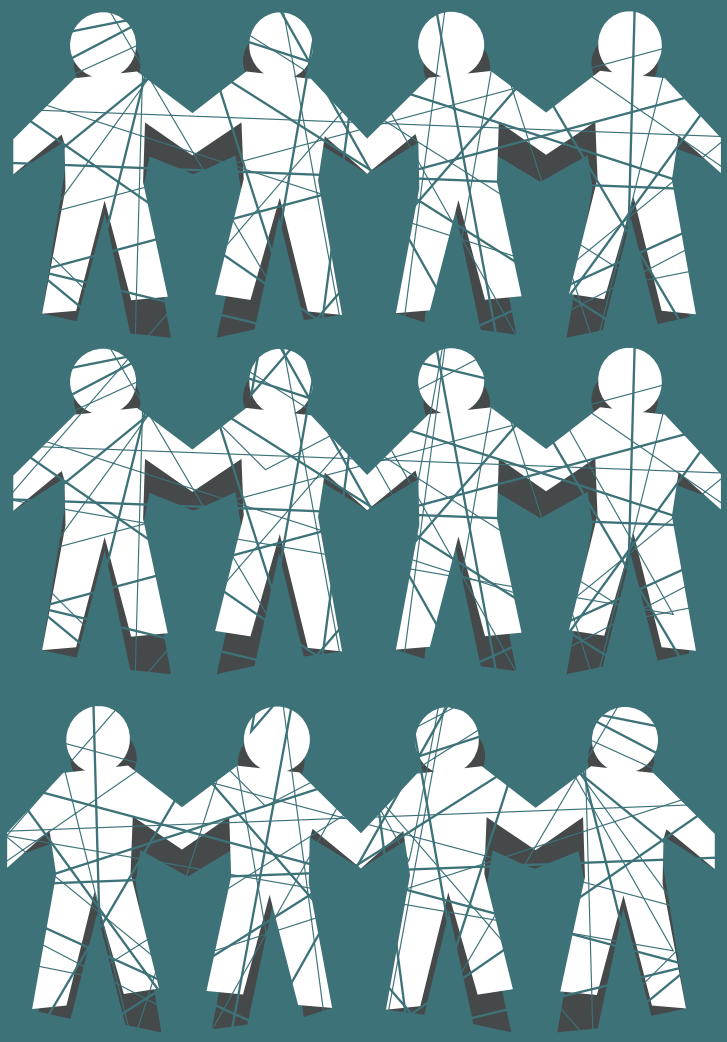
Where Law meets Artificial Intelligence:

Discrimination and Unfair Differentiation in Insurance



Marvin van Bekkum

Institute for Computing and Information Sciences

RADBOUD UNIVERSITY PRESS

Radboud Dissertation Series

Where Law meets Artificial Intelligence: Discrimination and Unfair Differentiation in Insurance

Where Law meets Artificial Intelligence:
Discrimination and Unfair Differentiation in Insurance
Marvin van Bekkum

Radboud Dissertation Series
ISSN: 2950-2772 (Online); 2950-2780 (Print)

Published by RADBOUD UNIVERSITY PRESS
Postbus 9100, 6500 HA Nijmegen, The Netherlands
www.radbouduniversitypress.nl

Design: Marvin van Bekkum
Cover: Proefschrift AIO | Guus Gijben
Printing: DPN Rikken/Pumbo

ISBN: 9789465152653
DOI: 10.54195/9789465152653
Free download at: https://doi.org/10.54195/9789465152653

RADBOUD
UNIVERSITY
PRESS

Where Law meets Artificial Intelligence:
Discrimination and Unfair Differentiation in Insurance

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.M. Sanders,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 11 mei 2026
om 14.30 uur precies

door

Marvin Silvester Leopold van Bekkum

geboren op 28 april 1994

te Kerkrade

Promotoren:
Prof. mr. F.J. Zuiderveen Borgesius
Prof. dr. T.M. Heskes

Manuscriptcommissie:
Prof. dr. mr. W.H. van Boom
Prof. dr. J.G.J. Rinkes (Open Universiteit)
Prof. dr. mr. J.V.J. van Hoboken (Universiteit van Amsterdam)
Prof. dr. A.B. Terlouw
Prof. mr. dr. M. Hildebrandt

Where Law meets Artificial Intelligence:
Discrimination and Unfair Differentiation in Insurance

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.M. Sanders,
according to the decision of the Doctorate Board
to be defended in public on

Monday, May 11, 2026
at 2.30 p.m.

by

Marvin Silvester Leopold van Bekkum

born on April 28, 1994

in Kerkrade, the Netherlands

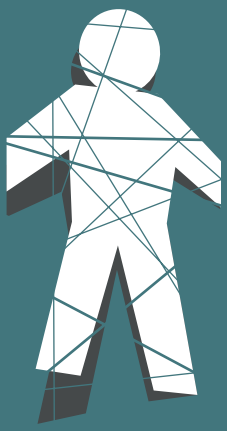
Supervisors:
Prof. mr. F.J. Zuiderveen Borgesius
Prof. dr. T.M. Heskes

Manuscript Committee:
Prof. dr. mr. W.H. van Boom
Prof. dr. J.G.J. Rinkes (Open Universiteit)
Prof. dr. mr. J.V.J. van Hoboken (University of Amsterdam)
Prof. dr. A.B. Terlouw
Prof. mr. dr. M. Hildebrandt

Table of Contents

List of Tables

List of Figures



I.

