between prejudice/stereotyping and the disadvantageous outputs that some patients will receive from the AI-CDS system. Whether it is strong enough to warrant a finding of direct discrimination, will depend on the circumstances. For example, consider the situation where a development team decides to alter the pain levels reported by women before feeding the data to a learning algorithm. They do so because they assume that women tend to exaggerate the pain they experience. This would constitute a blatant example of prejudice that could directly cause disadvantageous treatment of patients from a protected group. It may arguably be treated as direct discrimination under EU non-discrimination law. A pre-deployment

[1069] Section 4.4.2.6; Barocas and Selbst (2016) 675.

[1070] Section 4.4.3.5.

discrimination assessment should therefore consider whether prejudiced decisions may have been made during the development of an AI-CDS system.

10.5 Causation and Indirect Algorithmic Discrimination

10.5.1 Causation in Relation to Indirect Discrimination is About Attributing Disadvantage to an AI-CDS System

In contrast to direct discrimination, a finding of indirect discrimination requires causation between a provision, criterion or practice and a disadvantage incurred by a group. In relation to indirect discrimination, there is no requirement of causation between a protected characteristic and the disadvantage experienced by a group. According to the RED and the GSED, respectively, a provision, criterion or practice is indirectly discriminatory if it "would put persons of a racial or ethnic origin at a particular disadvantage compared with other persons,"[1071] or if it "would put persons of one sex at a particular disadvantage compared with persons of the other sex."[1072] This wording requires that a particular disadvantage must somehow be attributed to the disputed practice. In the context of a pre-deployment discrimination assessment, this requirement raises the question of *why* the outputs from an AI-CDS system is biased to the disadvantage of a protected group. In relation to indirect discrimination, the reason for the bias is not that a model uses on a protected characteristic as a feature variable. The assessment of legal causation in relation to indirect discrimination can therefore concentrate on other reasons for a biased distribution of disadvantageous outputs.

Importantly, the reasons why biases potentially constituting indirect discrimination occur in an AI-CDS system are relevant both under the causation assessment discussed in this chapter and under the objective justification test discussed in chapter 11. The close relationship between the objective justification test and the causation requirement is further considered in the following.

10.5.2 The Role of Counterfactual Reasoning in Relation to Indirect Discrimination

Unlike in the assessment of direct causation, counterfactual reasoning about a patient's characteristics is not relevant when assessing indirect discrimination in an AI-CDS system. It

[1071] Article 2(2)(b) RED.

[1072] Article 2(b) GSED.

does not matter whether one specific case within a test dataset would have been treated differently 'but for' the patient's sex or ethnicity.[1073] If one were to apply a 'but for' test for indirect discrimination, one would have to ask a different counterfactual question. For indirect discrimination, the counterfactual question would be: *Would a protected group be at a particular disadvantage but for the disputed practice?*[1074] This question emphasises a discussion of whether a particular disadvantage for a group is caused by an AI-CDS system or by a cause that is too detached from the AI-CDS system for attribution to be warranted. The distinction is difficult to make in circumstances where one might say that biases occur through a complex interplay between the AI development process[1075] and the societal conditions surrounding data generation, collection, and processing. In these cases, one could say that biases are "co-created" by AI systems.[1076]

Arguably, the question of whether a disadvantage must be attributed to an AI-CDS system is not a question that can be answered through causal, counterfactual reasoning alone. Primarily, the question is whether the connection between an AI-CDS system and a disadvantage pertaining to a protected group is of such a nature that discrimination should be attributable to the AI-CDS system. This question requires a broader balancing of conflicting arguments beyond those relating merely to 'causation.' It is not a question of causation as much as it is a question of objective justification, which also requires consideration of why a bias occurs.[1077]

[1073] In contrast, Ellis and Watson seem to assume that a 'but for' test is applicable both under direct and indirect discrimination: "… whether the alleged discrimination is direct or indirect, all that is required to establish causation is that 'but for' belonging to the protected category, the victim would not have sustained the disadvantage alleged": Ellis and Watson (2012) 167.

[1074] The UK Supreme Court suggests in *Essop* that such a 'but for' reasoning is appropriate: Judgment of the UK Supreme Court of 5 April, 2017, Essop and Others, [2017] UKSC 27, para. 26.

[1075] See section 1.5.9.

[1076] Niklas (2020) 535.

[1077] In relation to the disparate impact doctrine in US law, Packin and Lev-Aretz view this as a matter of causation, noting that a robust causation requirement increases the possibility that decisions made by machine learning algorithms may be able to escape liability "as long as they can demonstrate that their models only replicate and imitate present methods of systemic bias against minorities": Nizan Geslevich Packin and Yafit Lev-Aretz, "Learning Algorithms and Discrimination," in *Research Handbook on the Law of Artificial Intelligence*, ed. Woodrow Barfield and Ugo Pagallo (Edward Elgar Publishing Limited, 2018), 99.

This overlap renders the considerations of causation and justification inseparable in the context of assessing indirect discrimination.

The objective justification test allows the reasons why a bias occurs to be considered in a wider context where the magnitude of the potential consequences of the bias are also taken into account. The objective justification test is therefore better suited as a venue for the assessment of whether a bias should be attributed to an AI-CDS system. It is therefore contended that when assessing potential indirect discrimination in an AI-CDS system, the assessment should concentrate on whether a biased distribution of disadvantageous outputs is objectively justified rather than on whether it is 'caused' by the AI-CDS system that is being assessed. Given this view, I take a similar position to that of Hellborg, when it comes to the relationship between causation and objective justification in relation to indirect discrimination. Hellborg does not find that a separate consideration of causation is required for a finding of direct discrimination (ex post).[1078] She rather argues that the crux of an indirect discrimination assessment is to weigh the purpose of the disputed measure against the negative impacts of the measure, which is something that the objective justification test facilitates.[1079]

Methodological elements

Consideration for indirect discrimination assessment:

? Consider the strength of the connection between the source of bias and the disadvantage that a protected group might experience if the system is deployed.
  - Method: Assessment of training data, feature variables, target variables, and the intended population based on the known sources of bias discussed in chapter 4.
  - ➢ Proceed to objective justification test, where the strength of the connection is one of several considerations

Inside the frame of the objective justification test, considerations of the connection between an AI-CDS system and a disadvantage are important. It is arguably relevant to consider how closely a disadvantage is connected to the ML process. If a disadvantage is introduced through the ML process, there is arguably a stronger connection between the AI-CDS system and the disadvantage. With

[1078] Hellborg (2018) 297.

[1079] Hellborg (2018) 298.

reference to the sources of bias discussed in section 4.4, the following is an outline of how this reasoning could play out.

Data bias	Reasoning	Implication
Structural inequalities	If an AI-CDS system reflects structural inequalities, the bias and disadvantage is not introduced by the ML process. Rather, it exists in the real-world prior to the ML process.	Weak connection between ML process and bias/disadvantage
Sample bias	The bias and disadvantage have arisen because AI developers failed to collect data that properly reflects the relevant population.	Strong connection between ML process and bias/disadvantage
Unequal representation	The bias is a result of the composition of the relevant population and the disadvantage is a result of ML-based systems' tendency to perform better on groups on which it has seen more training data.	It is arguable that there is a strong connection, although it might vary how much developers can influence this type of bias and disadvantage in practice. This should be considered as part of the objective justification test.
Aggregation bias	Aggregation is a property of the ML process and typically means that developers have overlooked something they should have accounted for.	Strong connection between ML process and bias/disadvantage

Historical error disparities	The bias and disadvantage existed prior to the ML process and is repeated by the AI-CDS system.	Weak connection between ML process and bias/disadvantage
Bias in data processing and modelling choices	When bias that leads to a disadvantage is introduced through feature selection or the choice of target variable, this is a direct consequence of the ML process.	Strong connection between ML process and bias/disadvantage
Hidden inferences	When a learning algorithm makes hidden inferences and a disadvantage is a result thereof, there is a close connection between the ML process and the disadvantage.	Strong connection between ML process and bias/disadvantage
Deployment bias	Deployment bias is often similar to sample bias, as the relevant population upon deployment turns out not to correspond to the population an algorithm has learned from.	Strong connection between ML process and bias/disadvantage.

Table 8: The connection between bias from different sources and the disadvantage incurred by a protected group if an AI-CDS system is deployed.

The table above outlines causal considerations that are relevant when assessing indirect discrimination in an AI-CDS system. However, rather than concluding separately on whether or not there is sufficient 'causation,' these considerations should be fed into the objective justification test.

10.5.3 Causation, Justification, and Explanation

As indicated, to consider the connection between an AI-CDS system and a group-level disadvantage, an explanation is required of why a bias occurs in an AI-CDS system. This is a different type of explanation than the explanation of causal relations between features and between features and outputs, which are relevant in relation to the direct discrimination assessment. In an ex post enforcement context, the onus would be on the defendant of showing that the reason for a particular disadvantage is a reason that cannot be attributed to the disputed practice, due to the burden of proof in EU non-discrimination law. This is an important principle that allows the prohibition on indirect discrimination to redress discrimination that occurs as a result of complex structures which may be challenging to explain. Khaitan argues that "the entire point of indirect discrimination law is to unearth and redress discrimination that is typically harder to detect, let alone explain."[1080]

Ex post, a claimant must show that there is a particular disadvantage but is not required to explain the reasons why a disadvantage occurs.[1081] The CJEU holds in the *Danfoss* ruling that "where an undertaking applies a system of pay which is totally lacking in transparency, it is for the employer to prove that his practice in the matter of wages is not discriminatory."[1082] Accordingly, biases that cannot be plausibly explained arguably cannot be justified in an objective manner. This would apply in an ex post enforcement context, and the same principle should apply in the context of a pre-deployment discrimination assessment. A supporting argument is that such a principle would incite AI providers to consider, and produce explanations of, the potential disadvantages of an AI system.[1083]

[1080] Tarunabh Khaitan, "Indirect Discrimination Law: Causation, Explanation and Coat-Tailers," *Law Quarterly Review* 132 (January 2016): 37.

[1081] With reference to the EU Equality Directives, the UK Supreme Court notes in *Essop* that "[i]n none of the various definitions of indirect discrimination, is there any express requirement for an explanation of the reasons why a particular PCP puts one group at a disadvantage when compared with others": Essop and Others, para. 24.

[1082] Judgment of 17 October, 1989, Danfoss, C-109/88, ECLI:EU:C:1989:383, para. 11.

[1083] A similar argument is put forward by Khaitan in relation to the ex post enforcement context: Khaitan (2016) 5.

10.6 Conclusion

This chapter began by highlighting the distinction between causation in direct and indirect discrimination. The subsequent analysis further solidified this distinction, suggesting that the assessment of causation might play a crucial role in determining whether an AI-CDS system should be deployed.

With regards to direct discrimination, this chapter has highlighted a distinction between the assessment of discrimination in AI-CDS systems intended to assist in scarce resource allocation, on the one hand, and systems intended for diagnosis, treatment recommendation and preventive intervention, on the other. In relation to the latter categories of clinical decisions, the disadvantage of disparate performance is the primary disadvantage to consider.[1084] No disparate performance is, by definition, caused by reliance on a PILF, if there is an adequate medical justification for relying on the PILF.[1085] Any reliance on a PILF as a feature variable might therefore be medically justified in proportion to its influence on a model's outputs. Consideration of the medical justification for relying on a PILF should therefore be part of the direct causation assessment as far as assessing causation between a PILF and the disadvantage of disparate performance is concerned.

In contrast, if an AI-CDS system is intended for decision-making concerning scarce resource allocation, the disadvantage of interest is that of resource denial.[1086] Patients who receive undesirable outputs receive this type of disadvantage regardless of the medical justification for the differential treatment. A finding of direct causation when assessing these systems therefore indicates direct discrimination.

It follows that for all types of clinical decisions, a pre-deployment assessment of direct discrimination should attempt to establish the relative weight assigned to PILFs within an AI-CDS system. Even though the usefulness of asking which features 'cause' an output from an AI model has been questioned in ML literature,[1087] this is the appropriate way of framing the question as part of a discrimination assessment based on the non-discrimination principle in

---

[1084] Chapter 9.

[1085] Section 10.4.4.

[1086] Section 9.3.2.

[1087] Selbst (2021) 178; David Lehr and Paul Ohm, "Playing with the Data: What Legal Scholars Should Learn About Machine Learning," *UC Davis Law Review* 51 (2017): 707.

EU law. The causation requirement refers to the significance of a protected characteristic on the outputs produced by an AI-CDS system. In highly interpretable models, the weight of different feature variables may be readily observable upon scrutinization of the model. In contrast, deep learning models may rely on an incomprehensible number of feature variables, the relative importance of which can be challenging to determine.[1088] Lack of interpretability would restrict the feasibility of understanding how each of a potentially enormous number of features contribute to a model's outputs.[1089]

EU non-discrimination law does not clarify the threshold for determining whether there is direct causation. In its case law, the CJEU applies a 'but for' test, in which it is considered, retrospectively, whether the claimant would have been treated differently if it was not for the protected characteristic at issue. While challenging to apply in a pre-deployment context, the test encourages counterfactual reasoning. It was therefore argued in this chapter that counterfactual testing and explanations of a model's outputs should be included in a pre-deployment discrimination assessment. Counterfactual methods are primarily relevant to consider in relation to opaque models in which the relative weight of feature variables are not directly observable. Potential technical methods for counterfactual testing were identified in this chapter, while underscoring the need to further investigate the utility of those methods in a pre-deployment discrimination assessment context. Estimation of feature importance is an ongoing subject of research in the field of explainable AI (XAI).[1090]

In relation to the causation requirement for indirect discrimination, the focus shifts towards identifying different sources of bias rather than explaining the causal relationship between feature variables and outputs. It was concluded that considerations of causation in relation to indirect discrimination are best addressed as part of the objective justification test, which takes into account a broader set of considerations beyond mere causation. While the objective justification test is explored in the subsequent chapter, the present chapter has outlined the

---

[1088] Bathaee argues that it might be impossible to audit a neural network to "determine what information it found outcome-determinative or how it is making decisions": Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation," *Harvard Journal of Law & Technology* 31, no. 2 (Spring 2017): 927.

[1089] Mittelstadt et al. (2016) 4.

[1090] Alsaleh et al. (2023) 7.

causation-related considerations that should be integrated in the objective justification part of a pre-deployment discrimination assessment.

The methodological elements of pre-deployment discrimination assessment for AI-CDS systems developed in this chapter are summarised as follows:

Direct discrimination assessment

Objective: Determine whether an AI-CDS system is influenced by PILFs to such a degree that patients disadvantaged by the system would be directly discriminated against.

Relative importance of a PILF on model outputs:

- ? What is the relative importance of PILFs within a model?
  - □ Technical method: counterfactual testing and explanations
    - ➢ e.g., Individual Conditional Expectation, Partial Dependence Plots, or Causal Concept Effect (CaCE)

Intended purpose:

- ? Which type of clinical decision is the system intended for?
- □ Diagnosis, treatment recommendation, or preventive intervention;
  - ➢ The degree of influence from PILFs should be justified by the medical justification for relying on PILFs in the specific clinical decision
- □ Allocation of scarce resources
  - ➢ The non-discrimination principle does not permit direct causation between PILFs and disadvantageous outputs

Source of bias:

- ? Are biases likely to be caused by prejudice/stereotyping, and from which source?
  - ➢ Historical cognitive biases among past decision-makers, reflected in training data
    - ➢ No direct causation
  - ➢ Stereotyping/prejudice influencing choices made during development of the AI-CDS system
    - ➢ Consider potential direct discrimination

Indirect discrimination assessment

Source of bias:

? Consider the strength of the connection between the source of bias and the disadvantage that a protected group might experience if the system is deployed.
  - Method: Assessment of training data, feature variables, target variables, and the intended population based on the known sources of bias discussed in chapter 4.
  - ➢ Proceed to objective justification test, where the strength of the connection is one of several considerations

11 Objective Justification of Biased AI-CDS Systems

11.1 Introduction

According to the Equality Directives, a practice that puts a protected group at a particular disadvantage does not constitute indirect discrimination if the practice is "objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary."[1091] In a pre-deployment discrimination assessment, given that one follows the structure of the prohibition on indirect discrimination, the issue of objective justification is relevant in cases where it is established that an AI-CDS system is biased to such an extent that the system would be likely to put patients from a protected group at a particular disadvantage. In these cases, the question arises whether the bias may be objectively justified in accordance with the non-discrimination principle.

This chapter analyses the objective justification requirement based on CJEU jurisprudence and aims to develop the considerations that should be included when applying this requirement in the context of a pre-deployment discrimination assessment. The methodological elements of the objective justification assessment are not easily derived from the wording of the Equality Directives.[1092] Case law is therefore an important source of interpretation. To develop methodological elements, the insights gained from legal interpretation of the objective justification requirement are combined with knowledge of AI-CDS systems, the typical reasons for deploying these systems, and the sources of equality-related biases and potential discrimination that occurs in these systems.

This chapter finds that some of the considerations that should be included when applying the objective justification requirement as part of a pre-deployment discrimination assessment refer to the specific circumstances of individual deployers. This raises certain challenges when it comes to conducting such an assessment for the purpose of determining whether an

---

[1091] Article 2(2)(b) RED; Article 2(b) GSED.

[1092] Makkonen notes that the objective justification test so general that it has limited ability to inform the interpretation of the law and guide its application in practice Timo Makkonen, *Equal in Law, Unequal in Fact: Racial and Ethnic Discrimination and the Legal Response Thereto in Europe*, ed. Martti Koskenniemi, The Erik Castrén Institute Monographs on International Law and Human Rights Series, (Leiden: BRILL, 2012), 259. http://ebookcentral.proquest.com/lib/tromsoub-ebooks/detail.action?docID=842208.

AI-CDS system should be placed on the market. These challenges are discussed and solutions proposed, in section 11.7.

11.2 Overview of the Objective Justification Requirement

11.2.1 Three Components Reflected in the Wording of the Equality Directives: Legitimate Aim, Suitability, and Necessity

The wording of the objective justification requirement in the Equality Directives comprises three components. First, justification must be based on aims recognised as legitimate within the EU legal order. Second, the disputed practice must be a suitable means of pursuing the legitimate aims. Third, the disputed practice must be necessary to achieve those aims.

11.2.2 Proportionality Stricto Sensu as a Fourth Component

A fourth component – proportionality in a narrow sense, is not explicitly reflected in the wording of the Equality Directives.[1093] However, the Directives are specific expressions of the non-discrimination principle enshrined in Article 21 of the Charter. Article 52(1) of the Charter explicitly invokes the "principle of proportionality," stating that limitations on Charter rights may only be made if "they are necessary and genuinely meet objectives of general interest recognised by the Union or the need to protect the rights and freedoms of others."[1094] This wording points out the three core elements of objective justification in EU non-discrimination law – legitimacy of the aim, suitability, and necessity. Article 52(1) explicitly frames these components as facets of the principle of 'proportionality.' In the *Léger* ruling, proportionality under Article 52(1) of the Charter is interpreted to mean that:

> … the measures laid down by national legislation must not exceed the limits of what is appropriate and necessary in order to attain the objectives legitimately pursued by that legislation; when there is a choice between several appropriate measures, recourse

[1093] In some legal systems where there is a statutory objective justification rule that essentially corresponds to the rule found in EU non-discrimination law, a requirement of 'proportionality' is explicitly included in the statutory wording. For example, this is the case with the UK 2010 Equality Act, Section 19(2)(d) (requiring the disputed practice to be a "proportionate means of achieving a legitimate aim"), and the 2017 Norwegian Equality and Non-Discrimination Act, § 9 (requiring that the disputed practice must not be disproportionate towards the persons being subject to differential treatment).

[1094] Article 52(1) EUCFR.

must be had to the least onerous among them, and the disadvantages caused must not be disproportionate to the aims pursued.[1095]

The final part of the quoted statement, which emphasizes that the disadvantages must not be disproportionate to the aims pursued, is not always clearly expressed in the case law of the CJEU under the Equality Directives. This aspect of proportionality, known as 'proportionality in a narrow sense' or 'stricto sensu,' is considered in the remainder of this chapter as a fourth component of the objective justification requirement. The assessment of proportionality stricto sensu as part of a pre-deployment discrimination assessment for AI-CDS systems is further discussed in section 11.6.

11.2.3 Objectivity as a Fundamental Requirement Underpinning the Aforementioned Components of Objective Justification

When discussing whether a disputed practice is justified, a fundamental requirement is that the justification must be 'objective.' The 'objective' part of the requirement traces back to the *Bilka* ruling, according to which there must be "objectively justified factors unrelated to any discrimination."[1096] In the specific context of biases in an AI-CDS system, one implication of the objectivity requirement is that a biased system probably cannot be justified by reference to biased data. This is particularly relevant in relation to the suitability requirement. The evidence relied on to demonstrate the suitability of deploying a biased system to pursue a legitimate aim should satisfy certain quality standards, as section 11.4.5 elaborates.

Furthermore, the objectivity requirement may be interpreted as suggesting that the deployment of an AI-CDS system is not justified if it perpetuates biases or inequalities from the past, because this would imply that the reasons for deploying the system are not "unrelated to any discrimination."[1097] However, determining whether the reproduction of historical inequalities in an AI-CDS system is capable of being objectively justified, is not a question that can be answered based on the objectivity criterion alone. It necessitates an assessment that takes into account all aspects of the objective justification requirement. As subsequent sections discuss, this poses questions regarding the feasibility of addressing

---

[1095] Léger, C-528/13, 58.

[1096] Bilka, C-170/84, para. 30.

[1097] Ibid; Adams-Prassl, Binns, and Kelly-Lyth (2023) 153.

historical inequalities during the development of an AI-CDS system, and the associated costs involved.

11.3 Legitimate Aims of Deploying an AI-CDS System

11.3.1 Typical Reasons for Deploying an AI-CDS System

In order for the deployment of an AI-CDS system to be deemed justifiable, it must pursue a legitimate aim. Applying this aspect of the objective justification requirement is likely to be rather straightforward. The CJEU has, in its case law, recognized a wide range of aims as legitimate.[1098] Given the broad range of legitimate aims in EU non-discrimination law, health institutions or AI providers may have several valid reasons for implementing AI-CDS systems. Some of the primary aims that are likely to be invoked in practice, include:

Methodological element

Consideration:

? What legitimate aims may be pursued by deploying the AI-CDS system?

1. improving the quality of care by increasing predictive accuracy;
2. improving efficiency by decreasing the time and resources required to produce a clinical assessment;
3. facilitating preventive interventions and/or more personalised care;
4. increasing access to specialised care (e.g., by enabling general practitioners to conduct more specialised assessments with the assistance of AI-CDS systems or enabling assessments in areas or situations where there is a scarcity of healthcare professionals).

These reasons for deploying AI-CDS systems constitute a reference frame for the following analysis. When assessing whether deployment of a biased AI-CDS system is objectively justified, the reason for deploying the system is central to the assessment. The requirements of

[1098] For summaries of the aims that have been accepted in the CJEU's case law, see: Tobler (2008): 33; Maliszewska-Nienartowicz (2014) 44-45; Wachter (2020) 410-11.

suitability, necessity and proportionality all require consideration of the legitimate aim pursued.

11.3.2 Efficiency and Economic Considerations

The abovementioned aims are likely to be recognised as legitimate by the CJEU. However, it is worth noting that the Court often does not accept justifications based on strictly budgetary or economic considerations.[1099] On the other hand, there are also cases where the Court seems to accept purely economic considerations as legitimate, particularly for private entities.[1100] The general issue of to what extent EU non-discrimination law recognises purely economic considerations as legitimate is not explored in-depth here. However, it is worth examining the legitimacy of considerations related to economy, resource efficiency, etc., within the specific context of providing health services. In the *Szpital Kliniczny* ruling, the CJEU suggests that 'saving money' alone cannot constitute a legitimate aim:

[1099] When it comes to Member States pursuing an aim under its social policy, the CJEU is adamant about rejecting "strictly budgetary considerations." In *De Weerd*, the Court notes that "although budgetary considerations may influence a Member State's choice of social policy and affect the nature or scope of the social protection measures it wishes to adopt, they cannot themselves constitute the aim pursued by that policy and cannot, therefore, justify discrimination against one of the sexes": Judgment of 24 February, 1994, De Weerd, C-343/92, ECLI:EU:C:1994:71, para. 35. Similarly, in the *YS* case, the CJEU indicates that sex discrimination resulting from a national law cannot be justified by reference to "budgetary considerations": YS, C-223/19, para. 60. Furthermore, in the *Szpital Kliniczny* ruling, the CJEU categorically rejects the legitimacy of the aim of saving money in a case that does not directly concern the aims of a state's social policy: Szpital Kliniczny, C-16/19, para. 59. Tobler firmly rejects the legitimacy of purely economic considerations under EU non-discrimination law, "purely budgetary considerations can never serve as an objective justification," at least as far as legislative acts are concerned: Tobler (2008): 33; Dalenberg (2018) 623; Tobler (2022): 68.

[1100] Bilka, C-170/84, para. 36. Ellis and Watson note that the CJEU has not been clear about the importance of "budgetary considerations" in the shaping of member states' social policy measures: Ellis and Watson (2012) 467. Wachter notes that some economic justifications can be seen as legitimate, with reference to *Nikoloudi* and *Land Hessen*: Wachter (2020) 408-09. Zuiderveen Borgesius leaves the question of whether budgetary considerations are permissible for private companies open: Borgesius (2020 B) 415.

> … it seems to be apparent from the intended purpose of the practice at issue in the main proceedings, which is to save money, that the conditions for such justification are not satisfied, which it is for the referring court to verify, where appropriate.[1101]

According to the facts of the case, an employer decided to pay an allowance only to employees who presented documentation of their disabilities after a certain point in time. Employees that had presented this documentation before that specific point in time, did not get the allowance. This way, the employer saved money by excluding those who had already documented their disabilities at an earlier point in time. Importantly, given these peculiar facts, there seems to have been no plausible reason whatsoever for such a potentially discriminatory practice, other than saving money.

While economic considerations may be influential on the decision to deploy an AI-CDS system, this decision usually involves broader considerations beyond 'saving money.' For example, the aim of deploying an AI-CDS system is not solely economic if deployment is aimed at providing healthcare more swiftly and resource-efficiently. This intent aligns with the broader objective of increasing the availability of healthcare services and hastening their delivery while reducing clinicians' workload. The legitimacy of such aims arguably finds support in the CJEU's ruling in the *Jørgensen* case, where the Court accepts the legitimacy of "measures intended to ensure sound management of public expenditure on specialised medical care and to guarantee people's access to such care."[1102] Notably, this statement links expenditure to the need to guarantee access to care, a core function of any Member State's health system. Furthermore, a statement in the *Elchinov* ruling appears recognises a Member State's "desire to control costs, and to prevent, as far as possible, any wastage of financial, technical and human resources."[1103] The CJEU then points out that "such wastage would be all the more damaging because it is generally recognised that the healthcare sector generates considerable costs and must satisfy increasing needs, while the financial resources which may be made available for healthcare are not unlimited, whatever the mode of funding applied."[1104] This statement is referred to by the CJEU in *Veselības ministrija*,[1105] where it

---

[1101] Szpital Kliniczny, C-16/19, 59.