Introduction

Apparently, house numbers influence car insurance claims. In 2015, the Dutch consumer organisation *Consumentenbond* reported that consumers with different postal codes, or even different house numbers, or even different house number extensions, paid different prices for their car insurance. In some cases, for example, consumers living at house number 11A would pay a different price than consumers at house number 11B.[1] In 2022, the Consumentenbond still found differences within the same street.[2] A consumer with the same age, gender and car as their next-door neighbour paid up to €200 per year more for car insurance.[3]

What sets insurance apart from other sectors is underwriting: insurers 'give an activity financial support and take responsibility for paying any costs if it fails.'[4] While underwriting risks, insurers set a price that depends on a risk assessment, estimating the chance that a consumer may claim the costs of an accident or other damages. For example, car insurers may charge a higher price to a senior if consumers who had car accidents in the past were often

1 Consumentenbond, 'Premies verzekeringen verschillen tot op huisnummer ['Insurance premiums differ by house number']', https://www.consumentenbond.nl/inboedelverzekering/verzekeringspremies-verschillen-tot-op-huisnummer, 2015.

2 Consumentenbond, 'Zelfde auto, andere verzekeringspremie', https://www.consumentenbond.nl/nieuws/2022/zelfde-auto-andere-verzekeringspremie, March 2022.

3 The television programme Radar looked into the issue independently, coming with similar conclusions as the Consumentenbond. Avrotros Radar, 'Dezelfde autoverzekering als je buurman, toch honderden euro's meer kwijt', https://radar.avrotros.nl/artikel/dezelfde-autoverzekering-als-je-buurman-toch-honderden-euros-meer-kwijt-51290, 27 September 2021.

4 https://dictionary.cambridge.org/dictionary/english/underwrite.

seniors. Insurers establish such relationships by analysing vast datasets about the accident history of consumers.[5]

When consumers apply for insurance, insurers ask for specific information, such as the price of the consumer's car.[6] Insurers have been using statistical systems for a long time.[7] Over the past decade, insurers considered gathering data from different sources, such as Internet of Things devices (for example, wearables).[8] With the new data, insurers could, for example, better predict who is a low-risk or a high-risk consumer. Most European insurers seem hesitant to use machine learning, a specific type of AI system, for underwriting. Some European insurers embrace machine learning, using it for completely automatic risk underwriting or for augmenting the insurer's risk assessments.[9]

According to insurers, consumers may enjoy benefits if insurers use AI systems for underwriting: (i) Consumers with a safe lifestyle could save some money. (ii) Insurers could allow certain high-risk consumers to access insurance, such as insurance for people with high-risk sports. (iii) Insurers could 'improve customer satisfaction, [because] customers [no longer need to] fill out repetitive questionnaires.' (iv) Finally, insurers can insure risks that they previously could not if they use AI to find preventive measures consumers could take to prevent accidents.[10]

5 See Chapter 1 of the thesis.

6 Gandy 2016, p. 69.

7 Actuaries have been using statistics for insurance pricing since the mid-18th century. For a history since ancient Egypt, see Gupta, *The Actuary Magazine* 2024.

8 EIOPA 2019/4.2. The Geneva Association (Noordhoek) 2023, p. 26. Scholars have warned for insurance differentiation through smart devices. See e.g. Hildebrandt 2015.

9 A survey conducted by the European Insurance and Occupational Pensions Authority in June 2023 shows that 'to date insurance undertakings have predominantly opted for more simple and explainable AI algorithms such as advanced regressions or tree-based algorithms'. EIOPA 2024, p. 26. The report also states that 'it is [...] noteworthy that some insurers are already using AI to perform risks assessments and manage underwriting risks [...]'. EIOPA 2024, p. 27 & Fig. 19. See also Swiss Re (Ladva & Grasso) 2024. Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019, p. 14. For a review of scientific papers discussing AI applications in automotive, health, and property insurance, see e.g. Bhattacharya and others, *Frontiers in Artificial Intelligence* 2025/8.

10 Insurance Europe 2019.

1 Problem statement and central research question

Scholars warn that certain AI systems may pose a risk to fairness and non-discrimination in insurance.[11] Machine learning systems can find thousands of correlations in datasets containing many features such as weight and age. Due to the complex algorithms and datasets involved, insurers may find it difficult to understand the statistical relationships behind a machine learning system's predictions and decisions.[12]

Some insurers express concern about 'a shift from causation to correlation', where understanding relationships in data becomes less important for the insurer than finding correlations.[13] If insurers were to rely too much on machine learning, they might no longer question their assumptions about a dataset or scrutinize the output of machine learning systems.