[1102] Judgment of 6 April, 2000, Jørgensen, C-226/98, ECLI:EU:C:2000:191, para. 42.

[1103] Judgment (GC) of 5 October, 2010, Elchinov, C-173/09, ECLI:EU:C:2010:581, para. 43.

[1104] Ibid.

[1105] Veselības Ministrija, C-243/19, para. 46.

leads the Court to conclude that "it cannot be excluded that the possible risk of seriously undermining the financial balance of a social security system may constitute a legitimate objective capable of justifying a difference in treatment based on religion."[1106]

The cited case law entails two lessons. The first lesson is that it is indeed legitimate to strive to avoid wastage of financial, technical and human resources. This underscores the relevance of efficiency in clinical decision-making processes. The second lesson is that the CJEU may recognise strictly economic considerations as a sole legitimate aim, if the financial balance of a social security system is at stake. While it may be difficult to conceive how the deployment of an AI-CDS system could be directly decisive for the financial balance of a social security system, the first lesson is important in the context of deploying AI-CDS systems. Enhancing the efficiency of healthcare service provision is a legitimate aim for deployment of such systems.

11.4 Suitability

11.4.1 Suitability/Appropriateness/Effectiveness: Introduction

According to the Equality Directives, the means of achieving a legitimate aim must be "appropriate and necessary."[1107] This section deals with the requirement of appropriateness, often referred to as 'suitability' in the academic literature.[1108] To be justifiable, a potentially discriminatory practice must be a suitable means of pursing a legitimate aim.[1109] According to the CJEU, suitability requires that a measure is "appropriate for ensuring attainment of the objective pursued and genuinely reflects a concern to attain it in a consistent and systematic manner."[1110] This requirement of suitability is sometimes considered synonymous with the 'effectiveness' of the disputed measure in the CJEU's jurisprudence.[1111]

[1106] Veselības Ministrija, C-243/19, para. 47.

[1107] Article 2(2)(b) RED; Article 2(b) GSED.

[1108] e.g., Ellis and Watson (2012) 131.

[1109] For examples of instances where the CJEU does not find the suitability requirement to be satisfied, see: Judgment of 18 June, 2009, Hütter, C-88/08, ECLI:EU:C:2009:381; Judgment (GC) of 19 January, 2010, Kücükdeveci, C-555/07, ECLI:EU:C:2010:21.

[1110] e.g., Comune Di Gesturi, C-670/18, para. 50; WABE, Joined Cases C-804/18 and C-341/19, para. 68; Tobler (2022): 70.

[1111] YS, C-223/19, para. 63.

Hence, one relevant question to consider in a pre-deployment discrimination assessment is the extent to which an AI-CDS system is effective in achieving the aim pursued. In relation to AI-CDS systems, an important aspect of suitability is therefore the system's performance at the task it is intended for. The higher the performance of the system, the more effective it is at achieving its intended purpose and, thus, the legitimate aims that are pursued. Consequently, performance testing must be included in a pre-deployment discrimination assessment where an objective justification is required. In relation to the objective justification requirement, performance testing is about measuring the overall performance of an AI-CDS system, rather than measuring the extent of *disparate*, which is part of the disadvantage measurement discussed in chapter 9.

Methodological element

Consideration, suitability:

? How effective is the AI-CDS system at achieving the aim pursued?
- ➢ Performance testing
- ➢ See section 11.7 for implications

Performance is an important consideration, and a minimum level of performance is arguably required if an AI-CDS system is going to be deemed suitable. However, the suitability requirement does not refer only to how effective a measure is at achieving a legitimate aim. The suitability of deploying an AI-CDS system also depends on a broader set of considerations, which are discussed in the following sections.

11.4.2 Accuracy, Verifiability, and Human Oversight

A poorly performing AI-CDS system is not an effective way of conducting the task it is applied for. However, erroneous and inaccurate outputs are bound to occur from time to time. For guidance as to how the possibility of erroneous outputs should be treated in an objective justification assessment, this section turns to the CJEU's ruling in a case that concerns an interference with the fundamental rights to privacy and data protection enshrined in Articles 7 and 8 of the EU Charter of Fundamental Rights.

In the case of *Ligue des droits humains,*[1112] the CJEU was presented with the task of examining the validity of the EU Passenger Name Record (PNR) Directive,[1113] specifically with regard to its role in facilitating a surveillance regime that involved the automated assessment of the personal data of air passengers.[1114] At the heart of the matter, the CJEU was required to determine whether the interference with the rights to privacy and data protection could be justified in accordance with Article 52(1) of the Charter.[1115] It is unnecessary to elaborate the details of the surveillance regime at issue in this case. The automated assessment in question served the purpose of identifying passengers who posed a risk pertaining to their potential involvement in certain types of criminal activities. This automated assessment resulted in a considerable number of false positives.[1116]

The issue of false positives in the context of the disputed surveillance regime is examined by the CJEU as part of the broader question of whether this regime is an 'appropriate' means of achieving the legitimate aim of ensuring security in the EU, specifically by preventing and detecting terrorist offences or serious crime. The CJEU reasons that the appropriateness of a screening system which produces a significant number of false positives depends on the *proper functioning of the subsequent verification of the positive results through non-automated means*.[1117] The ruling thus confirms that the performance of an AI system is a crucial aspect to consider when determining the suitability of the system as a means of pursuing a legitimate aim.[1118] It further indicates that, if an AI system is used in a context where it is possible to verify the outputs produced by the system by non-automated means, this is a relevant argument supporting the suitability of the system.

[1112] Judgment (GC) of 21 June, 2022, Ligue Des Droits Humains, C-817/19, ECLI:EU:C:2022:491.

[1113] Directive (EU) 2016/681 of the European Parliament and of the Council of 27 April 2016 on the use of passenger name record (PNR) data for the prevention, detection, investigation and prosecution of terrorist offences and serious crime (PNR Directive).

[1114] Ligue Des Droits Humains, C-817/19, para. 111.

[1115] Ligue Des Droits Humains, C-817/19, para. 113.

[1116] Ligue Des Droits Humains, C-817/19, para. 123.

[1117] Ligue Des Droits Humains, C-817/19, para. 124.

[1118] Ligue Des Droits Humains, C-817/19, para. 123.

Methodological element

Consideration, suitability:

? Can disadvantageous outputs be verified by non-automated means?
  - ➢ Are state-of-the-art human oversight measures in place?

Implication:

If disadvantageous outputs cannot be verified, this provides an argument against deployment of the system.

Although the *Ligue des droits humains* ruling concerns Articles 7 and 8 of the Charter, the justification reasoning, which is based on Article 52(1) of the Charter, is arguably applicable also when assessing discrimination. The ruling specifically highlights the importance of verifying *positive* outputs from an automated system.[1119] This is due to the severe implications of positive outputs in the specific circumstances at issue in the case, where positive outputs indicated that a person should be deemed a security threat. In the context of the case, a positive output was disadvantageous, whereas a negative output was advantageous. Adapting CJEU's guidance from this case to the context of AI-CDS systems, it appears that a relevant question to raise is whether outputs defined as *disadvantageous* (see chapter 9) can be verified by non-automated means.

Consequently, the possibility of verifying disadvantageous outputs by non-automated means should be included as a consideration in a pre-deployment discrimination assessment methodology based on EU non-discrimination law. Specifically, in the context of AI-CDS systems, disadvantageous outputs could be subject to review by a clinician. In practice, such review should be guided by state-of-the-art human oversight measures for AI technologies. For example, techniques have been developed for the purpose of "visualizing the model's decision-making process and prediction confidence, allowing clinicians the ability to step in, in the case of a faulty prediction."[1120] However, human oversight measures for various AI technologies are an ongoing area of research.[1121] Future research efforts in this area could

[1119] Case law on data protection and privacy contains more examples of weight being given to specific safeguards as part of the justification assessment, see for example: Judgment (GC) of 21 December, 2016, Tele2 Sverige, Joined Cases C-203/15 and C-698/15, ECLI:EU:C:2016:970.

[1120] Diaz-Asper et al. (2023) 11.

[1121] e.g., Robert Munro Monarch, *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered AI* (Shelter Island: Simon and Schuster, 2021); Chelsea Chandler,

attempt to develop more specific methods, accounting for different types of AI-CDS systems, to enhance the type of human oversight that this section advocates for based on the non-discrimination principle.

11.4.3 Appropriateness of Target Variables

If it can be argued that a target variable is not 'appropriate' for the task at hand, this would suggest that the AI-CDS system may not be a suitable means of achieving the aim pursued. Therefore, the appropriateness of target variables is important to consider when determining the suitability of an AI-CDS system. The *Léger* ruling stresses the importance of assessing each person on the basis of their individual situation and behaviour, rather than relying on proxies.[1122] Moreover, one could argue that a target variable that perpetuates historical inequalities, does not carry the 'objectivity' that is required for the deployment of an AI-CDS system to be objectively justified. Based on these arguments, the target variables of an AI-CSD system may be deemed inappropriate based on EU non-discrimination law.

In the NORspine project,[1123] where an AI-CDS system is developed for the prediction of spine surgery outcomes, there are several possible target variables, each with different limitations. One potential target variable is the difference in a patient's ODI score before and after surgery.[1124] Because the ODI score is based on self-reported data, this is a target variable that is vulnerable to the inaccuracy of self-assessments. It might nonetheless be an appropriate target variable, provided that it represents the most relevant measurement of

Methodological element

Consideration, suitability:

Are target variables appropriate?

- Is there an adequate medical justification for the choice of target variables?
- Is it likely that a target variable reflects historical inequalities?

Peter W Foltz, and Brita Elvevåg, "Improving the Applicability of AI for Psychiatric Applications through 'Human-in-the-Loop' Methodologies," *Schizophrenia Bulletin* 48, no. 5 (2022), https://doi.org/10.1093/schbul/sbac038; Diaz-Asper et al. (2023) 9.

[1122] Léger, C-528/13, para. 67. Similar reasoning is found in the CJEU's ruling in *Specht and Others*, which concerns the justification of using age as a criterion: Judgment of 19 June, 2014, Specht and Others, C-501/12, ECLI:EU:C:2014:2005.

[1123] Section 5.2.

[1124] Oswestry Disability Index, cf. section 5.2.

the outcome, from a medical perspective. As long as there is a proper medical justification for the choice of a target variable, the target variable is compatible with the goal of assessing each patient's situation rather than relying on inappropriate proxies.

On the other hand, the hypothetical case of Simon Tesfay v. UHS illustrates a case of a less appropriate target variable.[1125] In this case, the utilisation of health services is relied on as a proxy for health needs. Given that this target variable might be ingrained with differences between ethnic groups that are not accounted for, one could argue that the target variable is not appropriate and that, consequently, the AI-CDS system used by UHS is not a suitable means of identifying patients eligible for the Preventive Care Program.

11.4.4 Importance of 'Consistency' in Suitability Assessment?

The CJEU requires that a potentially discriminatory measure must genuinely reflect a concern to attain a legitimate aim in a "consistent and systematic manner."[1126] The implication is that if an AI-CDS system does not demonstrate a consistent and systematic approach to clinical decision-making, its deployment may not be seen as a consistent and systematic way of improving clinical decision-making or realising other legitimate aims. It is therefore worth considering how the consistent/systematic requirement may be interpreted in relation to AI-CDS systems. There are a few possible interpretations of 'consistency' in this context.

First, consider an 'offline' AI-CDS system – one that does not have continuous learning capabilities.[1127] It can be expected that such a trained model will always deliver the same output when presented with the same input. In other words, its outputs are predetermined. This can be seen as one a form of consistency. Not all AI systems inherently possess this type of consistency. Some AI systems have a built-in element of randomness to its outputs, as a matter of design. Such an element of randomness would, of course, not be included in an AI-CDS system intended to support diagnosis, treatment selection, preventive interventions, or

[1125] Section 5.1.

[1126] Comune Di Gesturi, C-670/18, para. 50; Judgment (GC) of 14 March, 2017, G4s Secure Solutions, C-157/15 ECLI:EU:C:2017:203, para. 40; WABE, Joined Cases C-804/18 and C-341/19, para. 68.

[1127] Section 1.5.10.

resource allocation. In this sense, one may assume that AI-CDS systems will regularly display consistency.

Another type of consistency could be an AI-CDS system's tendency to produce similar outputs for similar – but not identical – inputs. In this respect, certain AI-CDS systems may be seen as inconsistent if slight variations in input data lead to significantly different outputs. However, inconsistency in a system's responses to similar inputs is only a problem in clinical decision-making if the inputs represent patients who are so similarly situated – in respect of clinically relevant factors – that they ought to receive similar outputs. This type of consistency is dealt with by the comparability assessment discussed in chapter 9.

Finally, consistency could also refer to an AI model producing expected outputs when prompted with new information that it has not seen during training. If the model's outputs deviate from expectations based on training and test results, one might regard the model as inconsistent with those results. This type of inconsistency relates to an AI-CDS system's performance. As already mentioned, an AI-CDS system's performance at the task it is intended for, is an important aspect of the suitability assessment. It follows that the requirement that a practice must be part of a consistent and systematic attempt to pursue a specific aim does not translate into any specific considerations to include in a pre-deployment discrimination assessment. An assessment of the suitability of deploying an AI-CDS system can concentrate on performance and the suitability of target variables, rather than the question of whether there is a 'consistent' or 'systematic' approach to the aims pursued.

11.4.5 Evidence Substantiating the Connection Between an AI-CDS system and a Legitimate Aim

When assessing the suitability of an AI-CDS system as a means to achieve a legitimate aim, the question arises as to what data or evidence is required. As alluded to in section 11.2.3, the connection between a practice and the aims pursued must be established 'objectively' through factual evidence. The CJEU states in the *Seymour-Smith* case that "[m]ere generalisations concerning the capacity of a specific measure" to achieve the legitimate aim are not sufficient.[1128] This principle is further exemplified in the *CHEZ* case, where an electricity provider installed electricity meters at higher positions in certain neighbourhoods to combat

[1128] Seymour-Smith, C-167/97, para. 76.

fraud. This practice put persons of Roma ethnicity at a particular disadvantage given that the targeted neighbourhood was predominantly inhabited by people of Roma origin. The CJEU instructs that, in such cases, the electricity provider must demonstrate the existence of the alleged fraudulent behaviour.[1129] Without specific evidence of fraudulent conduct that could be addressed through the higher placement of electricity meters, such a practice is not a suitable means of combating fraud.

While the relationship between means and ends must be established through objective, factual evidence, CJEU jurisprudence does not provide comprehensive guidance on how to assess the validity, representativeness and relevance of statistical or other evidence. This is routinely left for the national courts.[1130] Some guidance is nonetheless found in the *Léger* ruling. This ruling is based directly on the EU Charter rather than the Equality Directives. However, its guidance on assessing whether a practice is justified under Article 52(1) of the EU Charter remains relevant to objective justification assessment under the Equality Directives.

In *Léger*, a French decree regarding the conditions for blood donation is deemed potentially discriminatory as it made it more difficult for homosexual individuals to donate blood compared to heterosexual individuals.[1131] The justification asserted by French authorities was based on public health considerations, specifically the protection of blood recipients and others from HIV transmission. To justify the decree, French authorities referred to data from the period between 2003 and 2008, which showed that most HIV infections in France at the time resulted from sexual relations, with 48 % of new infections concerning men who had sex with other men.[1132] In keeping with the tradition for preliminary rulings, the CJEU refrains from making a conclusive determination of whether the French decree was objectively justified in light of this data. However, the Court stresses that it is a matter for the national court "to ascertain, in light of the current medical, scientific and epidemiological knowledge, whether [the data referred to by French authorities] is reliable and, if that is the case, whether it is still relevant."[1133]

[1129] CHEZ, C-83/14, para. 116.

[1130] e.g., YS, C-223/19, para. 49; Villar Láiz, C-161/18, para. 45.

[1131] Léger, C-528/13, paras. 49-50.

[1132] Léger, C-528/13, para. 42.

[1133] Léger, C-528/13, para. 44.

The *CHEZ* and *Léger* rulings underscore the importance of demonstrating the suitability of a potentially discriminatory measure through updated ("current"), reliable and relevant knowledge. Applying this principle to the context of assessing discrimination in an AI-CDS system before its deployment, the suitability of deploying a biased AI-CDS system should be supported by medical, scientific knowledge. This arguably implies that the performance (accuracy) of an AI-CDS system, as well as the appropriateness of its target variables, should be supported by evidence that satisfies certain scientific standards for reliability. The question of what the required standards are, finds a parallel discussion in contemporary discourse on the evidence standards for placing AI medical devices on the market in accordance with medical device regulations (in the EU and beyond).[1134] Determining these standards is a matter that should be addressed in future research and policy development. One question that should be addressed is whether the standards of evidence are stricter when it comes to justifying the deployment of a biased AI-CDS system, compared to when assessing whether the regular performance requirements in the MDR and (once it's adopted) the AIA.

Methodological element

Consideration, suitability:

? Is there adequate evidence substantiating the suitability of the system?

However, it is noteworthy that EU non-discrimination law does not accept just any method of performance testing that ML practitioners may apply. While it is common practice to assess the performance of ML-based models using a hold-out subset of the training data,[1135] this approach may not suffice as a means of demonstrating suitability.[1136] The reason is that it fails to demonstrate that the performance can be generalised to populations and circumstances that differ from those represented in the training data. It should be noted, however, that there may be

[1134] Line Farah et al., "Are Current Clinical Studies on Artificial Intelligence-Based Medical Devices Comprehensive Enough to Support a Full Health Technology Assessment? A Systematic Review," *Artificial Intelligence in Medicine* 140, no. 102547 (2023), https://doi.org/10.1016/j.artmed.2023.102547.

[1135] Section 1.5.9

[1136] The demographic composition of a test dataset is often similar to that of the training datasets, in which case models with disparate accuracy across different groups may achieve a high accuracy level when evaluated against the test dataset.

different views on this interpretation of the suitability requirement. Some scholars may hold a more lenient view of this requirement in relation to AI systems.[1137]

11.4.6 Conclusion: Suitability of Deploying a Biased AI-CDS System

Deployment of a biased AI-CDS system that potentially puts a protected group at a particular disadvantage, requires that the system nonetheless performs well enough to realise the specific aim it is intended to achieve. Furthermore, to be suitable, safeguards should be in place that are capable of mitigating the consequences of inaccurate, disadvantageous outputs from an AI-CDS system. Such safeguards should particularly include state-of-the-art human oversight solutions, enabling a clinician's review of disadvantageous outputs. Suitability also requires that target variables are medically appropriate and not ingrained with historical inequalities.

In order to objectively justify the deployment of a biased AI-CDS system, it is crucial to establish a substantiated connection between the system and the legitimate aim it is intended to achieve. For a biased AI-CDS system to be 'suitable', it must be capable of meeting a demonstrable need. This requires relying on data that objectively demonstrates the appropriateness of the system in pursuing the desired aim. The standards and scientific rigour required of such data are important topics of further research and development.

11.5 Necessity: Searching for Alternatives to an AI-CDS System

11.5.1 The Necessity Requirement

In addition to being suitable, the deployment of a biased and potentially discriminatory AI-CDS system must also be 'necessary.' To meet this criterion, the deployment of the system must not be a measure that goes beyond what is required to achieve the legitimate aim

[1137] In the context of commercial contracting, Hacker argues that the suitability criterion merely relates to the predictive accuracy of the model and that the standard is easily met in cases of algorithmic bias, because predictive accuracy is always measured: Hacker (2018) 1162; Similarly, Xenidis and Senden argue that "[a]lgorithms are in fact developed precisely to ensure a level of precision and granularity that human minds are not able to reproduce. Hence, the requirements of a legitimate aim and the appropriateness of an algorithm meeting that aim are likely to be satisfied": {Xenidis, 2020 #1971@22 (SSRN preprint, https://ssrn.com/abstract=3529524)}

pursued.[1138] According to established CJEU jurisprudence, the necessity test involves assessing whether there are alternative measures available that would be suitable to achieve the intended aim while being less discriminatory.[1139]

The CJEU states in some rulings that a measure must be "strictly necessary" to be justifiable, suggesting that there cannot exist less discriminatory means capable of achieving the same legitimate aim, at all.[1140] However, the Court's jurisprudence is not consistent in terms of how 'strictly' the necessity criterion should be interpreted. The *CHEZ* ruling appears to indicate that an alternative measure must be "as effective" as the disputed measure, to render the disputed measure discriminatory.[1141] In academic literature, Hacker assumes that the necessity assessment is oriented towards "similarly effective" measures.[1142] As a general rule, this is a reasonable interpretation. A 'stricter' interpretation would imply that the necessity assessment must consider any measure that might achieve the same aim, even if it does so less effectively than the disputed measure. While situations may occur where such a strict interpretation is warranted,[1143] this is arguably not the general rule.[1144]

[1138] e.g., Léger, C-528/13, para. 59; Comune Di Gesturi, C-670/18, para. 46; YS, C-223/19, para. 65.

[1139] Léger, C-528/13, para. 59.

[1140] e.g., G4s Secure Solutions, C-157/15 para. 42; WABE, Joined Cases C-804/18 and C-341/19, para. 69.

[1141] Léger, C-528/13, para. 123.

[1142] Hacker (2018) 1162. In a constitutional law perspective, Barak notes that an alternative that does not fully realise the pursued aim does not preclude a finding of necessity: Aharon Barak, "Proportionality (2)," in *The Oxford Handbook of Comparative Constitutional Law*, ed. Michel Rosenfeld and András Sajó (Oxford University Press, 2012), 745.

[1143] In *Léger,* the CJEU stresses that even where no equally or similarly effective measures can be identified, the disputed measure could only be justifiable if there were no less onerous methods of achieving the aim pursued: Léger, C-528/13, para. 65. However, this case is peculiar because the disputed practice constituted potential direct discrimination, cf. section 6.3.2. The strictness of the necessity test indicated here is therefore not necessarily transferrable to indirect discrimination assessments.

[1144] There may be different views on whether less effective measures must be considered. Scholars addressing proportionality and justification in EU privacy and data protection law have assumed otherwise: Lorenzo Dalla Corte, "On Proportionality in the Data Protection Jurisprudence of

In the context of a pre-deployment assessment of an AI-CDS system, a strict necessity requirement could, depending on the circumstances,[1145] mean that the existence of less biased and less effective measures would indicate that the system under assessment should not be deployed. Consequently, many patients may not obtain access to what may be the best performing system, for them. When an advantaged group is brought down to the level of the disadvantaged group, this is often referred to as 'levelling down.'[1146] For example, if an AI-CDS system for spine surgery decisions performs worse for women than for men, this bias could be mitigated either by improving the system's performance for women or reducing the system's performance for men.[1147] The latter solution would amount to 'levelling down.'[1148] EU non-discrimination law arguably does not support levelling down.[1149] Moreover, in the context of AI-CDS systems, levelling down could interfere with the right to health, cf. Article 35 of the EU Charter, of patients negatively affected by the choice of not deploying the best performing version of an AI-CDS system. 'Levelling down' would also align poorly with the EU's ambition of ensuring citizens a high level of protection from health and safety risks

---

the Cjeu," *International Data Privacy Law* 12, no. 4 (2022): 261, https://doi.org/10.1093/idpl/ipac014. ("The necessity test inquires whether there are other means that are suitable for the achievement of the same objective and, at the same time, both less restrictive towards the right restricted, and as effective as the measure tested"); in relation to CJEU jurisprudence under Articles 7 and 8 of the EU Charter, Kouvakas argues for a different approach: Ioannis Kouvakas, "The Watson Case: Another Missed Opportunity for Stricto Sensu Proportionality," *Cambridge L.aw Review* 2 (2017): 179. ("… the existence of a less injurious alternative does not necessarily imply that this alternative has to be chosen if it is less effective in advancing the means pursued by the choices of the legislature").

[1145] In some cases, the system can be deployed with additional safeguards addressing its biases, cf. section 11.5.4.

[1146] Moreau (2004) 311; Fredman (2022) 10;Tobler (2022): 90; Mittelstadt, Wachter, and Russell (2023).

[1147] Barocas and Selbst suggest that efforts to minimise the difference in error rates between groups may involve "unattractive tradeoffs": Barocas and Selbst (2016) 720.

[1148] Mittelstadt et al. point out how the use of certain 'fairness metrics' might lead to 'levelling down': Mittelstadt, Wachter, and Russell (2023).

[1149] Judgment of 14 March, 2018, Stollwitzer, C-482/16, ECLI:EU:C:2018:180, para. 30; Tobler (2022): 90.

associated with AI systems.[1150] It is therefore asserted that EU non-discrimination law does not generally imply that a less effective yet less biased system is preferable to a more biased, higher performing system. However, for the more biased system to be deemed 'necessary,' it is crucial that safeguards addressing the negative consequences of biases are in place, cf. section 11.5.4 below.

Another aspect of the necessity requirement that may depend on how strictly one interprets it, is the question of how onerous an alternative measure can be for an AI provider or deployer. Arguably, there are limits to the costs and efforts a provider or deployer must undertake. In the *Léger* case (decided on the basis of Article 21 of the Charter), the CJEU suggests that the cost of available options should be taken into account when considering alternative measures.[1151] This aligns with a position commonly taken in non-discrimination law scholarship, asserting that the alternatives to a disputed practice must be "reasonable."[1152] Similarly, in legal literature specifically discussing algorithmic discrimination, Hacker notes that one must incur costs that are not unreasonable if this makes it possible to apply a less discriminatory AI model (although Hacker situates this consideration under the 'proportionality *stricto sensu* assessment').[1153] This interpretation makes sense in the light of the CJEU's statements in *Léger*, as referred to above.

[1150] Mittelstadt, Wachter and Russell also believe that it is "inappropriate" to decrease the performance for advantaged groups in clinical decision-making contexts: Mittelstadt, Wachter, and Russell (2023) 15.

[1151] Léger, C-528/13, para. 64.

[1152] Connolly (2011) 184; Ellis and Watson (2012) 170.

[1153] Hacker (2018) 1164.

The existing guidance to the necessity requirement in EU non-discrimination law does not provide one clear and consistent direction for the search for alternative measures. Consequently, there are uncertainties when adapting this requirement for the purpose of developing methodological elements of a pre-deployment discrimination assessment for AI-CDS systems. However, the bottom line is arguably that the relevant alternative measures are those which constitute equally or similarly effective means of achieving the aims pursued. These are the alternative measures that a pre-deployment discrimination assessment should consider. Consequently, the question to raise during the assessment is whether the stated aim of an AI-CDS system can be achieved in a similarly effective, yet less biased manner, at the time of the assessment. If such an alternative measure can be identified, this suggests that the deployment of a biased AI-CDS system may not be necessary.

Methodological element

Consideration, necessity:

? Are there alternative measures that are *equally or similarly* effective in pursuing the legitimate aim while being less onerous for the protected group that is disadvantaged by the system?
- measures not involving the use of AI technologies;
- alternative AI models;
- additional safeguards

When assessing the necessity of deploying a biased AI-CDS system, at least three categories of relevant alternative measures can be identified: (i) measures not involving the use of AI technologies; (ii) the use of alternative AI models, and (iii) the implementation of additional safeguard while deploying the model being assessed. These alternatives are further discussed in the following sections.

11.5.2 Comparison with Decision-Making Without AI

At present, most clinical decisions are made without the assistance of AI systems. Traditional clinical support systems based on hard-coded computer programs are often used. Decision-making procedures that do not involve AI are relevant to consider as alternative measures when determining the necessity of deploying a biased AI-CDS system.

Methodological elements

Considerations – necessity:

- If the aim of deployment is to improve the accuracy of clinical assessments
  - Sufficient performance has already been established as part of the suitability assessment
- If the aim is preventive interventions/personalised care
  - Unlikely that measures not involving AI have similar effectiveness
- If the aim is efficiency or wider access to care
  - need to consider deployer-specific circumstances
- In all cases
  - consider whether there is reason to assume that an alternative not in involving AI is less biased

In relation to commercial AI applications, Hacker has argued that algorithmic decision-making often lacks equally effective alternatives, because AI systems are often used when they outperform humans in performing a given task.[1154] The same argument is applicable in relation to AI-CDS systems. If the purpose of deploying an AI-CDS system is to improve the accuracy of assessments compared to the current status quo, it is expected that an AI-CDS system will outperform human clinicians in its intended task.[1155] Given that such higher performance is demonstrated on the basis of reliable evidence as discussed in section 11.4.5, comparison with decision-making without AI does not have to be included as a separate step in a discrimination assessment of an AI-CDS system intended to improve the accuracy of clinical assessments. Sufficient performance will already have been established as part of the suitability assessment.

Increased accuracy of assessments is not the only rationale for implementing AI-CDS systems into clinical decision-making. However, also when the aim is to facilitate preventive interventions and/or more personalised care, it is unlikely that these objectives can be

[1154] Hacker (2018) 1162.

[1155] Diaz-Asper et al. (2023) 10.

achieved through other means not involving AI technologies. These legitimate aims are enabled by AI technologies, particularly due to AI-CDS systems' ability to autonomously process and interpret new information continuously. Some decisions in this category are made by clinicians under time constraints where they may not be able to adequately process available information, for example in an intensive care setting.

When considering the use of a biased AI-CDS system with the goal of reducing the time and resources required for clinical assessments and improving the efficiency of providing care, the consideration of alternative measures not involving AI could become more challenging. In such cases, one may explore potential strategies for enhancing efficiency through alternative organisational measures. This may involve the implementation of non-AI-based decision support software or the restructuring of clinical workflows to increase efficiency. Consequently, the necessity assessment might depend on factors that differ between healthcare institutions. This is an issue that potentially occurs in several instances when applying the objective justification test in a pre-deployment discrimination assessment. These issues are addressed collectively in section 11.7, which discusses how to approach considerations that depend on deployer-specific circumstances.

When considering alternative measures to a biased AI-CDS system, a fundamental premise is that the alternative measures must be less biased against the group or groups potentially discriminated against. This raises the issue that human clinical assessments themselves may be biased. The extent to which clinical assessments are biased in contexts where AI is not involved, may be impossible to measure. However, one may argue that if the bias observed in an AI-CDS system is the result of historical error disparities in clinical decision-making,[1156] the human alternative may not be likely to be less biased than an AI-CDS system. Regardless of what alternative measure one is considering, it is important to contemplate whether it is indeed likely to be less biased than the AI-CDS system being assessed.

[1156] Section 4.4.2.6.

11.5.3 Comparison with Alternative Models

Methodological elements

Consideration – necessity:

? Could an alternative, less biased model be deployed?
- ➢ Retrain model?
- ➢ Apply debiasing techniques?
- ➢ Replace with another existing model?

One alternative that must be considered is to substitute a model in the AI-CDS system being assessed with another model. In practice, substituting one model with another would mean to retrain the model being assessed, alter the model through debiasing techniques, or replace it with a different model that already exists. Debiasing techniques is an evolving field of research aiming to develop specific techniques for mitigating biases in different types of AI models and applications.[1157] A necessity assessment should include consideration of state-of-the-art debasing techniques relevant to the particular type of model and the intended use of the AI-CDS system.

Retraining a model can be a viable alternative if the bias in the model stems from biased training data,[1158] or if using more appropriate target variables could mitigate the bias.[1159] Retraining can involve running a learning algorithm on a combination of the original training data and new, additional training data, or running it on an entirely new dataset. However, retraining can be a complex and time-consuming process, and a less biased model is not guaranteed. Particularly, if the bias in an AI-CDS system is a result of unequal representation of protected groups in the training data, it might not be feasible to retrain a model based on a

[1157] e.g., Tolga Bolukbasi et al., "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," *Advances in Neural Information Processing Systems* 29 (NIPS 2016) (2016); Lucas Dixon et al., "Measuring and Mitigating Unintended Bias in Text Classification," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* AIES 18 (February 2-3 2018), https://doi.org/10.1145/3278721.3278729; Fahse, Huber, and van Giffen (2021) 103; Anne Lauscher et al., "A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces," *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 05 (2020); Vokinger, Feuerriegel, and Kesselheim (2021) 2.

[1158] Hacker (2018) 1163.

[1159] Retraining with alternative target variables to mitigate bias is demonstrated by Obermeyer et al: Obermeyer et al. (2019).

more diverse dataset. If feasible at all, constructing a dataset with equal representation of various ethnic groups could significantly delay deployment and be associated with substantial costs. It must be considered whether, given the circumstances, it may reasonably be required that an AI provider assumes the costs and efforts of such an alternative measure.

From the perspective of a healthcare institution conducting a pre-deployment discrimination assessment in connection with the purchase of an AI-CDS system from an external vendor, the relevant costs to consider are the costs of opting for such an alternative model. The costs that an AI deployer can reasonably be required to incur to opt for an alternative model must be considered in the light of how much less biased the alternative model would be. What amount of costs and burdens a provider or deployer may reasonably be required to take, is a highly contextual issue which does permit the articulation of a more precise, general criterion.

11.5.4 Additional Safeguards as Alternatives

As mentioned in section 11.4, the safeguards implemented to mitigate negative impacts of errors in an AI system are important to consider when determining an AI system's suitability. However, despite passing the suitability assessment, a biased AI-CDS system can fail the necessity assessment if additional safeguards are available which would make the system less biased towards a protected group. Therefore, the implementation of additional safeguards capable of mitigating the discriminatory effects of biases in an AI-CDS system should be considered as part of the necessity assessment.

Methodological element

Consideration, necessity:

? Could additional safeguards be implemented to mitigate the negative consequences of bias towards a protected group?
- ➢ Enabling personnel to identify errors and mitigate bias in final decisions
- ➢
- ➢

Safeguards in the form of measures that facilitate human oversight of the decision-making process are often relevant to consider. For instance, implementing improved training procedures for the personnel that will use an AI-CDS system can enhance their ability to identify errors. Particularly, the training of user personnel should rely on documentation of the biases that exist in an AI-CDS system, the sources of bias, and who the biases are likely affect.

Additionally, it may be possible to incorporate technical safeguards into an AI-CDS system, as well as human oversight techniques and methods.[1160] One example is the use of confidence measurements that enable the system to communicate its level of certainty for each output it generates.[1161]

Moreover, a post-deployment monitoring plan may serve as an additional safeguard. To some extent, post-deployment monitoring of an AI-CDS system is required by the AIA and the MDR. When post-deployment monitoring efforts are considered as part of an objective justification assessment, it is important to assess whether the post-deployment monitoring efforts would be effective in preventing or, at least, detecting discrimination. Post-deployment monitoring of an AI-CDS system's performance may be especially important if the system has displayed disparate performance across groups during pre-deployment testing, without exceeding the threshold for a "particular disadvantage." In such cases, pre-deployment monitoring should ascertain that the disparate performance does not turn out to be worse than indicated during pre-deployment testing.

Another example of a relevant safeguard is found in the NORspine case study, concerning an AI-CDS system for predicting the outcome of spine surgery. As noted in section 5.2, patients have different preferences when it comes to defining what a successful surgery outcome is. However, when target variables are defined by an AI development team, this means that the definition of a successful outcome is largely made by the development team on behalf of all patients. It is possible that target variables may be defined in a way that reflects the preferences of a development team. These preferences may, for example, be more aligned with the preferences of men than those of women. In the NORspine project, it is being considered how the preferences of individual patients can be included in the shared decision-making process when an AI-CDS system is used. Solutions to account for individual preferences can function as a safeguard against some of the bias that potentially stems from the choice of target variables.

Furthermore, technical solutions that enhance the interpretability and explainability of an AI-CDS system are particularly relevant. However, it should be noted that the safeguards

---

[1160] On the importance of human oversight measures, see section 11.4.2.

[1161] e.g., Guo et al. (2017).

mentioned here are just examples, and further research is needed to explore the potential for various techniques of mitigating the negative effects of biases.

11.5.5 Conclusion: Necessity of Deploying a Biased AI-CDS System

If an AI-CDS system appears to be biased to such an extent that deployment of the system could amount to indirect discrimination, it must be assessed whether deployment is necessary. In the context of a pre-deployment assessment based on EU non-discrimination law, the necessity requirement involves considering whether there are alternative measures available that would achieve the legitimate aim pursued with similar effectiveness as the AI-CDS system being assessed, while being less discriminatory. The most important alternative measures to consider are decision-making procedures that do not involve AI, alternative AI models, and the implementation of additional safeguards. To properly assess the necessity of deploying an AI-CDS system in the light of these alternative measures, a pre-deployment discrimination assessment should be based on updated scientific and technical knowledge. For example, technical safeguards related to human oversight, confidence estimation and bias mitigation are ongoing research topics in ML science and likely to improve in the future.

11.6 Proportionality S*tricto Sensu*

11.6.1 Does a Balancing Act Turn Out in Favour of Deployment?

As mentioned in section 11.2, the objective justification test reflects the principle that any interference with a fundamental right should be proportional to the aims pursued. Proportionality is a central principle in international law on human rights and in constitutional law, and it is recognised as a general principle of EU law.[1162] Despite the term 'proportionality' being absent from the definition of discrimination in the Equality Directives, the objective justification rule reflects key aspects of how proportionality is commonly understood, i.e., legitimacy, suitability and necessity. These aspects form part of the proportionality principle, in a wide sense.[1163] An aspect of proportionality that is less

---

[1162] Wolf Sauter, "Proportionality in Eu Law: A Balancing Act?," *Cambridge Yearbook of European Legal Studies* 15 (2013): 442, https://doi.org/10.5235/152888713809813611.

[1163] Francisco J. Urbina, "A Critique of Proportionality," *American Journal of Jurisprudence* 57 (2012): 49; Aharon Barak, "Proportionality and Principled Balancing," *Law & Ethics of Human Rights* 4, no. 1 (2010), https://doi.org/doi:10.2202/1938-2545.1041. 6; Sauter (2013) 447; Dalla Corte (2022) 261.

prominent in the Equality Directives and the CJEU's case law under these directives, is what is called 'proportionality in a narrow sense' or 'stricto sensu.'[1164] This involves weighing the reasons in favour of a measure against its detrimental effects.[1165] In other words, it is a balancing act, where the advantages of the disputed practice must justify its harm.[1166]

This type of balancing does not often appear as a separate step of the objective justification test in CJEU rulings. Rather, proportionality considerations are sometimes embedded in the Court's reasoning pertaining to the suitability and necessity of a disputed measure.[1167] It is possible – as Dalla Corte has suggested about the CJEU's case law in relation to the fundamental rights to privacy and data protection – that the Court leans towards the necessity requirement when concluding that national measures violate the EU Charter, because an assessment of necessity appears as a more judicial assessment, requiring a legally defined threshold of necessity to be met.[1168] In contrast, proportionality *stricto sensu* entails a more value-based and, thus, openly moral and political balancing of interests. In *CHEZ*, however, the CJEU highlights proportionality *stricto sensu* as a standalone part of the objective justification test under the RED: "Furthermore, assuming that no other measure as effective as the practice at issue can be identified, the referring court will also have to determine whether the disadvantages caused by the practice at issue are disproportionate to the aims pursued."[1169]

As often noted in academic literature on proportionality in international and constitutional law, seeing proportionality as a 'balancing' act involves the metaphor of a scale.[1170] However, while a scale can be attuned to a standardised metrics system and determine the weight of an

---

[1164] Barak (2012) 745.

[1165] "Proportionality Stricto Sensu (Balancing)," in *Proportionality: Constitutional Rights and Their Limitations*, Aharon Barak, Cambridge Studies in Constitutional Law (Cambridge: Cambridge University Press, 2012), 340.

[1166] Barak (2010) 6-7; Urbina (2012) 49.

[1167] e.g., proportionality considerations are clearly being taken in the *Ingeniørforeningen i Danmark* case: Ingeniørforeningen I Danmark, C-499/08, para. 39. Della Corte finds that the same occurs in data protection case law, although the prominence of *stricto sensu* proportionality is generally stronger in data protection case law: Sauter (2013) 447; Dalla Corte (2022) 270-71.

[1168] Dalla Corte (2022) 273.

[1169] CHEZ, C-83/14, para. 123.

[1170] Frank M Coffin, "Judicial Balancing: The Protean Scales of Justice," *New York University Law Review* 63, no. 1 (1988): 16 and 19; Barak (2010) 7.

entity in kilograms, how to determine the weight of arguments in favour of implementing a measure interfering with a fundamental right is a contentious issue.[1171]

The balancing required during a pre-deployment discrimination assessment can be seen as one between the interference with the right to non-discrimination and the benefits of deploying an AI-CDS system. These benefits relate to patients, healthcare providers, and society at large. However, such a balancing act is challenging due to the incommensurability of the entities on each side of the scale:[1172] one side of the scale contains interferences with the fundamental right to non-discrimination, whereas the other side contains the prospect of an innovation that would contribute to the realisation of some health-related interest. In this regard, it is worth noting the words used by the CJEU in the *Veselības ministrija* ruling.[1173] This ruling concerns national measures within the field of public health. The CJEU notes that "the health and life of humans rank foremost among the assets and interests protected by the [TFEU]."[1174] Hence, health-related interests are recognised as significant within the EU legal order, which implies that they should be given considerable weight in a proportionality assessment. It is arguable that *not* deploying an AI-CDS system interferes with the right to health of the patients who would benefit from the system. On the other hand, the non-discrimination principle is a fundamental principle of EU law, which must certainly be given due weight in any balancing exercise.

While methodological aspects of proportionality assessments are widely discussed in general, there is limited guidance in relevant jurisprudence or literature when it comes to assessing the proportionality of interferences with the right to non-discrimination specifically.[1175] The following sections reflect on how the proportionality of deploying a biased AI-CDS system can be assessed based on the non-discrimination principle in EU law.

[1171] Barak (2010) 7.

[1172] Incommensurability of the sizes is a common criticism of proportionality thinking in law: Urbina (2012) 54, and the references cited therein.

[1173] Veselības Ministrija, C-243/19.

[1174] Veselības Ministrija, C-243/19, para. 45.

[1175] Nilsson (2020) 127.

11.6.2 Consideration of Disadvantages

When assessing the proportionality of deploying a biased AI-CDS system, the disadvantages incurred by groups potentially discriminated against must be considered and placed on one side of the scale. The various disadvantages discussed in section 9.3 may all be considered at this stage of a pre-deployment assessment. This includes allocational harms such as the negative health impacts of receiving suboptimal treatment, as well as more intangible, representational harms. Of particular importance, representational harms can include potential stigmatisation or prejudice against a group.[1176] CJEU jurisprudence explicitly supports the notion that a stigmatising effect on the disadvantaged group is a strong indication that the negative effects of a practice are not proportionate.[1177] In relation to the assessment methodologies of risk assessment and impact assessment, cf. chapter 7, this suggests that such effects should be considered as severe impacts of deploying a biased AI-CDS system.

Furthermore, it is essential to consider the extent of disadvantage measured according to the quantitative disadvantage measurement discussed in chapter 9. When a pre-deployment discrimination assessment based on the methodological elements developed in this thesis reaches the step of proportionality stricto sensu, the extent of disadvantage has already been measured, and indications of a "particular disadvantage" have been established. Chapter 9 found that the notion of a "particular disadvantage" represents a contextual and somewhat flexible threshold. Arguably, if the disadvantage measured in an AI-CDS system barely exceeds the threshold for a "particular disadvantage" (however defined in each case), it should be easier to justify the system's deployment than if the particular disadvantage is more significant.

11.6.3 Consideration of Benefits

On the other side of the scale, weighing in favour of deploying a biased AI-CDS system, relevant considerations range from macro-level public health benefits to the positive health impacts the system can have for individual patients. The benefits must be considered specifically in the light of the aim that is pursued through deployment of an AI-CDS system. If the aim is to improve the accuracy of clinical assessments, the key consideration is the

---

[1176] The negative effects on the disadvantaged groups must be considered widely, according to CJEU case law, taking into account their legitimate interests: CHEZ, C-83/14, para. 128.

[1177] Ibid; Liu and O'Cinneide (2019): 59.

overall improvement provided by the system. How much is it expected to improve clinical assessments, overall, compared to the status quo at the time of deployment? A marginal overall improvement compared to the status quo should count less than a major leap in diagnostic or prognostic precision. Moreover, it is relevant to consider whether the deployment of an AI-CDS system leads to better assessments for the disadvantaged group compared to the status quo. Despite not being as beneficial for a protected group compared to other persons, does it still improve the accuracy of clinical assessments for the group compared to the current situation before deployment of the AI-CDS system? If it does, this means that the situation of the disadvantaged group is improved by the deployment of the AI-CDS system. This would arguably provide an argument of considerable weight in a proportionality assessment.

If the rationale for deployment is to facilitate more efficient use of healthcare resources, the benefits reaped through increased efficiency are central to the assessment. An AI-CDS system might free up clinicians' time to concentrate more on other tasks, which potentially allows them to pay more attention to each patient. This way, deployment of an AI-CDS system could improve both the quality and efficiency of healthcare delivery, in a broader perspective encompassing more than just the accuracy of clinical assessments. However, it is worth noting that one should probably be careful about not giving too much weight to efficiency-based arguments in a proportionality assessment under EU non-discrimination law.[1178]

If the aim of deployment is to make specialized assessments available to patients in rural or remote areas, one may argue that an AI-CDS system can fill a critical gap by improving the level of care that is locally available to patients in such areas. When it comes to AI-CDS systems intended to support the allocation of scarce resources, there may be arguments relating to efficiency as well as accuracy. Furthermore, if an AI-CDS system enables preventive interventions, the fact that these interventions are not possible without the AI-CDS system is arguably an important argument in favour of deployment.

[1178] Regarding arguments related to efficiency and economy, see section 11.3.2.

11.6.4 Conclusion: Proportionality of Deploying a Biased AI-CDS System

Assessing the proportionality (in a narrow sense) of deploying a biased AI-CDS system requires careful balancing of the abovementioned disadvantages and benefits, as well as other circumstance-specific considerations that come up during a pre-deployment assessment. While it is challenging to adapt the proportionality requirement into a streamlined process that can be applied in all situations, this section has outlined important considerations that should be included, depending on the intended purpose and legitimate aim pursued by deploying an AI-CDS system. A broad range of arguments in favour of deployment are relevant in a proportionality assessment. However, when a particular disadvantage is incurred by a protected group, this is an interference with the non-discrimination principle – a general principle of EU law that enjoys a high status within the EU legal order. While this principle generally weighs heavier than arguments pertaining to efficiency and expenditure, public health interests and the right to health of individual patients are also considerable.

Methodological element – proportionality stricto sensu

Considerations:

- □ Could the bias have a stigmatising effect on the protected group?
- □ Consider the quantitative extent of disadvantage measured in accordance with chapter 9;
- □ Would deployment improve the situation for the protected group, despite the comparative disadvantage?
- □ Consider the importance of the legitimate aims pursued
- □ A lack of alternatives indicate that the benefits of deployment may be considerable

In each case, the extent of disadvantage caused by a biased AI-CDS system must be balanced against the benefits achieved by realising the specific aim for which the system is intended. Where feasible, this section has indicated the relative weight of different arguments, based on how the CJEU can be expected to assess them. These indications may serve as guidance for those tasked with assessing the proportionality of deploying a biased AI-CDS system.

11.7 Discussion: Objective Justification Before Deployment of an AI-CDS System: Macro-Level and Local Justifications

This chapter has examined the objective justification requirement in EU non-discrimination law. On the basis thereof, relevant considerations pertaining to legitimate aims, suitability, necessity, and proportionality have been identified and adapted to suit the context of assessing discrimination in an AI-CDS system before its deployment. One fundamental challenge of developing methodological elements applicable in the various situations in which a pre-deployment discrimination assessment may be conducted, relates to the fact that the objective justification assessment requires considerations that depend on the specific circumstances of different deployers. Hence, prior to the deployment of an AI-CDS system, there are two perspectives that may be relevant. One is a 'local' perspective whereas the other is a 'macro' perspective.

The local perspective involves consideration of the legitimate aims pursued by the specific healthcare institution deploying an AI-CDS system, the alternative measures available to that institution, and the specific disadvantages and benefits of deploying the system within the institution. If a healthcare institution is conducting a pre-deployment discrimination assessment, this local perspective is the only relevant perspective. However, when a pre-deployment discrimination assessment is conducted by an AI provider looking to place an AI-CDS system on the EU market, or by a third party (notified body) as part of a certification process, a macro perspective is required: The assessment needs to contemplate how an AI-CDS system may be used by many different deployers.

What needs to be justified at the macro level is the deployment, or placing on the market, of an AI-CDS system, generally. In a macro perspective, the question is therefore not whether an individual user is pursuing a legitimate aim; the aims pursued may vary from one healthcare institution to another. For instance, consider an AI-CDS system that diagnoses diabetic retinopathy with 95 % accuracy according to the provider's documentation. In one institution, the historical accuracy of clinicians' assessments may be 90 %, in which case it is plausible that deployment of the AI-CDS system is a suitable means of achieving the aim of improving the accuracy of clinical assessments. In another institution, the historical accuracy of assessments may be 97 %, in which case it is less convincing that deployment of the AI-CDS system is a suitable means of improving clinical accuracy. On the other hand, the provider of an AI-CDS system might want to account for the possibility that the second institution does

not deploy the system to improve accuracy, but rather to decrease the time and resources required to produce clinical assessments with similar accuracy levels as the status quo.

When assessing discrimination in an AI-CDS system at the macro level, the need to account for deployer-specific circumstances could be addressed by the system's provider. The provider may include considerations pertaining to the various legitimate aims that deployers might rely on when deploying the system. The documentation accompanying an AI-CDS system could explain how deployment of the system, given its intended purpose, may achieve a legitimate aim in accordance with the non-discrimination principle. It could describe relevant legitimate aims and directions on how to determine whether deployment of the system is a suitable means of pursuing such aims at the local level. This type of information could become a part of the provider's description of the 'intended purpose' of the system.[1179] Arguably, if such guidance is included in the documentation, this could support that placement of the system on the market is objectively justified. It would demonstrate that the provider has reflected on the relevant justifications for deploying the system, and these justifications would become part of the system's intended purpose.

The following table provides an example of information that the provider of an AI-CDS system intended for a diagnostic task could include in the documentation, to facilitate a localised suitability assessment. In this example, the AI-CDS system has a 97 % accuracy rate.

AI-CDS system for diagnosis, 97 % accuracy rate	
Aim of deployment	Suitable for
Improvement of clinical accuracy	• *improvement of diagnostic accuracy in deployers where reported historical accuracy is below 97 %*

[1179] The intended purpose is "the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation": Article 3(12) AIA (EP). The 'intended purpose' is also an important term within the MDR, cf. Article 2(12) MDR.

Improvement of resource efficiency	• *increasing resource efficiency of clinical assessments in deployers where the reported historical accuracy is 97 % or lower;* • *increasing resource efficiency of clinical assessments in deployers where the reported historical accuracy is higher than 97 %, if the deployer finds that the benefits of increased efficiency outweighs the disadvantage of decreased accuracy, considering the safeguards that are in place to maintain the highest possible clinical accuracy*
Increasing access to specialised care	*increasing access to specialised assessments in deployers where patients would otherwise not have access to personnel qualified for the clinical assessments for which the AI-CDS system is intended.*

Table 9: Documentation facilitating a deployer-specific objective justification assessment.

One challenge with applying a macro perspective when assessing justification before deployment is that the opportunities for implementing alternative measures may vary among different deployers. The search for alternative measures is not addressed by the information in Table 9 above. There may be institutional constraints at issue, such as budget limitations, staff skills, infrastructure limitations, or availability of alternative or supplementary technological solutions. As a result, what may be a viable alternative for one deployer may not be feasible for another. For example, a large urban hospital may have more resources and alternative organisational or technical solutions available than a smaller clinic. Therefore, it is possible that the smaller clinic may find it necessary to deploy the same AI-CDS system that the larger institution decides not to deploy because less biased alternatives are available.

If the necessity assessment indicates that an AI-CDS system may be deployed, the final step of the objective justification test is the assessment of proportionality in a narrow sense – weighing the benefits of deploying the system against its disadvantages.[1180] Here, again, challenging considerations may arise at the local, deployer-specific level. For example, pre-

[1180] Section 11.6.

deployment testing of an AI-CDS system might indicate that there is disparate performance to the disadvantage of an ethnic minority group. One could argue that deployment of such a system causes more disadvantage, overall, if it is deployed in a hospital serving a population where many patients from the ethnic minority reside, compared to a hospital serving a more homogenous ethnic population.

Methodological elements

Consideration before placement on the market:

? Does the documentation provide sufficient guidance relating to the justifications that different deployers may rely on to deploy the AI-CDS system?

Consideration before a system is put into service within an individual deployer:

? What are the relevant deployer-specific circumstances in relation to suitability, necessity, and proportionality stricto sensu?

While an objective justification assessment at the macro-level may provide considerations that are crucial to a deployer's decision to deploy an AI-CDS system, it is vital that the macro-level considerations are supplemented by a localised, deployer-specific assessment. This section has outlined some considerations that may potentially be relevant. As a principle, a provider of an AI-CDS system should as far as possible contemplate the assessments that individual deployers have to make, and set out relevant information and guidance in the documentation accompanying an AI-CDS system, before placing it on the market.

11.8 Conclusion

Proportionality assessments are sometimes perceived as rather mysterious exercises of discretion – Urbina calls them "unconstrained moral reasoning,"[1181] and Kumm emphasises their aspect of "general practical reasoning."[1182] Indeed, adapting the objective justification test into methodological elements of a pre-deployment discrimination assessment for AI-CDS systems is not straightforward. As part of a pre-deployment discrimination assessment, the objective justification requirement is an arena for discussion of the advantages and disadvantages of deploying a biased AI-CDS system. Moreover, it encourages consideration

[1181] Urbina (2012) 49.

[1182] Mattias Kumm, "The Idea of Socratic Contestation and the Right to Justification: The Point of Rights-Based Proportionality Review," *Law & Ethics of Human Rights* 4, no. 2 (2010): 147.

of state-of-the-art solutions that may allow a system to be used in a way that strikes a fair balance between opposing arguments, interests and rights.

However, to facilitate a pre-deployment objective justification assessment in practice, there is a need to further integrate the considerations outlined in this chapter into broader assessment methodologies such as risk assessment and impact assessment. In these methodologies, considerations of probability and severity of consequences are central.[1183] It obvious how the considerations outlined here relate to probability of indirect discrimination occurring in an AI-CDS. There is also a need to further develop the relationship between the objective justification assessment and the estimation of severity. Arguably, the objective justification assessment encompasses certain considerations which overlap with the estimation of severity of deploying a biased AI-CDS system. For example, the type of clinical decision at issue, the nature of the harm experienced by disadvantaged patients, and whether there are any stigmatising effects involved, are relevant considerations in a pre-deployment objective justification assessment. At the same time, these considerations arguably illuminate the severity of the impact on the right to non-discrimination of deploying a biased AI-CDS system. However, further integrating the methodological elements of this thesis into assessment methodologies such as risk and impact assessments lies outside the scope of the thesis. As noted in section 1.1, the thesis aims to develop methodological elements of pre-deployment discrimination assessments, i.e., methodological elements which may be relevant to any type of pre-deployment assessment where the non-discrimination principle in EU law must be applied, whether such an assessment is part of a risk assessment, impact assessment or another assessment methodology.

Furthermore, this chapter has highlighted challenges associated with including both the 'macro' perspective and the 'local' perspective when assessing discrimination in an AI-CDS system.[1184] It was proposed that providers should contemplate different legitimate aims that deployers might have and include relevant guidance in the documentation accompanying an AI-CDS system, before placing it on the market. The purpose is to enable local discrimination

[1183] Section 7.6.2.

[1184] Section 11.7.

assessments where the specific circumstances of an individual deployer are accounted for, before the deployer puts the system into service.

Furthermore, this chapter found that safeguards to mitigate the negative effects of a biased AI-CDS system should be considered in a pre-deployment discrimination assessment. Such safeguards are particularly relevant in relation to the suitability and necessity requirements. In this regard, it is worth noting that when the AIA enters into force, an intersection occurs between the objective justification requirement in EU non-discrimination law and the AIA's mandatory requirements for high-risk AI systems. For example, as discussed in section 11.4, the CJEU's ruling in *Ligue des drotis humains* highlights the importance of human oversight measures. Such measures are also required by Article 14 AIA. According to this provision, human oversight shall aim at minimising the risks to health, safety and fundamental rights. In a pre-deployment discrimination assessment, the criteria for objective justification should be read in conjunction with Article 14 AIA, and vice versa. If human oversight measures foreseen by an AI provider are not sufficient to render a biased system objectively justified based on the non-discrimination principle, the provider arguably also fails its obligation pursuant to Article 14 AIA. On the other hand, it is not given that appropriate human oversight measures in accordance with the AIA are sufficient to demonstrate the suitability of a biased AI-CDS system. The relationship between AIA requirements and the objective justification part of a pre-deployment discrimination assessment is worth attention in future research efforts.

The methodological elements of pre-deployment discrimination assessment developed in this chapter are summarised as follows:

Objective Justification Assessment

Step 1: Legitimate aims

Consideration:

? What legitimate aims may be pursued by deploying the AI-CDS system?
- improving the quality of care by increasing predictive accuracy;
- decreasing the time and resources required to produce a clinical assessment;
- facilitating preventive interventions and/or more personalised care;
- increasing access to specialised care

Step 2: Suitability

Effectiveness:

? How effective is the AI-CDS system at achieving the aim pursued?
- Performance testing
  - For implications, the relevant legitimate aims and other deployer-specific circumstances must be consulted, see step 5.

Human Oversight:

? Can disadvantageous outputs be verified by non-automated means?
- Are state-of-the-art human oversight measures in place?
  Implication: If disadvantageous outputs cannot be verified, this provides an argument against deployment of the system.

Appropriateness of Target Variables:

? Are target variables appropriate?
- Is there an adequate medical justification for the choice of target variables?
- Is it likely that a target variable reflects historical inequalities?

Clinical evidence:

? Is there adequate evidence substantiating the suitability of the system?

Step 3: Necessity

Alternative measures:

? Are there alternative measures that are *equally or similarly* effective in pursuing the legitimate aim while being less onerous for the protected group that is disadvantaged by the system?
- measures not involving the use of AI technologies;
- alternative AI models;
- additional safeguards

Measures not involving AI:

□ If the aim of deployment is to improve the accuracy of clinical assessments
- Sufficient performance has already been established as part of the suitability assessment

□ If the aim is preventive interventions/personalised care
- Unlikely that measures not involving AI have similar effectiveness

□ If the aim is efficiency or wider access to care
- need to consider deployer-specific circumstances

□ In all cases
- consider whether there is reason to assume that an alternative not in involving AI is less biased

Alternative models:

? Could an alternative, less biased model be deployed?
- Retrain model?
- Apply debiasing techniques?
- Replace with another existing model?

Additional safeguards:

? Could additional safeguards be implemented to mitigate the negative consequences of bias towards a protected group?
- Enabling personnel to identify errors and mitigate bias in final decisons
-
-

Step 4: Proportionality stricto sensu

Considerations:

- □ Could the bias have a stigmatising effect on the protected group?
- □ Consider the quantitative extent of disadvantage measured in accordance with chapter 9;
- □ Would deployment improve the situation for the protected group, despite the comparative disadvantage?
- □ Consider the importance of the legitimate aims pursued
- □ A lack of alternatives indicate that the benefits of deployment may be considerable

Macro-level and local considerations

Consideration before placement on the market:

- ? Does the documentation provide sufficient guidance relating to the justifications that different deployers may rely on to deploy the AI-CDS system?

Consideration before a system is put into service by an individual deployer:

- ? What are the relevant deployer-specific circumstances in relation to suitability, necessity, and proportionality stricto sensu?

PART V: CONCLUSION

12 Conclusion

12.1 Summary

The deployment of AI systems in healthcare is in an early, yet booming phase. In the near future, the use of AI to support clinical decisions could become quite the mainstream activity for clinicians. The hope is that AI-CDS systems will contribute to more accurate and better-informed clinical decisions, increased accessibility of specialised clinical assessments, more efficient provision of healthcare services, and more personalised, preventive care. The first chapter of this thesis provided a basic introduction to the AI technologies that enable these potential benefits. It also introduced four categories of clinical decisions: diagnosis, treatment recommendation, preventive intervention, and allocation of scarce resources (prioritisation). These are clinical decisions for which the use of AI technologies is being explored in practice.

Amidst all the praise and promise, contemporary discourse and cross-disciplinary scholarship on AI technologies raise profound concerns about the risks associated with these technologies. One major concern relates to the risk that biases in AI systems might cause discrimination. This concern is shared by the EU legislature, which has proposed the AI Act – a common European framework setting out requirements that AI systems, their providers, and their deployers, must comply with before placing an AI system on the market or putting it into service. Aiming to ensure the effective protection of safety and fundamental rights, the AI Act requires that providers or deployers (as relevant) take certain preventive measures before the deployment of 'high-risk' AI systems, including AI-CDS systems. In keeping with what the EU legislature frames as a proportional, risk-based approach to the regulation of AI technologies, the standards determining whether an AI system should be deployed are not specified in the AIA. Rather, an AI system may be deployed if AI providers or deployers (as relevant) have taken the preventive measures required by law.

The thesis departed from the assumption that the proposed AI Act would require one or more assessments which, in one form or another, must encompass aspects of discrimination by applying the non-discrimination principle in EU law prior to deployment (termed a 'pre-deployment discrimination assessment'). Given this assumption, it was argued that there is an urgent need for the development of methodologies operationalising the non-discrimination

principle in a pre-deployment assessment context. Against this background, the main objective of the thesis was defined as developing the considerations, principles and criteria upon which a pre-deployment discrimination assessment should be predicated in accordance with the non-discrimination principle.

While political agreement has not been reached regarding the AI Act, an interpretation of the European Parliament's Compromise Text confirmed that the Parliament's position would entail three types of pre-deployment discrimination assessments: risk assessment, impact assessment, and data bias examination. These assessment requirements all refer to the non-discrimination principle in EU law and are, therefore, pre-deployment discrimination assessment requirements. The analysis in chapter 7 found that these requirements refer to three slightly different assessment methodologies, with a common emphasis on assessing the probability that discrimination may occur if an AI-CDS system is deployed. In addition, risk and impact assessment methodologies involve considering the consequences or impact of deployment on the right to non-discrimination. Hence, In the context of these requirements, a discrimination assessment is not aimed at labelling an AI-CDS system as either 'discriminatory' or 'non-discriminatory.' Instead, chapter 7 concluded that the proposed discrimination assessment requirements imply that probability of discrimination and impacts on right to non-discrimination will guide the decision on whether to deploy an AI-CDS system.

The backdrop explaining why pre-deployment discrimination assessments have been proposed in the AI Act, is the concern about biases in AI systems potentially causing discrimination. 'Bias' is, however, not a self-explanatory concept. Chapter 3 of this thesis examined this concept, finding that it refers to often related, yet slightly varying phenomena, which are studied across different scientific disciplines. As a foundational definition to be used within this thesis, chapter 3 defined bias as a systematic difference in treatment (including perception, representation, prediction, action and decision) of certain people or groups in comparison to others. This definition is inspired by the ISO's standard on bias in AI systems,[1185] which was deemed appropriate in the light of the conceptual analysis conducted

[1185] *Nek Iso/Iec Tr 24027:2021,* (ISO, 2021).

in chapter 3. Based on this definition, chapter 3 also outlined the relationship between bias and discrimination.

Chapter 4 introduced the notion of 'equality,' which was used in that chapter as a lens through which the focus of the thesis was sharpened. More specifically, the focus was directed towards the types of biases that are relevant from an equality perspective. Subsequently, to understand the extent and mechanisms of bias as an equality-related problem in the specific context of AI for clinical decision-making, the most important sources of such biases were discussed and categorised. The sources of bias outlined in chapter 4 have served as recurring reference points throughout the thesis.

Through two case studies, chapter 5 further demonstrated how equality-related biases may occur and cause potential discrimination in AI-CDS systems. The fictional case of Simon Tesfay v University Hospital of Storevik (UHS) presented the perspective of an individual discovering by chance that he might have been discriminated against in respect of the allocation of scarce resources. The other case study built on an actual innovation project led by the North Norwegian University Hospital (UNN), referred to in this thesis as the 'NORspine' project because the training data are collected from a Norwegian health registry called NORspine. Both case studies have been referred to, where relevant, throughout the thesis, to illustrate the considerations and challenges involved in assessing discrimination in AI-CDS systems.

Part IV of the thesis analysed the non-discrimination principle in EU law and developed methodological elements of assessing discrimination based on this principle. The sources of interpretation of the non-discrimination principle typically contemplate ex post enforcement contexts, where it is alleged or suspected that discrimination has occurred. Pre-deployment discrimination assessments are emerging as a new venue for application of this principle. Moreover, the CJEU's non-discrimination jurisprudence has dealt with cases involving practices that are markedly different from AI systems, and rarely have these cases concerned healthcare or clinical decision-making. Therefore, there is an acute need for the development of pre-deployment assessment methodologies that incorporate a proper interpretation of the non-discrimination principle, while being adapted to the specific context of AI-CDS systems. This thesis has contributed by developing considerations that should be included in a pre-deployment discrimination assessment, as well as the principles and criteria that should guide

the outcomes of those considerations and, thus, the decision on whether to deploy an AI-CDS system.

As the first step of a pre-deployment discrimination assessment, chapter 8 proposed that the assessor should examine whether any protected characteristics or inextricably linked factors (PILFs) are included as feature variables in a model employed in an AI-CDS system ('Feature Identification'). The purpose of Feature Identification is to determine whether the deployment of an AI-CDS system might lead to direct discrimination. Relevant methods of Feature Identification depend on whether a model is interpretable to such an extent that feature variables can be readily observed. In interpretable models, the assessor should consider whether the feature variables include protected characteristics such as sex or ethnicity, or factors that are so closely linked to these characteristics that they cannot reasonably be separated (e.g., the melanin levels in a patient's skin and ethnicity). In addition, Feature Identification involves checking that a model does not (based on test data) produce disadvantageous outputs exclusively for a protected group or entirely excludes a protected group from being advantaged. If any of these effects are found in test data, this suggests that the model relies on features or combinations of features that are inextricably linked to a protected characteristic. Consequently, there is a potential for direct discrimination, according to this thesis's analysis of CJEU jurisprudence.

In less interpretable models, it was suggested that Feature Identification should involve considering the possibility that PILFs might be relied on through hidden inferences. It was argued that hidden inferences in AI-CDS systems can lead to PILFs being used in ways that amount to direct discrimination. Neural networks based on deep learning can create internal representations that correspond to protected characteristics, even if developers are not aware of it. Certain relevant technical approaches to determining whether hidden inferences of protected characteristics have occurred, were identified. If a Feature Identification process concludes that there are no PILFs included in an AI-CDS system, this indicates that the system is what EU non-discrimination law calls 'apparently neutral.' In such cases, a pre-deployment discrimination assessment should proceed to assess indirect discrimination in the system.

In contrast, if PILFs are relied on as feature variables, the next step is to conduct a direct causation assessment. The purpose of direct causation assessment is to determine whether an AI-CDS system is influenced by PILFs to such a degree that patients disadvantaged by the

system would be directly discriminated against. It was proposed in chapter 10 that an approach in line with CJEU jurisprudence would be to ascertain the relative importance of PILFs within a model through counterfactual testing and explanations. Potentially relevant counterfactual testing methods from ML literature were pointed out.

If sufficient direct causation between a PILF and the outputs of an AI-CDS system is established, the implication depends on which type of clinical decision the system is intended for. If the system is intended to support clinical decisions pertaining to the allocation of scarce resources (as defined in section 1.7.5), direct causation between a PILF and a model's outputs indicates that deployment would be tantamount to direct discrimination. On the other hand, if an AI-CDS system is intended for diagnosis, treatment recommendation, or preventive intervention, it must be considered whether there is an adequate medical justification for the use of a PILF as a feature variable. The reason for this distinction is because receiving a medically justified diagnosis, treatment recommendation or intervention, does not qualify as a disadvantage in EU non-discrimination law. This was established in section 9.4.1. In contrast, not receiving a scarce resource that one could benefit from, is a disadvantage. Therefore, if the output that keeps one from receiving a certain resource is caused by the reliance on a PILF as a feature variable, there is direct causation and, consequently, direct discrimination.

The distinction between scarce resource allocation and the three other categories of clinical decisions mentioned above illustrates that some of the considerations developed in this thesis have clear implications for the decision on whether to deploy an AI-CDS system. If there is direct causation between a PILF and the outputs produced by a system intended for scarce resource allocation, the outcome of a pre-deployment discrimination assessment is arguably given, regardless of which underlying assessment methodology the assessor applies. In contrast, other considerations within the methodological elements developed herein lead to arguments for or against deployment, which feed into the broader assessment and must be balanced against each other.

When direct discrimination in an AI-CDS system has been assessed, the focus of the assessment should shift towards indirect discrimination. The indirect discrimination assessment should concentrate on determining whether an AI-CDS system is biased to such an extent that it would be likely to put a protected group at a particular disadvantage ('Disadvantage measurement') and, if so, whether there is an objective justification for deploying the system. The methodological elements of Disadvantage Measurement were

developed in chapter 9, based on an interpretation of the Equality Directives' prohibition on indirect discrimination.

The objective of Disadvantage Measurement is to measure the relative disadvantage produced by an AI-CDS system, based on pre-deployment testing. However, before the disadvantage can be measured, it was argued that an appropriate dataset should be constructed based on the geographical areas where an AI-CDS system is intended to be deployed. Within such a dataset, Disadvantage Measurement presupposes that it is possible to map the patients who are in comparable situations in respect of the relevant clinical decision. Two basic criteria for comparability were proposed in chapter 9: 'Ground Truth Comparability' and 'Feature Comparability.' Pros and cons associated with each criterion were discussed. Depending on the circumstances, they may be applied as complementary criteria.

Determining comparability according to these criteria is a complex matter that raises profound medical-ethical questions about which factors one should consider when comparing two patients. While this thesis has argued that EU non-discrimination law requires such a comparison, it is not clear how the relevant medical-ethical considerations can be effectively implemented into a pre-deployment discrimination assessment methodology. Nor is it certain that the technical methods suggested for comparability determination can be effectively applied for this purpose. Pointing out the need for such technical methods and identifying potential candidates, as this thesis has done, is an important step. However, the feasibility of applying those methods in the context of a pre-deployment discrimination assessment is recommended as a topic of future research.

When an appropriate dataset is constructed for comparison purposes and patients in comparable situations within this dataset have been mapped, the measurement of disadvantage can commence. It was argued in chapter 9 that several measurement methods may be compatible with EU non-discrimination law. However, one should consider whether protected groups are equally represented in the test dataset. If there is unequal representation, it is important that disadvantage is measured by considering both the advantaged and disadvantaged groups. Relevant methods may include considering the composition of the advantaged and disadvantaged group, and 'Probability Comparison,' i.e., comparing the probability that a patient from a protected group will receive an advantageous output to the probability that other persons will receive such an output. It was highlighted that, depending

on the clinical decision at issue, the measurement method may be adjusted to prioritise either false negatives or false positives.

The threshold for indirect discrimination in EU non-discrimination law is met when a protected group is put at a "particular disadvantage." While the assessment of whether there is a "particular disadvantage" may take a broader set of considerations into account, CJEU jurisprudence suggests that a "particular disadvantage" is primarily a quantitative threshold. In an ex post enforcement context, this threshold must be exceeded before indirect discrimination can be found. However, even though EU non-discrimination law implies that there is a quantitative threshold, it does not specify exactly where the threshold is. It was indicated in chapter 9 that the threshold for a "particular disadvantage" seems to lie within a certain range; if the probability of a patient from a protected group of receiving an advantageous output is less than 60 % of that of other patients, this is a strong indication that there is a particular disadvantage. This would suggest that an objective justification assessment is required before deployment. If the relative probability is closer to 75 %, it is more debatable whether there is a particular disadvantage, in which case qualitative considerations such as the severity of receiving a disadvantageous output should be taken into account. Moreover, the size of dataset used for comparison should be considered. In larger datasets where there are many cases available for comparison, it is arguable that the threshold for finding a particular disadvantage is lower than when a smaller dataset is applied.

If pre-deployment testing indicates that an AI-CDS system would be likely to cause a particular disadvantage, the implication is that an objective justification assessment should be conducted before a decision is made on whether to deploy an AI-CDS system. The objective justification assessment consists of four components: determining that the deployment of an AI-CDS system may pursue one or more legitimate aims, assessing the suitability of the system in pursuing such aims, assessing the necessity of pursuing the relevant aims through deployment of the AI-CDS systems versus alternative solutions, and an assessment of proportionality in a narrow sense. Methodological elements of conducting these assessments in a pre-deployment context were developed in chapter 11.

Chapter 11 highlighted the importance of distinguishing between considerations that AI providers should be expected to undertake before they place an AI-CDS system on the market, and considerations that must be conducted locally, taking into account the specific circumstances of each deployer. It was proposed that a pre-deployment assessment conducted

for the purpose of placing an AI-CDS system on the market should ascertain that the provider's documentation offers sufficient guidance relating to the justifications that different deployers may rely on. Pre-deployment assessments conducted in relation to individual deployers should consider what the relevant deployer-specific circumstances are in relation to suitability, necessity, and proportionality in a narrow sense. This applies regardless of whether an assessment is conducted by a healthcare institution acting as deployer or by an external auditor or authority. Furthermore, chapter 11 discussed relevant differences between deployers and their potential implications for the pre-deployment assessment.

12.2 Practical Implications

A pre-deployment discrimination assessment is far from a guarantee that an AI-CDS system will not cause discrimination when it is deployed. Certification based on the AI Act requirements should not be perceived as a certificate of non-discrimination. The CE marking on an AI-CDS system merely signifies that an assessment has been conducted which considers the risk of discrimination, the impact on the right to non-discrimination, and the probability of discrimination arising due to data biases.[1186] Therefore, post-deployment measures aimed at ensuring non-discrimination are important supplements to the pre-deployment discrimination assessments that this thesis has concentrated on.

The methodological elements developed in this thesis amount to a large set of considerations which are at times highly technical and at times partially medical in nature. These methodological elements are fundamentally based on legal interpretation. The application of them in practice should also be guided by legal interpretation. One practical implication is that conducting a pre-deployment discrimination assessment based on the proposed methodological elements is likely to be a complex and resource-demanding task. It requires highly specialised legal expertise, which must be combined with expertise in machine learning and medicine. For small to medium enterprises and public healthcare institutions with limited budgets, it is questionable whether it is realistic to demand an assessment based on the methodological elements suggested in this thesis.

[1186] This assumes that the pre-deployment assessment requirements found in the European Parliament's Compromise Text are adopted.

Another implication is that pre-deployment discrimination assessments (and subsequent post-deployment monitoring) based on the proposed methodological elements entail a continuous search for the most recent medical knowledge, as well as state-of-the-art machine learning techniques. This can particularly be illustrated by the suitability component of the objective justification assessment, discussed in chapter 11. The *CHEZ* and *Léger* rulings underscore the importance of demonstrating the suitability of a potentially discriminatory practice through updated, reliable and relevant knowledge. This knowledge might change between the time when a pre-deployment discrimination assessment is conducted and the time when an individual claim of discrimination is filed after an AI-CDS system has been deployed. In the example of developing an AI-CDS system to support spine surgery decisions, medical knowledge might initially suggest that sex is a predictive factor for surgery outcomes. The assumption might, for example, be that women have worse outcomes than men because their spines are different and more vulnerable to surgery. Before deployment, it might therefore be deemed suitable that an AI-CDS system incorporates sex as a feature variable. However, if subsequent research indicates that disparities in outcomes are due to the fact that male surgeons tend not to understand the female anatomy as well as they understand the male, the suitability of the AI-CDS system may be called into question. This highlights the importance of supplementing a pre-deployment discrimination assessment with measures ensuring the continuous reassessment of knowledge and evidence supporting the use of potentially biased AI-CDS systems.

Relatedly, awareness of potential patterns that might exist in training data is extremely important in the development of non-discriminatory AI-CDS systems and in the assessment of discrimination in these systems. Examples of patterns suggested in medical research, which have been highlighted in this thesis, include the correlation between patient-surgeon sex concordance,[1187] the tendency to under- or overestimate pain reported by different ethnic groups,[1188] the higher prevalence of false positives in psychiatric diagnosis of certain ethnic groups,[1189] and the higher prevalence of false negatives in diagnosis of women presenting

[1187] Wallis et al. (2022).

[1188] Section 4.4.2.6; Hoffman et al. (2016).

[1189] Trierweiler et al. (2000); Hampton (2007).

with symptoms of heart disease, compared to men.[1190] More such patterns probably exist which have not yet been detected by the medical scientific community.

Given the proliferation of AI-CDS systems, the detection of patterns that AI developers should be aware of, should be a prioritised topic of medical research. However, such patterns are also likely to be discovered during the development and testing of AI-CDS systems. Therefore, a database should be established for the registration of patterns that might cause equality-related biases in AI-CDS systems, to facilitate the sharing of knowledge about such patterns, even if the patterns are not demonstrated in scientific research publications. This way, development teams could check the database for patterns of relevance to the type of clinical decision that they are planning on developing an AI system for.

Furthermore, it is important to note that the methodological elements developed in this thesis do not amount to a consistent, directly applicable framework for conducting a pre-deployment assessment of discrimination in an AI-CDS system. They cannot easily be implemented into an internally coherent decision tree streamlining the process of assessing discrimination in a pre-deployment setting. These methodological elements are intended to be further developed and adapted so that they may be operationalised as part of broader pre-deployment assessment methodologies, such as risk assessment, impact assessment, and data bias examination. As such, the methodological elements offered in this thesis could provide a significant contribution to the development of various methodologies where the non-discrimination principle in EU law is to be applied.

Given the objective of the thesis, it has focussed on developing considerations that should be involved in *assessing* discrimination in AI-CDS systems rather than the considerations that AI developers should make when developing these systems. At the same time, assessing models during development is part of a typical AI development process. In section 1.5.9, a typical development process was illustrated by Figure 2:

[1190] Section 4.4.2.6; NOU 2023: 5, 51; Vokinger and Gasser (2021) 738.

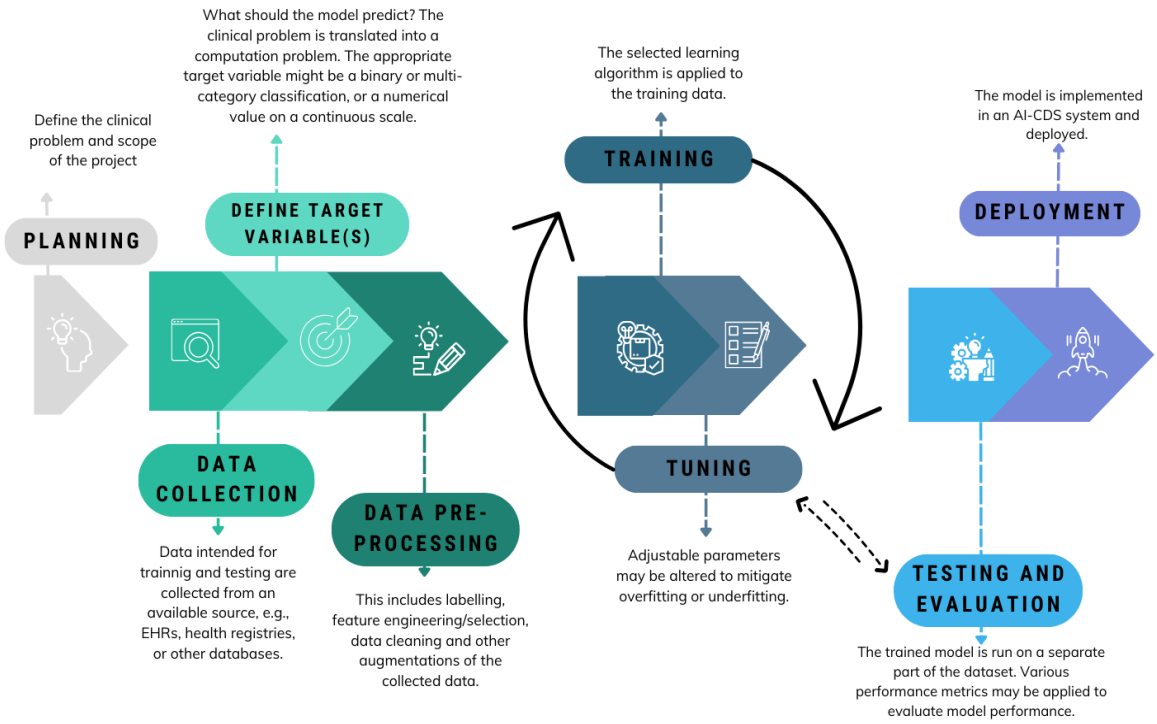


Figure 2.[1191]

The methodological elements of pre-deployment discrimination assessments developed in this thesis directly relate to the 'testing and evaluation' component of such a development process. The various methods that the thesis has suggested may be included in a pre-deployment discrimination assessment are, in practice, different ways of testing an AI-CDS system. In addition to the direct implications that the contributions from this thesis have for testing and evaluation, the insights produced in the thesis may impact other stages of the development process, given that developers aim to develop models that will be assessed positively. The objective of the thesis does not require a comprehensive mapping of the proposed methodological elements' implications for AI development as such. This could, instead, be an interesting topic of future research. However, with reference to the development stages in Figure 2, it is worth outlining some main implications in terms of considerations which should be included in each stage up until and including the data-preprocessing stage. These are the stages impacted by the most important implications of this thesis (not including the 'testing and evaluation' stage, which the thesis's methodological elements are directly aimed at). In addition, compared to Figure 2, a separate step titled 'safeguard implementation' is

[1191] Section 1.5.9.

added. To what extent the following points are already included in existing AI development methodologies may be a topic of future research:

- **Planning:** The intended purpose of an AI-CDS system should be defined in terms of the clinical decision it is intended for. Furthermore, the legitimate aims that may be pursued by deploying the system should be articulated.[1192]
- **Data collection:** Consider the sources of equality-related biases discussed in section 4.4.2 (data bias) to prevent or mitigate equality-related biases. Consider how a dataset for comparison purposes may be constructed according to section 9.5, to ensure that data is collected which enable the measurement of disadvantage during testing.
- **Define target variables:** Consider the sources of equality-related biases discussed in section 4.4.3.4 (choice of target variable) to prevent or mitigate equality-related biases. Document the medical rationale for the choice of target variables, to facilitate an assessment of its suitability.[1193] Consider what it is that defines whether an output from the AI-CDS system is disadvantageous to individual patients, to facilitate disadvantage measurement.[1194]
- **Data pre-processing:** Consider the sources of equality-related biases discussed in section 4.4.3.2 (feature selection), 4.4.3.5 (labelling), and 4.4.4 (hidden inferences), to prevent or mitigate equality-related biases. Consider whether PILFs are included as feature variables and, if so, document the medical rationale for using the relevant PILFs in relation to the clinical decision at issue, to prevent direct discrimination.[1195]
- **Safeguard implementation:** Consider the need for technical safeguards that may be integrated into an AI-CDS system and organisational safeguards that deployers may use to prevent biases in the system to result in discrimination. Review and document relevant state-of-the-art safeguards, to facilitate objective justification of biases. For example, as discussed in sections 11.4 and 11.5, the implementation of human oversight measures may be decisive in an objective justification assessment.

[1192] Section 11.3.

[1193] Section 11.4.

[1194] Section 9.3, cf. section 9.6.

[1195] Section 8.6, cf. section 10.4

12.3 Directions for Future Research

The thesis has pointed out several instances where the development of discrimination assessment methodologies could benefit from targeted research efforts aimed at developing such methodologies. Particularly, the thesis has identified potentially relevant technical methods of scrutinising the various aspects of an AI-CDS system that a discrimination assessment requires. The feasibility of applying these methods as envisaged herein, should be explored through further interdisciplinary research. The most important research developments that this thesis encourages, are the following:

- Research and development of technical methods to determine whether protected characteristics such as sex or ethnicity may be inferred from a dataset;
- Research and development of technical methods for the testing of model behaviour to determine whether protected characteristics are relied on by opaque models, including further research on the use of Testing with Concept Activation Vectors and related methods;[1196]
- Research and development of technical methods for mapping comparable patients within a dataset, including various forms of cluster analysis and propensity score matching;[1197]
- Research connecting the methods for measuring disadvantages in accordance with EU non-discrimination law[1198] with technical fairness metrics relied on in ML literature;
- Research and development of technical methods of establishing the relative importance of protected characteristics in a model. Particularly, counterfactual testing and explanations should be further developed, and it should be investigated to what extent it is feasible to do this through methods such as Individual Conditional Expectation, Partial Dependence Plots, or Causal[1199]

The thesis has encountered several instances where the interpretation of current law is highly uncertain. It must be expected that legal research on the application of EU non-discrimination law to AI-CDS systems in a pre-deployment setting will encounter uncertainties. All the

[1196] Section 8.6.5.

[1197] Section 9.5.4.5.

[1198] Chapter 9.

[1199] Section 10.4.5.

following issues may not be solvable through legal research, but they may nonetheless be worth pursuing, to improve the current understanding. It is equally important, however, that these issues are considered at the legislative level and in relation to the shaping of best practices in the field of pre-deployment discrimination assessments. It is therefore proposed that the following issues should be considered in future research, legislative efforts, or in the development of best practices:

- Clarifying the standards that apply in respect of the reliability of evidence supporting the suitability of a biased AI-CDS system. For instance, is it sufficient that the regular performance requirements in the AI Act and the MDR are satisfied, when the performance of an AI-CDS system is considered as part of an objective justification test?
- Clarifying how to define the thresholds for when a "particular disadvantage" occurs in relation to different types of clinical decisions;
- Clarifying how much influence a PILF must have on the outputs produced by an AI model for there to be direct causation;
- Clarifying to what extent AI providers should define an 'intended target population,' according to which criteria a target population can be defined, and the implications of defining a target population. Particularly, does the definition of a target population imply that an AI-CDS system can be placed on the EU market while only being intended to be used in certain geographical areas?
- Mapping the personal data processing required to operationalise the methodological elements proposed in this thesis, and investigating the legal basis and data protection implications of such processing;
- Analysing the mandatory scope of the DPIA requirement in Article 35 GDPR, particularly as regards the extent to which the consideration of discrimination is mandatory under this provision;
- Further development and integration of the methodological elements offered in this thesis into the methodologies of risk assessment, impact assessment, data bias examination and, potentially, other pre-deployment assessment methodologies.
- Analysis of case law from the European Court of Human Rights (which was not prioritised in this thesis), for the purpose of further developing the methodological elements;

- Adjusting the methodological elements proposed in this thesis to suit other sectors beyond healthcare (it may be assumed that the contributions of this thesis have considerable utility also outside of healthcare).

Table of References

(Excluding statutes, international treaties, and case law)

Books and electronic books:

Allport, Gordon W. *The Nature of Prejudice*. Reading, Mass: Addison-Wesley, 1954.

Aven, Terje, and Ortwin Renn. *Risk Management and Governance: Concepts, Guidelines and Applications.* Risk, Governance and Society. Edited by Jeryl L. Mumpower and Ortwin Renn. Heidelberg: Springer, 2010. doi:10.1007/978-3-642-13926-0.

Aven, Terje, Willy Røed, Hermann Steen Wiencke, and Eivind Vetlesen. *Risikoanalyse: Prinsipper Og Metoder, Med Anvendelser*. 2nd ed. Oslo: Universitetsforlaget, 2017.

Baldwin, Robert, Martin Cave, and Martin Lodge. *The Oxford Handbook of Regulation.* Oxford: Oxford University Press, 2010.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning.* fairmlbook.org, 2019. http://www.fairmlbook.org.

Beck, Ulrich. *Risk Society: Towards a New Modernity*. London: Sage Publications, 1992.

Becker, Gary S. *The Economics of Discrimination.* 2nd ed. Edited by Milton Friedman. Chicago: The University of Chicago Press, 1971. 1957.

Benjamin, Ruha. *Race after Technology: Abolitionist Tools for the New Jim Code.* Oxford: Polity, 2019.

Birkeland, Kåre I., Lars Gullestad, Lars Aabakken, and Kari C. Toverud. *Indremedisin : 1.* Vol. 1, Drammen: Vett & Viten, 2017.

Bishop, Christopher M, and Nasser M Nasrabadi. *Pattern Recognition and Machine Learning.* Vol. 4: Springer, 2006.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies.* 1st ed.: Oxford University Press, 3 September, 2014.

Brown, Rupert. *Prejudice: Its Social Psychology.* 2nd ed. Chichester: Wiley-Blackwell, 2010.

Bunge, Mario. *Causality and Modern Science*. 4th ed. New York: Routledge, 2017. doi:10.4324/9781315081656.

Bygrave, Lee Andrew. *Data Privacy Law: An International Perspective*. Oxford: Oxford University Press, 2014. doi:10.1093/acprof:oso/9780199675555.001.0001.

Connolly, Michael. *Discrimination Law*. 2nd ed. London: Sweet & Maxwell, 2011.

Delgado, Richard, Jean Stefancic, and Angela Harris. *Critical Race Theory (Third Edition): An Introduction.* New York: New York: NYU Press, 2017.

Dovidio, John F., Miles Hewstone, Peter Glick, and Victoria M. Esses. *The Sage Handbook of Prejudice, Stereotyping and Discrimination.* London: SAGE Publications, 2010.

Ellis, Evelyn, and Philippa Watson. *Eu Anti-Discrimination Law*. 2nd ed. ed. Oxford: Oxford University Press, 2012.

Flach, Peter. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data.* Cambridge University Press, 2012.

Foster, Nigel. *Foster on Eu Law.* 7th ed. Oxford: Oxford University Press, 2019.

Fredman, Sandra. *Discrimination Law.* 2nd ed. Oxford: Oxford University Press, 2011.

———. *Discrimination Law*. Clarendon Law Series. 3rd ed. Oxford: Oxford University Press, 2022. doi:10.1093/oso/9780198854081.001.0001.

Gellert, Raphaël. *The Risk-Based Approach to Data Protection.* Oxford University Press, 2020.
Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. New York, NY: Springer New York, 2017.
Hellborg, Sabina. *Diskrimineringsansvar : En Civilrättslig Undersökning Av Förutsättningarna För Ansvar Och Ersättning Vid Diskriminering.* Skrifter Från Juridiska Fakulteten I Uppsala. Vol. 136, Uppsala: Iustus förlag, 2018.
Hellum, Anne, and Vibeke Blaker Strand. *Likestillings- Og Diskrimineringsrett.* 1. utgave. ed. Oslo: Gyldendal, 2022.
Hepple, B. A. *Equality: The New Legal Framework.* Oxford: Hart, 2011.
Hervey, Tamara K. *Justifications for Sex Discrimination in Employment.* London: Butterworth, 1993.
Hillson, David, and British Standards Institution. *The Risk Management Universe: A Guided Tour.* 2nd ed. London: BSI, 2007.
Hoecke, Mark van. *Methodologies of Legal Research : Which Kind of Method for What Kind of Discipline?* 1st ed. Oxford,Portland, Oregon: Hart Publishing, 2011.
Hood, Christopher, Henry Rothstein, and Robert Baldwin. *The Government of Risk: Understanding Risk Regulation Regimes.* Oxford: Oxford University Press, 2010.
Kahneman, Daniel. *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux, 2011.
Kahneman, Daniel, Olivier Sibony, and Cass R Sunstein. *Noise: A Flaw in Human Judgment.* New York: Little, Brown Spark, 2021.
Kamath, Uday, John Liu, and James Whitaker. *Deep Learning for Nlp and Speech Recognition.* Cham, Switzerland: Springer Nature Switzerland AG, 2019. doi:https://doi.org/10.1007/978-3-030-14596-5.
Khaitan, Tarunabh. *A Theory of Discrimination Law.* Oxford: Oxford University Press, 2015.
Lee, Peter, Carey Goldberg, and Isaac Kohane. *The Ai Revolution in Medicine: Gpt-4 and Beyond.* Pearson, 2023.
Liddell, Roderick, and Michael O'Flaherty. *Handbook on European Non-Discrimination Law 2018 Edition.* European Union Agency for Fundamental Rights and Council of Europe, 2018. doi:doi:10.2811/792676.
Makkonen, Timo. *Equal in Law, Unequal in Fact: Racial and Ethnic Discrimination and the Legal Response Thereto in Europe.* The Erik Castrén Institute Monographs on International Law and Human Rights Series. Edited by Martti Koskenniemi. Leiden: BRILL, 2012. http://ebookcentral.proquest.com/lib/tromsoub-ebooks/detail.action?docID=842208.
McColgan, Aileen. *Discrimination, Equality and the Law*. Oxford: Hart Publishing, 2014.
Mitchell, Tom M. *Machine Learning.* Vol. 1: McGraw-hill New York, March 1, 1997.
Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* 2nd ed. Cambridge, Massachusetts: The MIT Press, 2018.
Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-Centered Ai.* Shelter Island: Simon and Schuster, 2021.
Nielsen, Ruth. *Civilretlige Diskriminationsforbud.* København: Jurist- og Økonomforbundets Forlag, 2010.
O'neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown, 2016.
Parker, Christine. *The Open Corporation: Effective Self-Regulation and Democracy.* Cambridge University Press, 2002. doi:10.1017/CBO9780511550034.010.
Pasquale, Frank. *Black Box Society: The Secret Algorithms That Control Money and Information.* Cambridge, Massachussetts: Harvard Universit Press, 2016.
Pearce, Dennis, Enid Campbell, and Don Harding. *Australian Law Schools: A Discipline Assessment for the Commonwealth Tertiary Education Commission* Canberra: Australian Government Publishing Service, 1987.
Rawls, John. *A Theory of Justice.* revised (1999) ed. Cambridge, Massachusetts: Harvard University Press, 1971.
Schauer, Frederick. *Profiles, Probabilities, and Stereotypes.* Harvard University Press, 2006.
Solanke, Iyiola. *Discrimination as Stigma: A Theory of Anti-Discrimination Law*. Oxford, England,Portland, Oregon: Hart Publishing, 2017.

Solove, Daniel J. *Understanding Privacy.* Cambridge, Massachussetts: Harvard University Press, 2008.
Tajfel, Henri. *Differentiation between Social Groups : Studies in the Social Psychology of Intergroup Relations.* European Monographs in Social Psychology. Vol. 14, London: Academic Press, 1978.
Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence.* Knopf Publishing Group, 2017.
Theobald, Oliver. *Machine Learning for Absolute Beginners: A Plain English Introduction.* 3rd edition, Kindle edition ed.: Scatterplot Press, 31 December, 2020.
Twining, William. *Law in Context: Enlarging a Discipline.* Oxford: Clarendon Press, 1997.
Weimer, Maria. *Risk Regulation in the Internal Market: Lessons from Agricultural Biotechnology.* Oxford Studies in European Law. Edited by Paul Craig and Gráinne de Búrca. Oxford: Oxford University Press, 2019.

Journal articles and conference papers:

Abdel-Jaber, Hussein, Disha Devassy, Azhar Al Salam, Lamya Hidaytallah, and Malak EL-Amir. "A Review of Deep Learning Algorithms and Their Applications in Healthcare." *Algorithms* 15, no. 71 (2022): 1-55. https://doi.org/10.3390/a15020071.
Alsaleh, Mohanad M., Freya Allery, Jung Won Choi, Tuankasfee Hama, Andrew McQuillin, Honghan Wu, and Johan H. Thygesen. "Prediction of Disease Comorbidity Using Explainable Artificial Intelligence and Machine Learning Techniques: A Systematic Review." *International Journal of Medical Informatics* 175, no. 105088 (2023): 1-9. https://doi.org/10.1016/j.ijmedinf.2023.105088.
Adams-Prassl, Jeremias, Reuben Binns, and Aislinn Kelly-Lyth. "Directly Discriminatory Algorithms." *The Modern Law Review* 86, no. 1 (2023): 144-75. https://doi.org/10.1111/1468-2230.12759.
Adamson, Adewole S, and Avery Smith. "Machine Learning and Health Care Disparities in Dermatology." *JAMA dermatology* 154, no. 11 (2018): 1247-48. https://doi.org/10.1001/jamadermatol.2018.2348.
Allen, Robin, and Dee Masters. "Artificial Intelligence: The Right to Protection from Discrimination Caused by Algorithms, Machine Learning and Automated Decision-Making." *ERA Forum* 20, no. 4 (2020): 585-98. https://doi.org/10.1007/s12027-019-00582-w. https://doi.org/10.1007/s12027-019-00582-w.
Anderson, K. O., C. R. Green, and R. Payne. "Racial and Ethnic Disparities in Pain: Causes and Consequences of Unequal Care." *Journal of Pain* 10, no. 12 (December 2009): 1187-204. https://doi.org/10.1016/j.jpain.2009.10.002.
Angell, Rico, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. "Themis: Automatically Testing Software for Discrimination." Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Lake Buena Vista, FL, USA, Association for Computing Machinery, 2018.
Anguera, A., J. M. Barreiro, J. A. Lara, and D. Lizcano. "Applying Data Mining Techniques to Medical Time Series: An Empirical Case Study in Electroencephalography and Stabilometry." *Computational and Structural Biotechnology Journal* 14 (2016): 185-99. https://doi.org/10.1016/j.csbj.2016.05.002.
Antoniadi, Anna Markella, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. "Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review." *Applied Sciences* 11, no. 5088 (2021): 1-23. https://doi.org/10.3390/ app11115088.
Arcaya, Mariana C., Alyssa L. Arcaya, and S. V. Subramanian. "Inequalities in Health: Definitions, Concepts, and Theories." *Global Health Action* 8, no. 27106 (2015): 1-12. https://doi.org/10.3402/gha.v8.27106.

Areia, Miguel, Yuichi Mori, Loredana Correale, Alessandro Repici, Michael Bretthauer, Prateek Sharma, Filipe Taveira, *et al.* "Cost-Effectiveness of Artificial Intelligence for Screening Colonoscopy: A Modelling Study." *The Lancet Digital Health* (April 2022): 1-9. https://doi.org/10.1016/ S2589-7500(22)00042-5.

Atkinson, Joe. "Automated Management, Digital Discrimination, and the Equality Act 2010." *Green's Employment Law Bulletin*, no. 159 (2020): 3-6.

Atrey, Shreya. "Race Discrimination in Eu Law after Jyske Finans: Case C-668/15, Jyske Finans a/S V. Ligebehandlingsnævnet, Acting on Behalf of Ismar Huskic, Judgment of the Court (First Chamber) of 6 April 2017 Eu: C: 2017: 278." *Common Market Law Review* 55, no. 2 (2018): 625-41.

Austin, P. C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46, no. 3 (May 2011): 399-424. https://doi.org/10.1080/00273171.2011.568786.

Avery, Joseph J., and Joel Cooper. "Racial Bias in Post-Arrest and Pre-Trial Decision Making: The Problem and a Solution." *Cornell Journal of Law and Public Policy* 29 (2019): 257-94.

Baert, Stijn, and Ann-Sophie De Pauw. "Is Ethnic Discrimination Due to Distaste or Statistics?". *Economics Letters* 125 (2014): 270-73. https://doi.org/10.1016/j.econlet.2014.09.020.

Balsa, Ana I., and Thomas G. McGuire. "Prejudice, Clinical Uncertainty and Stereotyping as Sources of Health Disparities." *Journal of Health Economics* 22, no. 1 (2003): 89-116. https://doi.org/10.1016/S0167-6296(02)00098-X.

Balsa, Ana I., Thomas G. McGuire, and Lisa S. Meredith. "Testing for Statistical Discrimination in Health Care." *Health Services Research* 40, no. 1 (2005): 227-52. https://doi.org/10.1111/j.1475-6773.2005.00351.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6773.2005.00351.x.

Bambauer, Jane R, Tal Zarsky, and Jonathan Mayer. "When a Small Change Makes a Big Difference: Algorithmic Fairness among Similar Individuals." *UC Davis Law Review* 55, no. 4 (2021): 2337-420.

Banerjee, Imon, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, *et al.* "Reading Race: AI Recognises Patient's Racial Identity in Medical Images." *arXiv preprint arXiv:2107.10356* (2021).

Barak, Aharon. "Proportionality and Principled Balancing." *Law & Ethics of Human Rights* 4, no. 1 (2010): 1-16. https://doi.org/doi:10.2202/1938-2545.1041.

Barnard, Catherine, and Bob Hepple. "Substantive Equality." *Cambridge L.aw Journal* 59, no. 3 (2000): 562-85.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact Essay." *California Law Review* 104, no. 3 (June 2016): 671-732. https://doi.org/10.15779/Z38BG31. 732.

Bartlett, Robert P, Adair Morse, Nancy Wallace, and Richard Stanton. "Algorithmic Discrimination and Input Accountability under the Civil Rights Acts." *Berkeley Technology Law Journal* 36, no. 2 (2021): 675-736. https://doi.org/10.15779/Z38N7XN5B.

Bartz, Deborah, Tanuja Chitnis, Ursula B. Kaiser, Janet W. Rich-Edwards, Kathryn M. Rexrode, Page B. Pennell, Jill M. Goldstein, *et al.* "Clinical Advances in Sex- and Gender-Informed Medicine to Improve the Health of All: A Review." *JAMA internal medicine* 180, no. 4 (2020): 574-83. https://doi.org/10.1001/jamainternmed.2019.7194.

Bathaee, Yavar. "The Artificial Intelligence Black Box and the Failure of Intent and Causation." *Harvard Journal of Law & Technology* 31, no. 2 (Spring 2017): 889-938.

Bau, David, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. "Network Dissection: Quantifying Interpretability of Deep Visual Representations." *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017): 6541-49.

Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, *et al.* "Think Your Artificial Intelligence Software Is Fair? Think Again." *IEEE Software* 36, no. 4 (2019): 76-80. https://doi.org/10.1109/MS.2019.2908514.

Besson, Samantha. "Gender Discrimination under Eu and Echr Law: Never Shall the Twain Meet?". *Human Rights Law Review* 8, no. 4 (2008): 647-82. https://doi.org/10.1093/hrlr/ngn023.

Bhopal, Raj S. "Racism in Health and Health Care in Europe: Reality or Mirage?". *European Journal of Public Health* 17, no. 3 (2007): 238-41. https://doi.org/10.1093/eurpub/ckm039.
Billig, Michael, and Henri Tajfel. "Social Categorization and Similarity in Intergroup Behaviour." *European Journal of Social Psychology* 3, no. 1 (1973): 27-52. https://doi.org/10.1002/ejsp.2420030103.
Binns, Reuben. "Fairness in Machine Learning: Lessons from Political Philosophy." *Proceedings of Machine Learning Research* 81, no. 1 (2018): 149-59.
Black, Eleanor, and Robyn Richmond. "Improving Early Detection of Breast Cancer in Sub-Saharan Africa: Why Mammography May Not Be the Way Forward." *Globalization and Health* 15, no. 1 (2019): 3. https://doi.org/10.1186/s12992-018-0446-6.
Black, Julia. "Decentring Regulation: Understanding the Role of Regulation and Self-Regulation in a 'Post-Regulatory' World." *Current Legal Problems* 54, no. 1 (2001): 103-46. https://doi.org/10.1093/clp/54.1.103.
Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach. "Language (Technology) Is Power: A Critical Survey of "Bias" in Nlp." *arXiv preprint arXiv: 2005:14050* (2020): 1-23.
Blount, Kelly. "Using Artificial Intelligence to Prevent Crime: Implications for Due Process and Criminal Justice." *AI & SOCIETY* (2022): 1-10. https://doi.org/10.1007/s00146-022-01513-z.
Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings." *Advances in Neural Information Processing Systems* 29 (NIPS 2016) (2016): 1-9. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.
Borgesius, Frederik J. Zuiderveen. "Strengthening Legal Protection against Discrimination by Algorithms and Artificial Intelligence." *The International Journal of Human Rights* 24, no. 10 (2020): 1572-93. https://doi.org/10.1080/13642987.2020.1743976. **(Borgesius (2020 A))**
Borgesius, Frederik Zuiderveen. "Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law." *European Business Law Review* 31, no. 3 (2020): 401-22. **(Borgesius (2020 B))**
Bornstein, Stephanie. "Antidiscriminatory Algorithms." *Alabama Law Review* 70 (2018): 519.
Bouchagiar, George. "The Long Road toward Tracking the Trackers and De-Biasing: A Consensus on Shaking the Black Box and Freeing from Bias." *Review of European Studies* 11, no. 1 (2019): 27-50. https://doi.org/10.5539/res.v11n1p27.
Braiek, Houssem Ben, and Foutse Khomh. "On Testing Machine Learning Programs." *Journal of Systems and Software* 164, no. 110542 (2020): 1-18. https://doi.org/10.1016/j.jss.2020.110542.
Braithwaite, John. "Enforced Self-Regulation: A New Strategy for Corporate Crime Control." *Michigan Law Review* 80, no. 7 (1982): 1466-507. https://doi.org/10.2307/1288556.
Breslau, Naomi, Glenn C. Davis, Patricia Andreski, Edward L. Peterson, and Lonni R. Schultz. "Sex Differences in Posttraumatic Stress Disorder." *Archives of General Psychiatry* 54, no. 11 (1997): 1044-48. https://doi.org/10.1001/archpsyc.1997.01830230082012.
Brown, Shea, Jovana Davidovic, and Ali Hasan. "The Algorithm Audit: Scoring the Algorithms That Score Us." *Big Data & Society* (January-June 2021): 1-8.
Browne, Annette J., Colleen M. Varcoe, Sabrina T. Wong, Victoria L. Smye, and Koushambhi B. Khan. "Can Ethnicity Data Collected at an Organizational Level Be Useful in Addressing Health and Healthcare Inequities?". *Ethnicity & Health* 19, no. 2 (2014): 240-54. https://doi.org/10.1080/13557858.2013.814766. https://doi.org/10.1080/13557858.2013.814766.
Bygrave, Lee A. "Security by Design: Aspirations and Realities in a Regulatory Context." *Oslo Law Review* 8, no. 3 (2021): 126-77. https://doi.org/10.18261/olr.8.3.2.
Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, Association for Computing Machinery, 2015.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356, no. 6334 (2017): 183-86. https://doi.org/10.1126/science.aal4230.

Calude, Cristian S, and Giuseppe Longo. "The Deluge of Spurious Correlations in Big Data." *Foundations of science* 22, no. 3 (2017): 595-612. https://doi.org/10.1007/s10699-016-9489-4.

Capuzzo, Giacomo. "A Comparative Study on Algorithmic Discrimination between Europe and North-America." *Comparative Law Review* 10, no. 2 (Fall 2019): 125-56.

Cartwright, Dorwin. "Contemporary Social Psychology in Historical Perspective." *Social Psychology Quarterly* 42, no. 1 (1979): 82-93. https://doi.org/10.2307/3033880.

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. "A Clarification of the Nuances in the Fairness Metrics Landscape." *Scientific Reports* 12, no. 4209 (2022): 1-21. https://doi.org/10.1038/s41598-022-07939-1.

Chandler, Chelsea, Peter W Foltz, and Brita Elvevåg. "Using Machine Learning in Psychiatry: The Need to Establish a Framework That Nurtures Trustworthiness." *Schizophrenia Bulletin* 46, no. 1 (2019): 11-14. https://doi.org/10.1093/schbul/sbz105.

———. "Improving the Applicability of AI for Psychiatric Applications through 'Human-in-the-Loop' Methodologies." *Schizophrenia Bulletin* 48, no. 5 (2022): 949-57. https://doi.org/10.1093/schbul/sbac038.

Chen, Po-Hsuan Cameron, Yun Liu, and Lily Peng. "How to Develop Machine Learning Models for Healthcare." *Nature Materials* 18, no. 5 (2019): 410-14. https://doi.org/10.1038/s41563-019-0345-0.

Chen, Richard J, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. "Synthetic Data in Machine Learning for Medicine and Healthcare." *Nature Biomedical Engineering* 5, no. 6 (2021): 493-97. https://doi.org/10.1038/s41551-021-00751-8.

Chicco, Davide, Matthijs J Warrens, and Giuseppe Jurman. "The Coefficient of Determination R-Squared Is More Informative Than Smape, Mae, Mape, Mse and Rmse in Regression Analysis Evaluation." *PeerJ Computer Science* 7, no. e623 (2021). https://doi.org/10.7717/peerj-cs.623.

Choi, Dong-Ju, Jin Joo Park, Taqdir Ali, and Sungyoung Lee. "Artificial Intelligence for the Diagnosis of Heart Failure." *Npj Digital Medicine* 3, no. 54 (2020). https://doi.org/10.1038/s41746-020-0261-3.

Christodoulou, Evangelia, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models." *Journal of clinical epidemiology* 110 (2019): 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

Chynoweth, Paul. "Legal Research." *Advanced research methods in the built environment* 1 (2008).

Cihon, Peter, Moritz J Kleinaltenkamp, Jonas Schuett, and Seth D Baum. "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries." *IEEE Transactions on Technology and Society* 2, no. 4 (2021): 200-09.

Cirillo, Davide, Silvina Catuara-Solarz, Czuee Morey, Emre Guney, Laia Subirats, Simona Mellino, Annalisa Gigante, *et al.* "Sex and Gender Differences and Biases in Artificial Intelligence for Biomedicine and Healthcare." *Npj Digital Medicine* 3, no. 1 (2020): 81. https://doi.org/10.1038/s41746-020-0288-5.

Clayton, Paul D, and George Hripcsak. "Decision Support in Healthcare." *International journal of bio-medical computing* 39 (1995): 59-66.

Clifford, Damian, and Jef Ausloos. "Data Protection and the Role of Fairness." *Yearbook of European Law* 37 (2018): 130-87.

Cobbe, Jennifer, Michelle Seng Ah Lee, and Jatinder Singh. "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems." *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021): 598-609. https://doi.org/10.1145/3442188.3445921.

Coffin, Frank M. "Judicial Balancing: The Protean Scales of Justice." *New York University Law Review* 63, no. 1 (1988): 16-42.

Cofone, Ignacio N. "Algorithmic Discrimination Is an Information Problem." *Hastings Law Journal* 70, no. 6 (2018): 1389.

Coglianese, Cary, and David Lehr. "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era." *Georgetown Law Journal* 105, no. 5 (2016): 1147-224.
Cohen, I. Glenn, Ruben Amarasingham, Anand Shah, Bin Xie, and Bernard Lo. "The Legal and Ethical Concerns That Arise from Using Complex Predictive Analytics in Health Care." *Health Affairs* 33, no. 7 (July 2014): 1139-47. https://doi.org/10.1377/hlthaff.2014.0048.

Combalia, Marc, Noel Codella, Veronica Rotemberg, Cristina Carrera, Stephen Dusza, David Gutman, Brian Helba, *et al.* "Validation of Artificial Intelligence Prediction Models for Skin Cancer Diagnosis Using Dermoscopy Images: The 2019 International Skin Imaging Collaboration Grand Challenge." *The Lancet Digital Health* 4, no. 5 (2022): e330-e39. https://doi.org/10.1016/S2589-7500(22)00021-8.
Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." *Proceedings of KDD '17* (August 13-17 2017): 797-806. https://doi.org/10.1145/3097983.3098095.
Corona Hernández, Hugo, Cheryl Corcoran, Amélie M Achim, Janna N de Boer, Tessel Boerma, Sanne G Brederoo, Guillermo A Cecchi, *et al.* "Natural Language Processing Markers for Psychosis and Other Psychiatric Disorders: Emerging Themes and Research Agenda from a Cross-Linguistic Workshop." *Schizophrenia Bulletin* 49, no. Supplement_2 (2023): S86-S92. https://doi.org/10.1093/schbul/sbac215.
Cowan, Sharon. "“Gender Is No Substitute for Sex”: A Comparative Human Rights Analysis of the Legal Regulation of Sexual Identity." *Feminist Legal Studies* 13, no. 1 (2005): 67-96. https://doi.org/10.1007/s10691-005-1457-2.
Culyer, A. J., and Adam Wagstaff. "Equity and Equality in Health and Health Care." *Journal of Health Economics* 12, no. 4 (1993): 431-57. https://doi.org/10.1016/0167-6296(93)90004-X.
Cutillo, Christine M., Karlie R. Sharma, Luca Foschini, Shinjini Kundu, Maxine Mackintosh, Kenneth D. Mandl, Tyler Beck, *et al.* "Machine Intelligence in Healthcare—Perspectives on Trustworthiness, Explainability, Usability, and Transparency." *Npj Digital Medicine* 3, no. 47 (2020): 1-5. https://doi.org/10.1038/s41746-020-0254-2.
Dalenberg, David Jacobus. "Preventing Discrimination in the Automated Targeting of Job Advertisements." *Computer Law & Security Review* 34, no. 3 (2018): 615-27. https://doi.org/10.1016/j.clsr.2017.11.009.
Dalla Corte, Lorenzo. "On Proportionality in the Data Protection Jurisprudence of the Cjeu." *International Data Privacy Law* 12, no. 4 (2022): 259-75. https://doi.org/10.1093/idpl/ipac014.
Danielsen, Andreas, Morten Fenger, Søren Østergaard, Kristoffer Nielbo, and Ole Mors. "Predicting Mechanical Restraint of Psychiatric Inpatients by Applying Machine Learning on Electronic Health Data." *Acta Psychiatrica Scandinavica* 140 (2019): 147–57. https://doi.org/10.1111/acps.13061.
Danks, David, and Alex John London. "Algorithmic Bias in Autonomous Systems." *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence* 17 (August 2017).
Davenport, Thomas, and Ravi Kalakota. "The Potential for Artificial Intelligence in Healthcare." *Future healthcare journal* 6, no. 2 (2019): 94-98. https://doi.org/10.7861/futurehosp.6-2-94.

Davis, Peter, and Sebastian Felix Schwemer. "Rethinking Decisions under Article 22 of the Gdpr: Implications for Semi-Automated Legal Decision-Making." Paper presented at the Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023), held in conjunction with ICAIL, 2023.

DeCamp, Matthew, and Charlotta Lindvall. "Latent Bias and the Implementation of Artificial Intelligence in Medicine." *Journal of the American Medical Informatics Association* 27, no. 12 (2020): 2020-23. https://doi.org/10.1093/jamia/ocaa094.

Delgado-Rodríguez, M, and J Llorca. "Bias." *Journal of Epidemiology and Community Health* 58, no. 8 (2004): 635-41. https://doi.org/10.1136/jech.2003.008466.
den Heijer, Maarten, Teun van Os van den Abeelen, and Antanina Maslyka. "On the Use and Misuse of Recitals in European Union Law." *Amsterdam Law School Research Paper*, no. 2019-31 (2019).
de Souza Nascimento, Elizamary, Iftekhar Ahmed, Edson Oliveira, Márcio Piedade Palheta, Igor Steinmacher, and Tayana Conte. "Understanding Development Process of Machine Learning Systems: Challenges and Solutions." Paper presented at the 2019 acm/ieee international symposium on empirical software engineering and measurement (ESEM), 2019.
Diaz-Asper, Catherine, Mathias K Hauglid, Chelsea Chandler, Alex S Cohen, Peter W Foltz, and Brita Elvevåg. "A Framework for Language Technologies in Behavioral Research and Clinical Applications: Ethical Challenges, Implications and Solutions (Preprint)." (2023). https://doi.org/10.1037/amp0001195.
Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. "Measuring and Mitigating Unintended Bias in Text Classification." *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* AIES 18 (February 2-3 2018): 67-73. https://doi.org/10.1145/3278721.3278729.
Dobbe, Roel, Sarah Dean, Thomas Gilbert, and Nitin Kohli. "A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics." *arXiv preprint arXiv:1807.00553* (2018). https://arxiv.org/abs/1807.00553.
Dongare, AD, RR Kharde, and Amit D Kachare. "Introduction to Artificial Neural Network." *International Journal of Engineering and Innovative Technology (IJEIT)* 2, no. 1 (July 2012): 189-94.
Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through Awareness." Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, Massachusetts, Association for Computing Machinery, 2012.
Dworkin, Ronald. "What Is Equality? Part 1: Equality of Welfare." *Philosophy & Public Affairs* 10, no. 3 (Summer 1981): 185-246. http://www.jstor.org/stable/2264894.
Elaziz, Mohamed Abd, Khalid M Hosny, Ahmad Salah, Mohamed M Darwish, Songfeng Lu, and Ahmed T Sahlol. "New Machine Learning Method for Image-Based Diagnosis of Covid-19." *Plos one* 15, no. 6 (26 June 2020): 1-18. https://doi.org/10.1371/journal.pone.0235187.
Eriksson, Andrea. "European Court of Justice: Broadening the Scope of European Nondiscrimination Law." *International Journal of Constitutional Law* 7, no. 4 (2009): 731-53. https://doi.org/10.1093/icon/mop025.
Fahse, Tobias, Viktoria Huber, and Benjamin van Giffen. "Managing Bias in Machine Learning Projects." Paper presented at the Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues, 2021.
Fajkovic, Harun, Joshua A. Halpern, Eugene K. Cha, Atessa Bahadori, Thomas F. Chromecki, Pierre I. Karakiewicz, Eckart Breinl, Axel S. Merseburger, and Shahrokh F. Shariat. "Impact of Gender on Bladder Cancer Incidence, Staging, and Prognosis." *World Journal of Urology* 29, no. 4 (2011): 457-63. https://doi.org/10.1007/s00345-011-0709-9.
Farah, Line, Julie Davaze-Schneider, Tess Martin, Pierre Nguyen, Isabelle Borget, and Nicolas Martelli. "Are Current Clinical Studies on Artificial Intelligence-Based Medical Devices Comprehensive Enough to Support a Full Health Technology Assessment? A Systematic Review." *Artificial Intelligence in Medicine* 140, no. 102547 (2023): 1-13. https://doi.org/10.1016/j.artmed.2023.102547.
Favaretto, Maddalena, Eva De Clercq, and Bernice Simone Elger. "Big Data and Discrimination: Perils, Promises and Solutions. A Systematic Review." *Journal of Big Data* 6, no. 12 (2019): 1-27. https://doi.org/10.1186/s40537-019-0177-4. https://doi.org/10.1186/s40537-019-0177-4.
Ferguson, Andrew Guthrie. "Policing Predictive Policing." *Washington University Law Review* 94 (2016): 1109.
Ferretti, Maria Teresa, Maria Florencia Iulita, Enrica Cavedo, Patrizia Andrea Chiesa, Annemarie Schumacher Dimech, Antonella Santuccione Chadha, Francesca Baracchi, *et al.* "Sex Differences in Alzheimer Disease — the Gateway to Precision Medicine." *Nature Reviews Neurology* 14, no. 8 (2018): 457-69. https://doi.org/10.1038/s41582-018-0032-9.

Foy, Kevin Chu, James L. Fisher, Maryam B. Lustberg, Darrell M. Gray, Cecilia R. DeGraffinreid, and Electra D. Paskett. "Disparities in Breast Cancer Tumor Characteristics, Treatment, Time to Treatment, and Survival Probability among African American and White Women." *NPJ breast cancer* 4 (2018 2018): 7. https://doi.org/10.1038/s41523-018-0059-5.

Fredman, Sandra. "The Reason Why: Unravelling Indirect Discrimination." *Industrial Law Journal* 45, no. 2 (2016): 231-43. **(Fredman (2016 A))**

———. "Substantive Equality Revisited." *International Journal of Constitutional Law* 14, no. 3 (2016): 712-38. **(Fredman (2016 B))**

Friedman, Batya, and Helen Nissenbaum. "Bias in Computer Systems." *ACM Transactions on Information Systems* 14, no. 3 (1996): 330-47.

Fusar-Poli, Paolo, Dominic Stringer, Alice MS Durieux, Grazia Rutigliano, Ilaria Bonoldi, Andrea De Micheli, and Daniel Stahl. "Clinical-Learning Versus Machine-Learning for Transdiagnostic Prediction of Psychosis Onset in Individuals at-Risk." *Translational psychiatry* 9, no. 259 (2019): 1-11. https://doi.org/10.1038/s41398-019-0600-9.

Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness Testing: Testing Software for Discrimination." Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, Paderborn, Germany, Association for Computing Machinery, 2017.

Gandy, Oscar H. "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems." *Ethics and Information Technology* 12, no. 1 (2010): 29-42. https://doi.org/10.1007/s10676-009-9198-6.

Gao, Caroline X., Dominic Dwyer, Ye Zhu, Catherine L. Smith, Lan Du, Kate M. Filia, Johanna Bayer, *et al.* "An Overview of Clustering Methods with Guidelines for Application in Mental Health Research." *Psychiatry Research* 327, no. 115265 (2023): 1-28. https://doi.org/10.1016/j.psychres.2023.115265.

Garland, David. "The Rise of Risk." *Risk and morality* 1 (2003): 48-86.

Gautam, Srishti, Marina M-C Höhne, Stine Hansen, Robert Jenssen, and Michael Kampffmeyer. "Demonstrating the Risk of Imbalanced Datasets in Chest X-Ray Image-Based Diagnostics by Prototypical Relevance Propagation." *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* (2022): 1-5. https://doi.org/10.1109/ISBI52829.2022.9761651.

Gellert, R. M. "The Role of the Risk-Based Approach in the General Data Protection Regulation and in the European Commission's Proposed Artificial Intelligence Act. Business as Usual ?". *Journal of Ethics and Legal Technologies* 3 (2021): 15-33.

Gellert, Raphaël Maurice. "Why the Gdpr Risk-Based Approach Is About Compliance Risk, and Why It's Not a Bad Thing." *Trends and Communities of legal informatics: IRIS 2017 - Proceedings of the 20th International Legal Informatics Symposium* (2017): 527-32.

Gerards, Janneke, and Frederik Zuiderveen Borgesius. "Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence Articles and Essays." *Colorado Technology Law Journal* 20 (2022): 1.

Gigerenzer, Gerd. "The Bias Bias in Behavioral Economics." *Review of Behavioral Economics* 5, no. 3-4 (2018): 303-36.

Glauner, Patrick, Petko Valtchev, and Radu State. "Impact of Biases in Big Data." *arXiv preprint arXiv:1803.00897* (2018): 1-10.

Gleser, Malcolm A., and Morris F. Collen. "Towards Automated Medical Decisions." *Computers and Biomedical Research* 5, no. 2 (1972): 180-89. https://doi.org/https://doi.org/10.1016/0010-4809(72)90080-8.

Gonçalves, Maria Eduarda. "The Risk-Based Approach under the New Eu Data Protection Regulation: A Critical Perspective." *Journal of Risk Research* 23, no. 2 (2020): 139-52. https://doi.org/10.1080/13669877.2018.1517381.

Goodman, Bryce. "Discrimination, Data Sanitisation and Auditing in the European Union's General Data Protection Regulation." *European Data Protection Law Review* 2, no. 4 (2016): 493-506.

Gopal, Dipesh P, Ula Chetty, Patrick O'Donnell, Camille Gajria, and Jodie Blackadder-Weinstein. "Implicit Bias in Healthcare: Clinical Practice, Research and Decision Making." *Future healthcare journal* 8, no. 1 (2021): 40-48. https://doi.org/10.7861/fhj.2020-0233.

Goyal, Yash, Amir Feder, Uri Shalit, and Been Kim. "Explaining Classifiers with Causal Concept Effect (Cace)." *arXiv preprint arXiv:1907.07165* (2019): 1-10. https://doi.org/10.48550/arXiv.1907.07165.

Greaves, Lorraine, and Stacey A. Ritz. "Sex, Gender and Health: Mapping the Landscape of Research and Policy." *International Journal of Environmental Research and Public Health* 19, no. 5 (2022): 2563. https://doi.org/10.3390/ijerph19052563.

Green, Alexander R, Dana R Carney, Daniel J Pallin, Long H Ngo, Kristal L Raymond, Lisa I Iezzoni, and Mahzarin R Banaji. "Implicit Bias among Physicians and Its Prediction of Thrombolysis Decisions for Black and White Patients." *Journal of general internal medicine* 22, no. 9 (2007): 1231-38.

Greenwald, Anthony G., Mahzarin R. Banaji, Laurie A. Rudman, Shelly D. Farnham, Brian A. Nosek, and Deborah S. Mellott. "A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem, and Self-Concept." *Psychological Review* 109 (2002): 3-25. https://doi.org/10.1037/0033-295X.109.1.3.

Greenwald, Anthony G., and Linda Hamilton Krieger. "Implicit Bias: Scientific Foundations." *California Law Review* 94, no. 4 (2006): 945-67. https://doi.org/10.2307/20439056.

Grote, Thomas, and Philipp Berens. "On the Ethics of Algorithmic Decision-Making in Healthcare." *Journal of medical ethics* 46 (2020): 205-11. https://doi.org/10.1136/medethics-2019-105586.

Grozdanovski, Ljupcho. "In Search of Effectiveness and Fairness in Proving Algorithmic Discrimination in Eu Law." *Common Market Law Review* 58, no. 1 (2021): 99-136.

Grozev, Rossen. "A Landmark Judgment of the Court of Justice of the Eu - New Conceptual Contributions to the Legal Combat against Ethnic Discrimination." *The Equal Rights Review* Fifteen (2015): 168-87.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. "On Calibration of Modern Neural Networks." Paper presented at the International conference on machine learning, 2017.

Gupta, Rohan, Smita Kumari, Anusha Senapati, Rashmi K Ambasta, and Pravir Kumar. "New Era of Artificial Intelligence and Machine Learning-Based Detection, Diagnosis, and Therapeutics in Parkinson's Disease." *Ageing Research Reviews* 90, no. 102013 (2023): 1-24. https://doi.org/10.1016/j.arr.2023.102013.

Guryan, Jonathan, and Kerwin Kofi Charles. "Taste‐Based or Statistical Discrimination: The Economics of Discrimination Returns to Its Roots." *The Economic Journal* 123, no. 572 (November 2013): 417-32. https://doi.org/10.1111/ecoj.12080.

Hacker, Philipp. "Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under Eu Law." *Common Market Law Review* 55, no. 4 (2018).

Hallinan, Dara Martin, Nicholas. "Fundamental Rights, the Normative Keystone of Dpia." *European Data Protection Law Review* 6, no. 2 (2020): 178-93. https://doi.org/10.21552/edpl/2020/2/6.

Hampton, Michelle DeCoux. "The Role of Treatment Setting and High Acuity in the Overdiagnosis of Schizophrenia in African Americans." *Archives of psychiatric nursing* 21, no. 6 (2007): 327-35. https://doi.org/10.1016/j.apnu.2007.04.006.

Hansson, Sven Ove. "Decision Theory." *A brief introduction. Department of Philosophy and the History of technology. Royal Institute of Technology. Stockholm* (1994).

———. "Ethical Criteria of Risk Acceptance." *Erkenntnis* 59, no. 3 (2003): 291-309. https://doi.org/10.1023/A:1026005915919.

Hauglid, Mathias K. "What's That Noise? Interpreting Algorithmic Interpretation of Human Speech as a Legal and Ethical Challenge." *Schizophrenia Bulletin* 48, no. 5 (2022): 960-62. https://doi.org/10.1093/schbul/sbac008.

Hauglid, Mathias K, and Karl Øyvind Mikalsen. "Tilgang Til Helseopplysninger I Maskinlæringsprosjekter." *Lov og Rett*, no. 7 (2022): 419-39. https://doi.org/10.18261/lor.61.7.3.

Hauglid, Mathias Karlsen, and Tobias Mahler. "Doctor Chatbot: The Eu's Regulatory Prescription for Generative Medical AI." *Oslo Law Review* 10, no. 1 (2023): 1-23. https://doi.org/10.18261/olr.10.1.1.

Hawath, Mariam. "Regulating Automated Decision-Making: An Analysis of Control over Processing and Additional Safeguards in Article 22 of the Gdpr." *European Data Protection Law Review* 7, no. 2 (2021): 161. https://doi.org/10.21552/edpl/2021/2/6.

Henderson, Bradley, Colleen M Flood, and Teresa Scassa. "Artificial Intelligence in Canadian Healthcare: Will the Law Protect Us from Algorithmic Bias Resulting in Discrimination?". *Canadian Journal of Law and Technology* 19, no. 2 (2022): 475-504.

Hepple, Bob. "Enforcing Equality Law: Two Steps Forward and Two Steps Backwards for Reflexive Regulation." *Industrial Law Journal* 40, no. 4 (2011): 315-35. https://doi.org/10.1093/indlaw/dwr020.

Herman, Edward S. "The Institutionalization of Bias in Economics." *Media, Culture & Society* 4, no. 3 (1982): 275-91.

Hofer, Ira S., Christine Lee, Eilon Gabel, Pierre Baldi, and Maxime Cannesson. "Development and Validation of a Deep Neural Network Model to Predict Postoperative Mortality, Acute Kidney Injury, and Reintubation Using a Single Feature Set." *Npj Digital Medicine* 3, no. 58 (2020): 1-10. https://doi.org/10.1038/s41746-020-0248-0.

Hoffman, Kelly M., Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences between Blacks and Whites." *Proceedings of the National Academy of Sciences* 113, no. 16 (2016): 4296-301. https://doi.org/10.1073/pnas.1516047113. https://www.pnas.org/content/pnas/113/16/4296.full.pdf.

Hoffmann, Anna Lauren. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society* 22, no. 7 (2019): 900-15.

Hovy, Dirk, and Shrimai Prabhumoye. "Five Sources of Bias in Natural Language Processing." *Language and Linguistics Compass* 15, no. 8 (2021): e12432. https://doi.org/https://doi.org/10.1111/lnc3.12432.

Hunt, Dereck L, R Brian Haynes, Steven E Hanna, and Kristina Smith. "Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review." *JAMA* 280, no. 15 (1998): 1339-46.

Hunter, Rosemary C, and Elaine W Shoben. "Disparate Impact Discrimination: American Oddity or Internationally Accepted Concept." *Berkeley Journal of Employment and Labor Law* 19, no. 1 (1998): 108-52.

Hutchinson, Terry, and Nigel Duncan. "Defining and Describing What We Do: Doctrinal Legal Research." *Deakin L. Rev.* 17, no. 1 (October 2012): 83-119.

Hyland, Stephanie L., Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, *et al.* "Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning." *Nature Medicine* 26, no. 3 (March 2020): 364-73. https://doi.org/10.1038/s41591-020-0789-4.

Irving, Jessica, Rashmi Patel, Dominic Oliver, Craig Colling, Megan Pritchard, Matthew Broadbent, Helen Baldwin, *et al.* "Using Natural Language Processing on Electronic Health Records to Enhance Detection and Prediction of Psychosis Risk." *Schizophrenia Bulletin* 47, no. 2 (2020): 405-14. https://doi.org/10.1093/schbul/sbaa126.

Janssen, Heleen L. "An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making." *International Data Privacy Law* 10, no. 1 (2020): 76-106. https://doi.org/10.1093/idpl/ipz028.

Jensen, Peter B, Lars J Jensen, and Søren Brunak. "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care." *Nature Reviews Genetics* 13, no. 6 (2012): 395-405. https://doi.org/10.1038/nrg3208.

Jneid, Hani, Gregg C. Fonarow, Christopher P. Cannon, Adrian F. Hernandez, Igor F. Palacios, Andrew O. Maree, Quinn Wells, *et al.* "Sex Differences in Medical Care and Early Death after Acute Myocardial Infarction." *Circulation* 118, no. 25 (2008): 2803-10. https://doi.org/10.1161/CIRCULATIONAHA.108.789800.

Johnson, Kipp W, Jessica Torres Soto, Benjamin S Glicksberg, Khader Shameer, Riccardo Miotto, Mohsin Ali, Euan Ashley, and Joel T Dudley. "Artificial Intelligence in Cardiology." *Journal*

of the American College of Cardiology 71, no. 23 (2018): 2668-79. https://doi.org/https://doi.org/10.1016/j.jacc.2018.03.521.
Kapur, Supriya. "Reducing Racial Bias in AI Models for Clinical Use Requires a Top-Down Intervention." *Nature Machine Intelligence* 3 (2021): 460-60. https://doi.org/10.1038/s42256-021-00362-7.
Kasirzadeh, Atoosa, and Damian Clifford. "Fairness and Data Protection Impact Assessments." Paper presented at the Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021.
Kaushal, Amit, Russ Altman, and Curt Langlotz. "Geographic Distribution of Us Cohorts Used to Train Deep Learning Algorithms." *JAMA* 324, no. 12 (2020): 1212-13. https://doi.org/10.1001/jama.2020.12067.
Khaitan, Tarunabh. "Indirect Discrimination Law: Causation, Explanation and Coat-Tailers." *Law Quarterly Review* 132 (January 2016): 35-41.
Kim, Pauline T. "Data-Driven Discrimination at Work." *William & Mary Law Review* 58, 3 (2016): 857.
———. "Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action." *California Law Review* 110, no. 5 (October 2022): 1539-96. https://doi.org/10.15779/Z387P8TF1W.
Kiseleva, Anastasiya. "AI as a Medical Device: Is It Enough to Ensure Performance Transparency and Accountability in Healthcare?". *European Pharmaceutical Law Review*, no. 1 (2020). https://doi.org/DOI: 10.21 552/epIr/2020/1/4.
Kiseleva, Anastasiya, and Paul Quinn. "Are You AI's Favorite? Eu Legal Implications of Biased Ai Systems in Clinical Genetics and Genomics." *European Pharmaceutical Law Review* 5, no. 4 (2021): 155-74. https://doi.org/10.21552/epIr/2021/4/4.
Kitamura, Felipe C. "Chatgpt Is Shaping the Future of Medical Writing but Still Requires Human Judgment." *Radiology* 307, no. 2 (2023). https://doi.org/10.1148/radiol.230171.
Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10 (2018): 113-74. https://doi.org/10.1093/jla/laz001.
Klimas, Tadas, and Jurate Vaiciukaite. "The Law of Recitals in European Community Legislation." *ILSA J. Int'l & Comp. L.* 15 (2008): 61.
Kline, Adrienne, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. "Multimodal Machine Learning in Precision Health: A Scoping Review." *Nature: NPJ Digital Medicine* 5, no. 1 (2022): 171. https://doi.org/https://doi.org/10.1038/s41746-022-00712-8.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, *et al.* "Racial Disparities in Automated Speech Recognition." *Proceedings of the National Academy of Sciences* 117, no. 14 (2020): 7684-89. https://doi.org/10.1073/pnas.1915768117.
Kohler-Hausmann, Issa. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination." *Northwestern University Law Review* 113, no. 5 (2018): 1163-228.
Koumakis, Lefteris. "Deep Learning Models in Genomics; Are We There Yet?". *Computational and Structural Biotechnology Journal* 18 (2020): 1466-73. https://doi.org/https://doi.org/10.1016/j.csbj.2020.06.017.
Kouvakas, Ioannis. "The Watson Case: Another Missed Opportunity for Stricto Sensu Proportionality." *Cambridge L.aw Review* 2 (2017): 173-82.

Krieger, Linda Hamilton. "The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity." *Stanford Law Review* 47, no. 6 (July 1995): 1161-248.
Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (Tcav)." Paper presented at the International conference on machine learning, 2018.

Kumm, Mattias. "The Idea of Socratic Contestation and the Right to Justification: The Point of Rights-Based Proportionality Review." *Law & Ethics of Human Rights* 4, no. 2 (2010): 142-75.
Lai, Calvin K., Kelly M. Hoffman, and Brian A. Nosek. "Reducing Implicit Prejudice." *Social and Personality Psychology Compass* 7, no. 5 (2013): 315-30. https://doi.org/https://doi.org/10.1111/spc3.12023.
Landers, Richard N, and Tara S Behrend. "Auditing the AI Auditors: A Framework for Evaluating Fairness and Bias in High Stakes AI Predictive Models." *American Psychologist* 78, no. 1 (2023): 36-49. https://doi.org/10.1037/amp0000972.
Landry, Latrice G, and Heidi L Rehm. "Association of Racial/Ethnic Categories with the Ability of Genetic Tests to Detect a Cause of Cardiomyopathy." *JAMA cardiology* 3, no. 4 (2018): 341-45. https://doi.org/10.1001/jamacardio.2017.5333.
Lauscher, Anne, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. "A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces." *Proceedings of the AAAI Conference on Artificial Intelligence* 34, no. 05 (2020): 8131-38.
Lebret, Audrey. "Allocating Organs through Algorithms and Equitable Access to Transplantation—a European Human Rights Law Approach." *Journal of Law and the Biosciences* 10, no. 1 (2023). https://doi.org/10.1093/jlb/lsad004.
LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature* 521 (28 May 2015): 436-44. https://doi.org/10.1038/nature14539. https://doi.org/10.1038/nature14539.
Ledley, Robert S, and Lee B Lusted. "Reasoning Foundations of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason." *Science* 130, no. 3366 (1959): 9-21.
Lee, Peter, Sebastien Bubeck, and Joseph Petro. "Benefits, Limits, and Risks of Gpt-4 as an AI Chatbot for Medicine." *New England Journal of Medicine* 388, no. 13 (2023): 1233-39.
Legato, Marianne J., Paula A. Johnson, and JoAnn E. Manson. "Consideration of Sex Differences in Medicine to Improve Health Care and Patient Outcomes." *JAMA* 316, no. 18 (2016): 1865-66. https://doi.org/10.1001/jama.2016.13995.

Lehr, David, and Paul Ohm. "Playing with the Data: What Legal Scholars Should Learn About Machine Learning." *UC Davis Law Review* 51 (2017): 653-717.
Lenaerts, Koen, and Jose A. Gutierrez-Fons. "To Say What the Law of the Eu Is: Methods of Interpretation and the European Court of Justice." *Columbia Journal of European Law* 20, no. 3 (2014): 3-61.
Lewis, David. "Causation." *The journal of philosophy* 70, no. 17 (1974): 556-67.
Li, Feng, Yuguang Wang, Tianyi Xu, Lin Dong, Lei Yan, Minshan Jiang, Xuedian Zhang, *et al.* "Deep Learning-Based Automated Detection for Diabetic Retinopathy and Diabetic Macular Oedema in Retinal Fundus Photographs." *Eye* 36, no. 7 (2021): 1433-41. https://doi.org/10.1038/s41433-021-01552-8.
MacCarthy, Mark. "Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms." *Cumberland Law Review* 48, no. 1 (2017): 67.
Macenaite, Milda. "The “Riskification” of European Data Protection Law through a Two-Fold Shift." *European Journal of Risk Regulation* 8, no. 3 (2017): 506-40. https://doi.org/10.1017/err.2017.40.
Malgieri, Gianclaudio. "The Concept of Fairness in the Gdpr: A Linguistic and Contextual Interpretation." Paper presented at the Proceedings of the 2020 Conference on fairness, accountability, and transparency, 2020.
Maliszewska-Nienartowicz, Justyna. "Direct and Indirect Discrimination in European Union Law–How to Draw a Dividing Line." *International Journal of Social Sciences* 3, no. 1 (2014): 41-55.
Mann, Monique, and Tobias Matzner. "Challenging Algorithmic Profiling: The Limits of Data Protection and Anti-Discrimination in Responding to Emergent Discrimination." *Big Data & Society* 6, no. 2 (2019), https://doi.org/10.1177/2053951719895805.

Mantelero, Alessandro, and Maria Samantha Esposito. "An Evidence-Based Methodology for Human Rights Impact Assessment (Hria) in the Development of AI Data-Intensive Systems."

Computer Law & Security Review 41, no. 105561 (2021): 1-35. https://doi.org/10.1016/j.clsr.2021.105561.
Marcum, James A. "Clinical Decision-Making, Gender Bias, Virtue Epistemology, and Quality Healthcare." *Topoi* 36, no. 3 (2017): 501-08. https://doi.org/10.1007/s11245-015-9343-2.
Martínez-Ramil, Pablo. "Discriminatory Algorithms. A Proportionate Means of Achieving a Legitimate Aim?". *Journal of Ethics and Legal Technologies* 4, no. 1 (2022).

Mazoué, J. G. "Diagnosis without Doctors." *The Journal of medicine and philosophy* 15, no. 6 (1990): 559-79. https://doi.org/10.1093/jmp/15.6.559.
McCrudden, Christopher. "The New Architecture of Eu Equality Law after Chez: Did the Court of Justice Reconceptualise Direct and Indirect Discrimination?". *European Equality Law Review* preprint (2016): 1-16. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2797587.
Megerian, Jonathan T., Sangeeta Dey, Raun D. Melmed, Daniel L. Coury, Marc Lerner, Christopher J. Nicholls, Kristin Sohl, *et al.* "Evaluation of an Artificial Intelligence-Based Medical Device for Diagnosis of Autism Spectrum Disorder." *Npj Digital Medicine* 5, no. 57 (2022): 1-11. https://doi.org/10.1038/s41746-022-00598-6.
Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys (CSUR)* 54, no. 6 (2021): 1-35. https://doi.org/10.1145/3457607.
Miettinen, Samuli, and Merita Kettunen. "Travaux to the Eu Treaties: Preparatory Work as a Source of Eu Law." *Cambridge Yearbook of European Legal Studies* 17 (2015): 145-67. https://doi.org/10.1017/cel.2015.6.
Mikalsen, Karl Øyvind, Cristina Soguero-Ruiz, Kasper Jensen, Kristian Hindberg, Mads Gran, Arthur Revhaug, Rolv-Ole Lindsetmo, *et al.* "Using Anchors from Free Text in Electronic Health Records to Diagnose Postoperative Delirium." *Computer Methods and Programs in Biomedicine* 152 (2017): 105-14. https://doi.org/10.1016/j.cmpb.2017.09.014.
Mikkelsen, E., T. Ingebrigtsen, A. M. Thyrhaug, L. R. Olsen, P. Nygaard Ø, I. Austevoll, J. I. Brox, *et al.* "The Norwegian Registry for Spine Surgery (Norspine): Cohort Profile." *European Spine Journal* (Sep 17 2023): 3713-30. https://doi.org/10.1007/s00586-023-07929-5.
Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3, no. 2 (2016): 2053951716679679. https://doi.org/10.1177/2053951716679679.
Mittelstadt, Brent, Sandra Wachter, and Chris Russell. "The Unfairness of Fair Machine Learning: Levelling Down and Strict Egalitarianism by Default." *arXiv preprint arXiv:2302.02404* (2023). https://arxiv.org/abs/2302.02404.
Mittermaier, Mirja, Marium M. Raza, and Joseph C. Kvedar. "Bias in AI-Based Models for Medical Applications: Challenges and Mitigation Strategies." *Npj Digital Medicine* 6, no. 113 (2023/06/14 2023): 1-3. https://doi.org/10.1038/s41746-023-00858-z.
Moreau, Sophia R. "The Wrongs of Unequal Treatment." *University of Toronto Law Journal* 54, no. 3 (2004): 291-326.
Mullainathan, Sendhil, and Ziad Obermeyer. "Does Machine Learning Automate Moral Hazard and Error?". *American Economic Review* 107, no. 5 (2017): 476-80. https://doi.org/10.1257/aer.p20171084.
Mökander, Jakob, Maria Axente, Federico Casolari, and Luciano Floridi. "Conformity Assessments and Post-Market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation." *Minds and Machines* 32, no. 2 (2022): 241-68. https://doi.org/10.1007/s11023-021-09577-4.
Nachbar, Thomas B. "Algorithmic Fairness, Algorithmic Discrimination." *Florida State University Law Review* 48, no. 2 (2020): 509-59.

Nelson, Alan. "Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care." *Journal of the National Medical Association* 94, no. 8 (2002): 666-68. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2594273/.
Nemati, Shamim, Andre Holder, Fereshteh Razmi, Matthew D. Stanley, Gari D. Clifford, and Timothy G. Buchman. "An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in

the ICU." *Critical Care Medicine* 46, no. 4 (2018): 547-53. https://doi.org/10.1097/CCM.0000000000002936.

Niessen, Jan. "Making the Law Work-the Enforcement and Implementation of Anti-Dicrimination Legislation." *European Journal of Migration & Law* 5, no. 2 (2003): 249-58.

Nilsson, Anna. "Same, Same but Different: Proportionality Assessments and Equality Norms." *Oslo Law Review* 7, no. 3 (2020): 126-44. https://doi.org/10.18261/ISSN.2387-3299-2020-03-01.

Nordling, Linda. "A Fairer Way Forward for AI in Health Care." 573 (26 September 2019): s103-s05. https://doi.org/10.1038/d41586-019-02872-2.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, *et al.* "Bias in Data-Driven Artificial Intelligence Systems—an Introductory Survey." *WIREs Data Mining and Knowledge Discovery* 10, no. e1356 (2020). https://doi.org/10.1002/widm.1356.

O'Cinneide, Colm. "Fumbling Towards Coherence: The Slow Evolution of Equality and Anti-Discrimination Law in Britain Special Issue on Human Rights and Equality." *Northern Ireland Legal Quarterly* 57, no. 1 (Spring 2006): 57-101.

O'Hare, Ursula. "Enhancing European Equality Rights: A New Regional Framework." *Maastricht Journal of European and Comparative Law* 8, no. 2 (2001): 133-65.

Obermeyer, Ziad, Rebecca Nissan, Michael Stern, Stephanie Eaneff, Emily Joy Bembeneck, and Sendhil Mullainathan. "Algorithmic Bias Playbook." *Center for Applied AI at Chicago Booth* (2021).

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464 (2019): 447-53. https://doi.org/10.1126/science.aax2342.

Onitiu, Daria. "The Limits of Explainability & Human Oversight in the Eu Commission's Proposal for the Regulation on AI- a Critical Approach Focusing on Medical Diagnostic Systems." *Information & Communications Technology Law* 32, no. 2 (2023): 170-88. https://doi.org/10.1080/13600834.2022.2116354.

Orzechowski, Marcin, Marianne Nowak, Katarzyna Bielińska, Anna Chowaniec, Robert Doričić, Mojca Ramšak, Paweł Łuków, *et al.* "Social Diversity and Access to Healthcare in Europe: How Does European Union's Legislation Prevent from Discrimination in Healthcare?". *BMC Public Health* 20, no. 1399 (2020): 1-10. https://doi.org/10.1186/s12889-020-09494-8.

Panch, Trishan, Heather Mattie, and Rifat Atun. "Artificial Intelligence and Algorithmic Bias: Implications for Health Systems." *Journal of Global Health* 9, no. 02318 (2019): 1-5. https://doi.org/10.7189/jogh.09.020318.

Pasquinelli, Matteo. "How a Machine Learns and Fails." *Spheres: Journal for Digital Cultures*, no. 5 (2019): 1-17.

Paulus, Jessica K, and David M Kent. "Race and Ethnicity: A Part of the Equation for Personalized Clinical Decision Making?". *Circulation: Cardiovascular Quality and Outcomes* 10, no. 7 (July 2017). https://doi.org/10.1161/CIRCOUTCOMES.117.003823.

Paulus, Jessica K., and David M. Kent. "Predictably Unequal: Understanding and Addressing Concerns That Algorithmic Clinical Prediction May Increase Health Disparities." *NPJ Digital Medicine* 3, no. 99 (2020): 1-7. https://doi.org/10.1038/s41746-020-0304-9.

Peña Gangadharan, Seeta, and Jędrzej Niklas. "Decentering Technology in Discourse on Discrimination." *Information, Communication & Society* 22, no. 7 (2019): 882-99. https://doi.org/10.1080/1369118X.2019.1593484.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *The american economic review* 62, no. 4 (1972): 659-61.

Placido, Davide, Bo Yuan, Jessica X. Hjaltelin, Chunlei Zheng, Amalie D. Haue, Piotr J. Chmura, Chen Yuan, *et al.* "A Deep Learning Algorithm to Predict Risk of Pancreatic Cancer from Disease Trajectories." *Nature Medicine* 29, no. 5 (2023): 1113-22. https://doi.org/10.1038/s41591-023-02332-5.

Poon, Carmen C. Y., Yuqi Jiang, Ruikai Zhang, Winnie W. Y. Lo, Maggie S. H. Cheung, Ruoxi Yu, Yali Zheng, *et al.* "AI-Doscopist: A Real-Time Deep-Learning-Based Algorithm for

Localising Polyps in Colonoscopy Videos with Edge Computing Devices." *Npj Digital Medicine* 3, no. 1 (2020): 73. https://doi.org/10.1038/s41746-020-0281-z.
Pope, Devin G, and Justin R Sydnor. "Implementing Anti-Discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy* 3, no. 3 (2011): 206-31. https://doi.org/10.1257/pol.3.3.206.
Price, W. Nicholson, II. "Black-Box Medicine." *Harvard Journal of Law & Technology* 28, no. 2 (2015): 419.
Prosperi, Mattia, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, *et al.* "Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare." *Nature Machine Intelligence* 2, no. 7 (2020): 369-75. https://doi.org/10.1038/s42256-020-0197-y.
Purtova, Nadezhda. "The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law." *Law, Innovation and Technology* 10, no. 1 (2018): 40-81. https://doi.org/10.1080/17579961.2018.1452176.
Quelle, Claudia. "Does the Risk-Based Approach to Data Protection Conflict with the Protection of Fundamental Rights on a Conceptual Level? (Preprint)." Available at SSRN 2726073 (2015).
Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine* 169, no. 12 (2018): 866-72. https://doi.org/10.7326/M18-1990.

Rajkomar, Alvin, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, *et al.* "Scalable and Accurate Deep Learning with Electronic Health Records." *Npj Digital Medicine*, no. 18 (2018): 1-10. https://doi.org/10.1038/s41746-018-0029-1.
Raub, McKenzie. "Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices." *Arkansas Law Review* 71, no. 2 (2018): 529-70.
Rezaii, Neguine, Phillip Wolff, and Bruce H Price. "Natural Language Processing in Psychiatry: The Promises and Perils of a Transformative Approach." *The British Journal of Psychiatry*, no. 220 (2022): 251-53. https://doi.org/10.1192/bjp.2021.188.
Ricci Lara, María Agustina, Rodrigo Echeveste, and Enzo Ferrante. "Addressing Fairness in Artificial Intelligence for Medical Imaging." *Nature Communications* 13, no. 4581 (2022): 1-6. https://doi.org/10.1038/s41467-022-32186-3.
Richens, Jonathan G., Ciarán M. Lee, and Saurabh Johri. "Improving the Accuracy of Medical Diagnosis with Causal Machine Learning." *Nature Communications* 11, no. 3923 (2020): 1-9. https://doi.org/10.1038/s41467-020-17419-7.
Rostadmo, Martine. "Svart Hud Er Tykkere Enn Hvit." *Tidsskrift for Den norske legeforening* (2021). https://doi.org/10.4045/tidsskr.21.0058.
Rudovsky, David. "Law Enforcement by Stereotypes and Serendipity: Racial Profiling and Stops and Searches without Cause Symposium: Race Crime and the Constitution." *University of Pennsylvania Journal of Constitutional Law* 3, no. 1 (2001): 296-366.
Röösli, Eliane, Brian Rice, and Tina Hernandez-Boussard. "Bias at Warp Speed: How AI May Contribute to the Disparities Gap in the Time of Covid-19." *Journal of the American Medical Informatics Association* 28, no. 1 (2020): 190-92. https://doi.org/10.1093/jamia/ocaa210.
Sackett, David L, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. "Evidence Based Medicine: What It Is and What It Isn't." 312 (13 January 1996): 71-72.
Salcito, Kendyl, Jürg Utzinger, Gary R Krieger, Mark Wielga, Burton H Singer, Mirko S Winkler, and Mitchell G Weiss. "Experience and Lessons from Health Impact Assessment for Human Rights Impact Assessment." *BMC international health and human rights* 15, no. 24 (2015): 1-12. https://doi.org/10.1186/s12914-015-0062-y.
Sallam, Malik. "Chatgpt Utility in Health Care Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns." *Healthcare* 11, no. 887 (2023): 1-20. https://doi.org/doi.org/10.3390/ healthcare11060887.
Samad, M. D., A. Ulloa, G. J. Wehner, L. Jing, D. Hartzel, C. W. Good, B. A. Williams, C. M. Haggerty, and B. K. Fornwalt. "Predicting Survival from Large echocardiography and Electronic health record Datasets: Optimization with Machine Learning." *JACC Cardiovasc Imaging* 12, no. 4 (April 2019): 681-89. https://doi.org/10.1016/j.jcmg.2018.04.026.

Samuel, Arthur L. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal of research and development* 3, no. 3 (1959): 210-29.
Sanchez, Pedro, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. "Causal Machine Learning for Healthcare and Precision Medicine." *Royal Society Open Science* 9, no. 220638 (2022): 1-15. https://doi.org/10.1098/rsos.220638.
Sarkar, Rahuldeb, Christopher Martin, Heather Mattie, Judy Wawira Gichoya, David J Stone, and Leo Anthony Celi. "Performance of Intensive Care Unit Severity Scoring Systems across Different Ethnicities in the USA: A Retrospective Observational Study." *The Lancet Digital Health* 3, no. 4 (2021): e241-e49. https://doi.org/10.1016/S2589-7500(21)00022-4.
Sauter, Wolf. "Proportionality in Eu Law: A Balancing Act?". *Cambridge Yearbook of European Legal Studies* 15 (2013): 439-66. https://doi.org/10.5235/152888713809813611.
Scherer, Matthew U. "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies." *Harvard Journal of Law & Technology* 29, no. 2 (Spring 2015): 353-400.
Scherer, Matthew U, Allan G King, and Marko J Mrkonich. "Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms." *South Carolina Law Review* 71, no. 2 (2019): 449-522.
Schmidhuber, Jürgen. "Deep Learning in Neural Networks: An Overview." *Neural networks* 61 (2015): 85-117. https://doi.org/http://dx.doi.org/10.1016/j.neunet.2014.09.003.
Schonberg, Soren, and Karin Frick. "Finishing, Refining, Polishing: On the Use of Travaux Preparatoires as an Aid to the Interpretation of Community Legislation." *European law review* 28, no. 2 (2003): 149-71.
Schulman, Kevin A., Jesse A. Berlin, William Harless, Jon F. Kerner, Shyrl Sistrunk, Bernard J. Gersh, Ross Dubé, *et al.* "The Effect of Race and Sex on Physicians' Recommendations for Cardiac Catheterization." *New England Journal of Medicine* 340, no. 8 (1999): 618-26. https://doi.org/10.1056/nejm199902253400806.
Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." *NIST Special Publication* 1270 (2022): 1-77. https://doi.org/10.6028/NIST.SP.1270.
Seiner, Joseph. "Disentangling Disparate Impact and Disparate Treatment: Adapting the Canadian Approach." *Yale Law & Policy Review* 25, no. 1 (Fall 2006): 95. www.jstor.org/stable/40239673.
Selbst, Andrew D. "An Institutional View of Algorithmic Impact Assessments." 35, no. 1 (2021): 117-90.

Sesen, M. Berkan, Timor Kadir, Rene-Banares Alcantara, John Fox, and Michael Brady. "Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer." *AMIA Annual Symposium proceedings* 2012 (2012): 838-47. https://pubmed.ncbi.nlm.nih.gov/23304358

Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in under-Served Patient Populations." *Nature Medicine* 27, no. 12 (2021): 2176-82. https://doi.org/10.1038/s41591-021-01595-0.
Shah, Deven, H Andrew Schwartz, and Dirk Hovy. "Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview." *arXiv preprint arXiv:1912.11078* (2019).
Shavers, Vickie L., Pebbles Fagan, Dionne Jones, William M. P. Klein, Josephine Boyington, Carmen Moten, and Edward Rorie. "The State of Research on Racial/Ethnic Discrimination in the Receipt of Health Care." *American Journal of Public Health* 102, no. 5 (2012): 953-66. https://doi.org/10.2105/ajph.2012.300773.
Sikstrom, Laura, Marta M Maslej, Katrina Hui, Zoe Findlay, Daniel Z Buchman, and Sean L Hill. "Conceptualising Fairness: Three Pillars for Medical Algorithms and Health Equity." *BMJ Health & Care Informatics* 29, no. 1 (2022): 1-11. https://doi.org/10.1136/bmjhci-2021-100459.
Silva, Selena, and Martin Kenney. "Algorithms, Platforms, and Ethnic Bias: An Integrative Essay." *Phylon* 55, no. 1 & 2 (2018): 9-37. https://doi.org/10.2307/26545017.

Sjoding, Michael W, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. "Racial Bias in Pulse Oximetry Measurement." *New England Journal of Medicine* 383, no. 25 (2020): 2477-78. https://doi.org/10.1056/NEJMc2029240.

Smith, Belinda. "How Might Information Bolster Anti-Discrimination Laws to Promote More Family-Friendly Workplaces?". *Journal of Industrial Relations* 56, no. 4 (2014): 547-65. https://doi.org/10.1177/0022185614540128.

Smuha, Nathalie A. "Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea." *Philosophy & Technology* 34, no. 1 (May 2020): 1-14. https://doi.org/10.1007/s13347-020-00403-w.

Snipes, Shedra Amy, Sherrill L Sellers, Adebola Odunlami Tafawa, Lisa A Cooper, Julie C Fields, and Vence L Bonham. "Is Race Medically Relevant? A Qualitative Study of Physicians' Attitudes About the Role of Race in Treatment Decision-Making." *BMC Health Services Research* 11, no. 183 (2011): 1-10. https://doi.org/10.1186/1472-6963-11-183.

Stiggelbout, A M, T Van der Weijden, M P T De Wit, D Frosch, F Légaré, V M Montori, L Trevena, and G Elwyn. "Shared Decision Making: Really Putting Patients at the Centre of Healthcare." *BMJ* 344, no. e256 (2012): 1-6. https://doi.org/10.1136/bmj.e256.

Strümke, Inga, Marija Slavkovik, and Clemens Stachl. "Against Algorithmic Exploitation of Human Vulnerabilities." *arXiv preprint arXiv:2301.04993* (2023).

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, *et al.* "Mitigating Gender Bias in Natural Language Processing: Literature Review." *arXiv preprint arXiv:1906.08976* (2019).

Suresh, Harini, and John V Guttag. "A Framework for Understanding Unintended Consequences of Machine Learning." *arXiv preprint arXiv:1901.10002* (2019).

Taekema, Sanne, and Prof van Klink. "On the Border: Limits and Possibilities of Interdisciplinary Research." *Bart van Klink and Sanne Taekema, Law and Method. Interdisciplinary research into Law (Series Politika, nr 4), Tübingen: Mohr Siebeck* 2011 (2011): 7-32.

Tandon, Neeraj, and Rajiv Tandon. "Will Machine Learning Enable Us to Finally Cut the Gordian Knot of Schizophrenia." *Schizophrenia Bulletin* 44, no. 5 (2018): 939-41. https://doi.org/10.1093/schbul/sby101.

Tannenbaum, Cara, Colleen M. Norris, and M. Sean McMurtry. "Sex-Specific Considerations in Guidelines Generation and Application." *Canadian Journal of Cardiology* 35, no. 5 (2019): 598-605. https://doi.org/10.1016/j.cjca.2018.11.011.

Thompson, Hale M., Brihat Sharma, Sameer Bhalla, Randy Boley, Connor McCluskey, Dmitriy Dligach, Matthew M. Churpek, Niranjan S. Karnik, and Majid Afshar. "Bias and Fairness Assessment of a Natural Language Processing Opioid Misuse Classifier: Detection and Mitigation of Electronic Health Record Data Disadvantages across Racial Subgroups." *Journal of the American Medical Informatics Association* 28, no. 11 (2021): 2393-403. https://doi.org/10.1093/jamia/ocab148.

Timmer, Alexandra. "Toward an Anti-Stereotyping Approach for the European Court of Human Rights." *Human Rights Law Review* 11, no. 4 (2011): 707-38. https://doi.org/10.1093/hrlr/ngr036.

Tobler, Christa, and Kees Waaldijk. "Case C-267/06 Tadao Maruko V Versorgungsanstalt Der Deutschen Bühnen: Judgement of the Grand Chamber of the Court of Justice of 1 April 2008, Not yet Reported." *Common Market Law Review* 46, no. 2 (2009): 723-46.

Tomašev, Nenad, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, *et al.* "A Clinically Applicable Approach to Continuous Prediction of Future Acute Kidney Injury." *Nature* 572, no. 7767 (2019): 116-19. https://doi.org/10.1038/s41586-019-1390-1. https://doi.org/10.1038/s41586-019-1390-1.

Tosoni, Luca. "The Right to Object to Automated Individual Decisions: Resolving the Ambiguity of Article 22 (1) of the General Data Protection Regulation." *International Data Privacy Law* 11, no. 2 (2021): 145-62. https://doi.org/10.1093/idpl/ipaa024.

Tramer, Florian, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. "Fairtest: Discovering Unwarranted Associations in Data-Driven Applications." Paper presented at the 2017 IEEE European Symposium on Security and Privacy (EuroS&P), 2017.

Tran, Khoa A, Olga Kondrashova, Andrew Bradley, Elizabeth D Williams, John V Pearson, and Nicola Waddell. "Deep Learning in Cancer Diagnosis, Prognosis and Treatment Selection." *Genome Medicine* 13, no. 152 (2021): 1-17. https://doi.org/10.1186/s13073-021-00968-x.

Trierweiler, Steven J, Harold W Neighbors, Cheryl Munday, Estina E Thompson, Victoria J Binion, and John P Gomez. "Clinician Attributions Associated with the Diagnosis of Schizophrenia in African American and Non-African American Patients." *Journal of consulting and clinical psychology* 68, no. 1 (2000): 171. https://doi.org/10.1037/0022-006X.68.1.171.

Udeshi, Sakshi, Pryanshu Arora, and Sudipta Chattopadhyay. "Automated Directed Fairness Testing." Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering, Montpellier, France, Association for Computing Machinery, 2018.

Urbina, Francisco J. "A Critique of Proportionality." *American Journal of Jurisprudence* 57 (2012): 49-80.

Vaccarino, Viola, Harlan M Krumholz, and Jorge Yarzebski. "Sex Differences in 2-Year Mortality after Hospital Discharge for Myocardial Infarction." *Annals of Internal Medicine* 134, no. 3 (2001): 173-81. https://doi.org/10.7326/0003-4819-134-3-200102060-00007.

Vale, Daniel, Ali El-Sharif, and Muhammed Ali. "Explainable Artificial Intelligence (Xai) Post-Hoc Explainability Methods: Risks and Limitations in Non-Discrimination Law." *AI and Ethics*, no. 2 (2022): 815-26. https://doi.org/10.1007/s43681-022-00142-y.

van Bekkum, Marvin, and Frederik Zuiderveen Borgesius. "Using Sensitive Data to Prevent Discrimination by Artificial Intelligence: Does the Gdpr Need a New Exception?". *Computer Law & Security Review* 48, no. 105770 (2023): 1-12. https://doi.org/10.1016/j.clsr.2022.105770.

van Dijk, Niels, Raphaël Gellert, and Kjetil Rommetveit. "A Risk to a Right? Beyond Data Protection Risk Assessments." *Computer Law & Security Review* 32, no. 2 (2016): 286-306. https://doi.org/10.1016/j.clsr.2015.12.017.

van Kolfschooten, Hannah. "Eu Regulation of Artificial Intelligence: Challenges for Patients' Rights." *Common Market Law Review* 59, no. 1 (2022): 81-112.

Veale, Michael, and Reuben Binns. "Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data." *Big Data & Society* 4, no. 2 (2017): 2053951717743530.

Veale, Michael, and Lilian Edwards. "Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling." *Computer Law & Security Review* 34, no. 2 (2018): 398-404. https://doi.org/10.1016/j.clsr.2017.12.002.

Vokinger, Kerstin N., Stefan Feuerriegel, and Aaron S. Kesselheim. "Mitigating Bias in Machine Learning for Medicine." *Communications Medicine* 1, no. 25 (2021): 1-3. https://doi.org/10.1038/s43856-021-00028-w.

Vokinger, Kerstin N., and Urs Gasser. "Regulating AI in Medicine in the United States and Europe." *Nature Machine Intelligence* 3, no. 9 (2021): 738-39. https://doi.org/10.1038/s42256-021-00386-z.

Vranas, Kelly C., Jeffrey K. Jopling, Timothy E. Sweeney, Meghan C. Ramsey, Arnold S. Milstein, Christopher G. Slatore, Gabriel J. Escobar, and Vincent X. Liu. "Identifying Distinct Subgroups of Icu Patients: A Machine Learning Approach*." *Critical Care Medicine* 45, no. 10 (2017): 1607-15. https://doi.org/10.1097/ccm.0000000000002548.

Vyas, Darshali A., Leo G. Eisenstein, and David S. Jones. "Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms." *New England Journal of Medicine* 383, no. 9 (2020): 874-82. https://doi.org/10.1056/NEJMms2004740.

Wachter, Sandra. "Affinity Profiling and Discrimination by Association in Online Behavioural Advertising." *Berkeley Technology Law Journal* 35, no. 2 (2020): 367-430. https://doi.org/10.15779/Z38JS9H82M.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the Gdpr." *Harvard Journal of Law & Technology* 31, no. 2 (Spring 2018): 841-88.

———. "Bias Preservation in Machine Learning: The Legality of Fairness Metrics under Eu Non-Discrimination Law." *West Virginia Law Review* 123, no. 3 (Spring 2021): 735-90. (**Wachter, Mittelstadt, and Russell (2021 A))**

———. "Why Fairness Cannot Be Automated: Bridging the Gap between Eu Non-Discrimination Law and AI." *Computer Law & Security Review* 41 (2021): 1-31. https://doi.org/10.1016/j.clsr.2021.105567. **(Wachter, Mittelstadt, and Russell (2021 B)).**

Waldum, Åsa Henning, Anne Flem Jacobsen, Mirjam Lukasse, Anne Cathrine Staff, Ragnhild Sørum Falk, Siri Vangen, and Ingvil Krarup Sørbye. "The Provision of Epidural Analgesia During Labor According to Maternal Birthplace: A Norwegian Register Study." *BMC Pregnancy and Childbirth* 20, no. 321 (2020): 1-10. https://doi.org/10.1186/s12884-020-03021-8.

Wallis, Christopher JD, Angela Jerath, Natalie Coburn, Zachary Klaassen, Amy N Luckenbaugh, Diana E Magee, Amanda E Hird, *et al.* "Association of Surgeon-Patient Sex Concordance with Postoperative Outcomes." *JAMA surgery* 157, no. 2 (2022): 146-56. https://doi.org/10.1001/jamasurg.2021.6339.

Ward, Angela. "The Impact of the EU Charter of Fundamental Rights on Anti-Discrimination Law: More a Whimper Than a Bang?". *Cambridge Yearbook of European Legal Studies* 20 (2018): 32-60. https://doi.org/10.1017/cel.2018.11.

Weberpals, Janick, Tim Becker, Jessica Davies, Fabian Schmich, Dominik Rüttinger, Fabian J. Theis, and Anna Bauer-Mehren. "Deep Learning-Based Propensity Scores for Confounding Control in Comparative Effectiveness Research: A Large-Scale, Real-World Data Study." *Epidemiology* 32, no. 3 (2021): 378-88. https://doi.org/10.1097/ede.0000000000001338.

Weissman, Myrna M., and Gerald L. Klerman. "Sex Differences and the Epidemiology of Depression." *Archives of General Psychiatry* 34, no. 1 (1977): 98-111. https://doi.org/10.1001/archpsyc.1977.01770130100011.

Wiens, Jenna, W Nicholson Price, and Michael W Sjoding. "Diagnosing Bias in Data-Driven Algorithms for Healthcare." *Nature Medicine* 26, no. 1 (2020): 25-26. https://doi.org/10.1038/s41591-019-0726-6.

Williams, Betsy Anne, Catherine F. Brooks, and Yotam Shmargad. "How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications." *Journal of Information Policy* 8 (2018): 78-115. https://doi.org/10.5325/jinfopoli.8.2018.0078.

Wójcik, Malwina Anna. "Algorithmic Discrimination in Health Care: An EU Law Perspective." *Health and human rights* 24, no. 1 (June 2022): 93-103.

Wynants, Laure, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc M J Bonten, *et al.* "Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal." *BMJ* 369 (2020): m1328. https://doi.org/10.1136/bmj.m1328.

Waaldijk, Kees, and Christa Tobler. "Case C–267/06, Tadao Maruko V. Versorgungsanstalt Der Deutschen Bühnen, Judgment of the Grand Chamber of the Court of Justice of 1 April 2008." *Common Market Law Review* 46, no. 2 (2009): 723-46.

Xenidis, Raphaële. "Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience." *Maastricht Journal of European and Comparative Law* 27, no. 6 (2020): 736-58. https://doi.org/10.1177/1023263X20982173.

Xie, Junyuan, Ross Girshick, and Ali Farhadi. "Unsupervised Deep Embedding for Clustering Analysis." *Proceedings of the 33rd International Conference on Machine Learning* 48 (2016): 478-87.

Yala, Adam, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. "A Deep Learning Mammography-Based Model for Improved Breast Cancer Risk Prediction." *Radiology* 292, no. 1 (2019): 60-66. https://doi.org/10.1148/radiol.2019182716.

Yeung, Karen, and Lee A. Bygrave. "Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship." *Regulation & Governance* 16 (2022): 137-55. https://doi.org/10.1111/rego.12401.

Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. "Recent Trends in Deep Learning Based Natural Language Processing." *ieee Computational intelligenCe magazine* 13, no. 3 (2018): 55-75. https://doi.org/10.1109/MCI.2018.2840738.

Zarlenga, Mateo Espinosa, Zohreh Shams, Michael Edward Nelson, Been Kim, and Mateja Jamnik. "Tabcbm: Concept-Based Interpretable Neural Networks for Tabular Data." *Transactions on Machine Learning Research* 7 (2023): 1-35.

Zarsky, Tal. "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making." *Science, Technology, & Human Values* 41, no. 1 (2015): 118-32. https://doi.org/10.1177/0162243915605575.

Zarsky, Tal Z. "Correlation Versus Causation in Health‐Related Big Data Analysis." *Big Data, Health Law, and Bioethics* (2018): 42-55.

Zehlike, Meike, Alex Loosley, Philipp Hacker, Håkan Jonsson, and Emil Wiedemann. "Beyond Incompatibility: Interpolation between Mutually Exclusive Fairness Criteria in Classification Problems." *arXiv preprint arXiv:2212.00469* (2022).

Zou, James, and Londa Schiebinger. "Ensuring That Biomedical Ai Benefits Diverse Populations." *EBioMedicine* 67, no. 103358 (2021): 1-6. https://doi.org/10.1016/j.ebiom.2021.103358.

Zulqarnain, Fatima, S Fisher Rhoads, and Sana Syed. "Machine and Deep Learning in Inflammatory Bowel Disease." *Current Opinion in Gastroenterology* 39, no. 4 (July 2023): 294-300. https://doi.org/https://doi.org/10.1097%2FMOG.0000000000000945.

Özdemir, Berna C, and Anna Dorothea Wagner. "Consideration of Sex and Gender Aspects in Oncology: Rationale, Current Status, and Perspectives." *Italian Journal of Gender-Specific Medicine* 8, no. 1 (2022): 55-58. https://doi.org/10.1723/3769.37567.

Theses:

Ikdahl, Ingunn. "Securing Women's Homes: The Dynamics of Women's Human Rights at the International Level and in Tanzania."PhD thesis, University of Oslo, 2010.

Ofstad, Eirik Hugaas. "Medical Decisions in 372 Hospital Encounters." University of Oslo, 2015.

Storvik, Marius. "Rettslig Vern Av Pasienters Integritet I Psykisk Helsevern." PhD, UiT Norges arktiske universitet, 2017.

Strand, Vibeke Blaker. "Diskrimineringsvernets Rekkevidde I Møte Med Religionsutøvelse."PhD thesis, Universitetet i Oslo, 2011.

Von Grafenstein, Maximilian. "The Principle of Purpose Limitation in Data Protection Laws."PhD thesis, Hamburg University.

Reports, guidelines, policy documents, and law proposals:

Administration, U.S. Food & Drug. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SAMD) - Discussion Paper and Request for Feedback.* (2 April 2019). https://www.fda.gov/media/122535/download?attachment.

Article 29 Data Protection Working Party. *Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks.* (30 May 2014). https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf.

———.*Guidelines on Data Protection Impact Assessment (DPIA).* (13 October 2017). https://ec.europa.eu/newsroom/article29/items/611236.

———.*Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679* (Wp251rev.01). (6 February 2018). https://ec.europa.eu/newsroom/article29/items/612053.

Baeten, Rita, Slavina Spasova, Bart Vanhercke, and Stéphanie Coster. *Inequalities in Access to Healthcare - a Study of National Policies.* European Commission (Brussels: November 2018). https://ec.europa.eu/social/main.jsp?catId=738&langId=en&pubId=8152&furtherPubs=yes.

Bell, Mark, and Sara Kjellstrand. *Critical Review of Academic Literature Relating to the EU Directives to Combat Discrimination.* European Commission Directorate-General for Employment and Social Affairs (2004). https://www.antigone.gr/wp-

content/uploads/library/documentation-of-EU-and-international-organizations/policy-documents/en/criticrevaclit.pdf.
Brown, Lydia X.Z., Michelle Richardson, Ridhi Shetty, Andrew Crawford, and Timothy Hoagland. *Challenging the Use of Algorithm-Driven Decision-Making in Benefits Determinations for People with Disabilities.* Center for Democracy and Technology (October 2020). https://cdt.org/insights/-challenging-the-use-of-algorithm-driven-decision-making-in-benefits-determinations-affecting-people-with-disabilities/.
Centre for Information Policy Leadership (CIPL). *Risk, High Risk, Risk Assessments and Data Protection Impact Assessments under the GDPR.* (21 December 2016). https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_gdpr_project_risk_white_paper_21_december_2016.pdf.
Commission on Social Determinants of Health. *Closing the Gap in a Generation: Health Equity through Action on the Social Determinants of Health* (Final /Executive Summary). World Health Organization (Geneva, Switzerland: WHO Press).
Committee, House of Commons Science and Technology. *Algorithms in Decision-Making.* (23 May 2018). https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/35104.htm.
Committee of Ministers. *Resolution (73) 22 on the Protection of the Privacy of Individuals Vis-a-Vis Electronic Data Banks in the Private Sector.* Council of Europe (26 September 1973). https://rm.coe.int/1680502830.
———. *Recommendation of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems.* Council of Europe (8 April 2020). https://rm.coe.int/09000016809e1154.
Crowley, Niall. *Equality Bodies Making a Difference.* European network of legal experts in gender equality and non-discrimination (2018).

———. *Equality, Diversity and Non-Discrimination in Healthcare: Learning from the Work of Equality Bodies.* European Network of Equality Bodies (EQUINET) (2021). https://equineteurope.org/publications/equality-diversity-and-non-discrimination-in-healthcare-learning-from-the-work-of-equality-bodies/.

Datatilsynet. *Ahus, Sluttrapport: Hjerterom for Etisk Ai.* (Februar 2023). https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/ahus-sluttrapport-ekg-ai/.

Edwards, Lilian. *Regulating AI in Europe: Four Problems and Four Solutions.* Ada Lovelace Institute (March 2022). https://www.adalovelaceinstitute.org//regulating-ai-in-europe/.
European Agency for Fundamental Rights (FRA). *Getting the Future Right - Artificial Intelligence and Fundamental Rights.* (Luxembourg: 2020). https://fra.europa.eu/sites/default/files/fra_uploads/fra-2020-artificial-intelligence_en.pdf.
European Commission. *Proposal for a Regulation of the European Parliament and of the Council on Medical Devices, and Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009.* (2012).
———. *White Paper on Artificial Intelligence - a European Approach to Excellence and Trust* (Brussels: 19 February 2020). https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
———. *Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* (Brussels: 21 April 2021).
———. *Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive)* (2022).
———. *Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products* (2022).
European Coordination Committee of the Radiological, Electromedical and Healthcare IT Industry (COCIR). *COCIR Analysis on AI in Medical Device Legislation.* (4 September 2020).

https://www.cocir.org/fileadmin/Position_Papers_2020/COCIR_Analysis_on_AI_in_medical_Device_Legislation_-_Sept._2020_-_Final_2.pdf.
European Data Protection Board, and European Data Protection Supervisor. *Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (2021).* https://edpb.europa.eu/our-work-tools/ourdocuments/edpbedps-joint-opinion.
European Data Protection Board (EDPB). *Guidelines 4/2019 on Article 25 Data Protection by Design and by Default.* (20 October 2019). https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf.
European Parliament. *European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies*, 2020.
European Union Agency for Fundamental Rights. *Bias in Algorithms: Artificial Intelligence and Discrimination.* (Vienna: 2022 2022). https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf
European Union Agency for Fundamental Rights. *Inequalities and Multiple Discrimination in Access to and Quality of Healthcare.* (Luxembourg: Publications Office of the European Union, 2013).
European Parliamentary Research Service: Study Panel for the Future of Science and Technology. *Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts* (June 2022).
Equalities Review Panel. *Fairness and Freedom: The Final Report of the Equalities Review.* (Crown, 2007).
Executive Office of the President, and John Podesta. *Big Data: Seizing Opportunities, Preserving Values.* United States: White House (2014). https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy__may_1_2014.pdf.
Ferryman, Kadija, and Mikaela Pitcan. *Fairness in Precision Medicine.* Data & Society (February 2018). https://datasociety.net/wp-content/uploads/2018/02/DataSociety_Fairness_In_Precision_Medicine_Feb2018.pdf.
Frykman, Jonas, Weini Kahsai Nobel, Johanna Ahnquist, Anna Schölin, and Marie Byskov Lindberg. *Discrimination - a Threat to Public Health Final - Health and Discrimination Project.* (2007).
Gerards, Janneke, and Raphaële Xenidis. *Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law.* European Commission Directorate-General for Justice and Consumers (Brussels: Publications Office, 2021). https://op.europa.eu/en/publication-detail/-/publication/082f1dbc-821d-11eb-9ac9-01aa75ed71a1.
Helsedirektoratet. *Forprosjekt. Utredning Om Bruk Av Kunstig Intelligens I Helsesektoren.* (December 2019). https://www.ehelse.no/publikasjoner/utredning-om-bruk-av-kunstig-intelligens-i-helsesektoren.
High-Level Expert Group on Artificial Intelligence. *A Definition of AI: Main Capabilities and Disciplines.* (European Commission, 8 April 2019).
Information Commissioner's Office. *Big Data, Artificial Intelligence, Machine Learning and Data Protection.* (2017).
Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* (Washington DC: The National Academic Press, 2003). http://nap.edu/12875.
Likestillings- og diskrimineringsombudet. *Veileder for Innebygd Diskrimineringsvern.* (2022). https://ldo.no/globalassets/_ldo_2019/_bilder-til-nye-nettsider/ki/ldo.-innebygd-diskrimineringsvern.pdf.
Liu, Kimberly, and Colm O'Cinneide. *The Ongoing Evolution of the Case-Law of the Court of Justice of the European Union on Directives 2000/43/Ec and 2000/78/EC.* European Commission

Directorate-General for Justice and Consumers (Luxembourg: Publications Office of the European Union, November 2019).
Makkonen, Timo. *Measuring Discrimination: Data Collection and Eu Equality Law.* European Commission Directorate-General for Employment, Social Affairs and Equal Opportunities (Luxembourg: Publications Office of the European Union, 2007).
Mantelero, Alessandro. *on Artificial Intelligence (Convention 108).* Council of Europe Consultative Committee of the Convention for the Protection of Individuals with Regard to Auotmatic Processing of Personal Data (2019). https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6.
Medical Device Coordination Group. *MDCG 2019-11 Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 - MDR and Regulation (EU) 2017/746 - IVDR* (October 2019).
Mittelstadt, Brent. *The Impact of Artificial Intelligence on the Doctor-Patient Relationship.* Council of Europe (December 2021). https://rm.coe.int/inf-2022-5--impact-of-ai-on-doctor-patient-relations-e/1680a68859.
Tobler, Christa. *Limits and Potential of the Concept of Indirect Discrimination.* European Commission Directorate-General for Employment, Social Affairs and Equal Opportunities (Luxembourg: Office for Official Publications of the European Communities, 2008).
———. *Indirect Discrimination under Directives 2000/43 and 2000/78.* European Commission Directorate-General for Justice and Consumers (Luxembourg: Publications Office of the European Union, 2022).
UN Committee on Economic, Social and Cultural Rights. *Cescr General Comment No. 14: The Right to the Highest Attainable Standard of Health (Art. 12).* (11 August 2000). https://www.refworld.org/pdfid/4538838d0.pdf.
UN Committee on Economic, Social and Cultural Rights (CESCR). *General Comment No. 20: Non-Discrimination in Economic, Social and Cultural Rights.* (2009). https://www.globalhealthrights.org/instrument/cescr-general-comment-no-20-non-discrimination-in-economic-social-and-cultural-rights/.
UN Committee on the Elimination of Discrimination Against Women (CEDAW). *General Recommendation No. 25, on Article 4, Paragraph 1, of the Convention on the Elimination of All Forms of Discrimination against Women, on Temporary Special Measures.* (2004). https://www.un.org/womenwatch/daw/cedaw/recommendations/General%20recommendation%2025%20(English).pdf.
van Der Sloot, Bart, Esther Keymolen, Merel Noorman, The Netherlands Institute for Human Rights, Hilde Weerts, Yvette Wagensveld, and Bram Visser. *Non-Discrimination by Design.* (2023). https://www.tilburguniversity.edu/about/schools/law/departments/tilt/research/handbook.
World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance.* (Geneva: 2021). https://apps.who.int/iris/bitstream/handle/10665/341996/9789240029200-eng.pdf.
Yeung, Karen. *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework.* Council of Europe (2019).
Zuiderveen Borgesius, Frederik. *Council of Europe Study on Discrimination, Artificial Intelligence, and Algorithmic Decision-Making.* Council of Europe: Anti-Discrimination Department (2018). https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.

Blog posts:

Brownlee, Jason, "Difference between Classification and Regression in Machine Learning," *Machine Learning Mastery*, 22 May, 2019, https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/.
Kozyrkov, Cassie, "What Is Bias?," *Towards Data Science*, 24 January, 2019, https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814.
Lindvall, Charlotta, Christine K. Cassel, Steven Z. Pantilat, and Matthew DeCamp, "Ethical Considerations in the Use of Ai Mortality Predictions in the Care of People with Serious Illness," *Health Affairs Blog*. *Health Affairs*, 2020, https://www.healthaffairs.org/do/10.1377/hblog20200911.401376/full/?MessageRunDetailID=3353581596&PostID=19618763&af=R&content=blog&mi=3egtxy&sortBy=Earliest&target=do-blog.
Hali, Brenda, "Understanding Bias-Variance Trade-Off in 3 Minutes," *Towards Data Science*, 2 December, 2019, https://towardsdatascience.com/understanding-bias-variance-trade-off-in-3-minutes-c516cb013513.
Hardt, Moritz, "How Big Data Is Unfair," *Medium*. *Medium*, 26 September, 2014, https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.
Kenton, Will, "Internal Controls: Definition, Types, and Importance," Julius Mansa and Suzanne Kvilhaug eds. *Investopedia*, 24 May, 2023, https://www.investopedia.com/terms/i/internalcontrols.asp.
———, "What Is Behavioral Economics? Theories, Goals, and Applications," Toby Walters and Marcus Reeves eds. *Investopedia*, 16 January, 2023, https://www.investopedia.com/terms/b/behavioraleconomics.asp.
Hidvegi, Fanny, Daniel Leufer, and Estelle Massé, "The Eu Should Regulate Ai on the Basis of Rights, Not Risks," *Access Now*, 17 February, 2021, https://www.accessnow.org/eu-regulation-ai-risk-based-approach/.
Mcleod, Saul, "Social Identity Theory in Psychology (Tajfel & Turner, 1979)," Olivia Guy-Evans ed. *Simply Psychology*, 2 October, 2023, https://www.simplypsychology.org/social-identity-theory.html.
Suleyman, Mustafa, and Dominic King, "Using Ai to Give Doctors a 48-Hour Head Start on Life-Threatening Illness," *deepmind.com*. *DeepMind*, 31 July, 2019, https://deepmind.com/blog/article/predicting-patient-deterioration.

Web pages:

"Treatment, Lung Cancer." 2019, accessed 29 July, 2021, https://www.nhs.uk/conditions/lung-cancer/treatment/.
"Computer Science." 2022, accessed 28 August, 2022, https://www.britannica.com/science/computer-science.
"Gender and Health." accessed 7 November, 2023, https://www.who.int/health-topics/gender#tab=tab_1.
"The Complete Guide on Overfitting and Underfitting in Machine Learning." Simplilearn Solutions, Updated 20 February 2023, 2023, accessed 7 November, 2023, https://www.simplilearn.com/tutorials/machine-learning-tutorial/overfitting-and-underfitting.
"Conformity Assessment." accessed 12 November, 2023, https://www.iso.org/conformity-assessment.html.
"Nasjonalt Kvalitetsregister for Ryggkirurgi." University Hospital of North Norway, Updated 8 September 2023, accessed 10 November, 2023, https://www.unn.no/fag-og-forskning/medisinske-kvalitetsregistre/nasjonalt-kvalitetsregister-for-ryggkirurgi.
"New Legislative Framework." European Commission, accessed 11 November, 2023, https://single-market-economy.ec.europa.eu/single-market/goods/new-legislative-framework_en.

Videos and podcasts:

Bowles, Cennydd. What Is AI? 2. Podcast audio. Machine Ethics podcast 24:30.
Crawford, Kate. "The Trouble with Bias - Nips 2017 Keynote." 2017. https://www.youtube.com/watch?v=fMym_BKWQzk.
Video Journal of Biomedicine. "Propensity Score Matching Methodology: Why and How It Is Used." 00:03:02 Video Journal of Biomedicine, 13 December 2022. Youtube. https://www.youtube.com/watch?v=40U_0DtNrQc.

Other references:

American Academy of Orthopaedic Surgeons. "Oswestry Low Back Disability Questionnaire." https://www.aaos.org/globalassets/quality-and-practice-resources/patient-reported-outcome-measures/spine/oswestry-2.pdf.
Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." (23 May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
Engler, Alex. "A Guide to Healthy Skepticism of Artificial Intelligence and Coronavirus."2023. (2 April 2020). https://www.brookings.edu/articles/a-guide-to-healthy-skepticism-of-artificial-intelligence-and-coronavirus/.
Flugstad Eriksen, Kjersti. "Skal Forske På Hvorfor Innvandrere Fra Sør-Asia Oftere Får Diabetes." *Dagens medisin*, 2023. https://www.dagensmedisin.no/diabetes-oslo-universitetssykehus-ous-overvekt/skal-forske-pa-hvorfor-innvandrere-fra-sor-asia-oftere-far-diabetes/565230.
Knight, Will. "AI Can Help Diagnose Some Illnesses - If Your Country Is Rich." *Wired*, 18 November, 2020. https://www.wired.com/story/ai-diagnose-illnesses-country-rich/?
O'Cinneide, Colm. "Positive Action and Eu Law." (2011): 1-25. Accessed November. https://www.era-comm.eu/oldoku/SNLLaw/04_Positive_action/2011-111DV20-O'Cinneide_EN.pdf.