While insurers might be concerned about the effect AI systems could have on their processes, they could perhaps trust more in the AI system if they could assess that the system does not have discrimination-related risks.[14] Such a 'non-discrimination-audit' generally requires two steps: (i) an insurer can measure possible discrimination using metrics and second, (i) the insurer can weigh whether discrimination has taken place.[15]

(i) Measuring discrimination using metrics involves calculating the extent to which vulnerable groups and individuals are represented in the AI system's predictions or decisions.[16] Insurers cannot always perform such a test, facing a legal hurdle: if an insurer wishes to test whether an AI system

11 Swedloff, *61 B.C. L. REV.* 2020/2031, p. 2083. European Insurance and Occupational Pensions Authority (EIOPA) 2021/VI.

12 The Geneva Association (Noordhoek) 2023, p. 24.

13 The Geneva Association (Noordhoek) 2023, p. 17, 21, 24. For a book criticizing the objectivity of data and AI, see Chapter 1 of Hong 2020.

14 The Geneva Association marks this approach as 'A specific approach highlighted by several insurers is to examine, ex-post, whether the outcomes of AI models are discriminatory.' See The Geneva Association (Noordhoek) 2023, p. 24.

15 For a paper that makes this distinction, see Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023.

16 Outside of the AI context, insurers can also measure whether certain groups are correctly represented in predictions and decisions. For example, in the US, by comparing public data for specific groups, the US Urban Institute found that 'Because of intentional government policies and industry practices, racial and ethnic disparities in access to

discriminates against, for example, a certain ethnicity, the insurer needs to know the ethnicities of consumers.[17] Yet in the European Union, the General Data Protection Regulation (GDPR), in principle, prohibits insurers from gathering information about ethnicities.[18] An exception in the EU AI Act could, potentially, allow insurers to gather the necessary ethnicity data.[19]

(ii) Next, the insurer must assess whether discrimination might take place. This requires weighing all the circumstances of the case. For such a weighing, insurers need more than their own perspective. For example, an insurer must engage in dialogue with other stakeholders or hold surveys to gather a different perspective. Fairness is 'not something owned by the market, decided by the market, shaped by the market. It is owned by everyone with a stake in what insurance does, such as consumers, society, regulators, trade bodies, professional bodies, politicians and, of course, market firms.'[20] Fairness and non-discrimination are societal values, which cannot be expressed solely in terms of numbers.[21]

As the previous paragraphs show, in the thesis, *law meets artificial intelligence*.[22] In writing the first thesis about the discrimination-related risks of AI in insurance, I aim to explore whether insurers could introduce machine learning into underwriting while mitigating possible discrimination-related

credit persist today'. Urban Institute 2024, p. 20. This is an example from outside of the AI context, that could also apply to the AI context.

17 For an overview of hurdles insurers could have in introducing AI systems into underwriting, see Charpentier & Vamparys, *Big Data & Society* 2025/12.

18 Art. 9(1) GDPR.

19 Article 10(5) AI Act.

20 Minty 2023, p. 11. Social scholars include similar arguments: ''Insights into citizens' fairness perceptions are essential to facilitate human-centric AI by informing developers entrusted with designing ethical ADM and decision-makers tasked with implementing such systems in social contexts.' Kieslich, Keller & Starke, *Big Data & Society* 2022/9. Fairness perceptions can contribute to answering the call for a society-in-the-loop approach that emphasizes embedding societal values in the design of ADM systems. Rahwan, *Ethics and Information Technology* 2018/20.

21 Some scholars in the AI Fairness community find that 'expressing fairness in a single metric seems overzealous'. De Jonge & Hiemstra, *ACM Conference on Fairness, Accountability, and Transparency* 2023, p. 837. See also Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023.

22 See further Section 5 (Method).

risks, investigating the central research question:

> To what extent do insurers introduce discrimination-related risks by using AI systems for underwriting?

A few notes about the scope of the thesis. Where I write 'insurer', I mean any type of company offering insurance. This could include a tech startup just entering the insurance business, but also a centuries-old insurance company. 'Discrimination-related' refers to (i) forms of discrimination that are prohibited under EU law and (ii) other unfair forms of differentiation.[23] Since I focus on discrimination and other unfair differentiation, many questions are outside the scope of the thesis. For example, I do not discuss the right to healthcare (relevant for health insurance).

Where I write about 'AI systems', I do *not* focus on Generative AI models such as ChatGPT.[24] In particular, where the technical type of AI system is relevant, the thesis focuses mostly on machine learning.[25]

I will answer the central research question by investigating the gaps (subtopics) following the general line I described in Section 1: First, I discuss how insurers currently use AI systems for underwriting and identify possible discrimination-related threats. Second, I will investigate whether the GDPR could hinder insurers from testing whether their AI systems pose discrimination-related threats. These two subtopics are, in a sense, the 'literature review' of the thesis: they review the existing literature about the problem I worked out in this section, to gain a deeper understanding.[26] Third, I will look into the AI Act's exception in more detail. Finally, I will investigate how the consumer thinks about the discrimination-related threats I identified.

23 See Tables 1 and 2.

24 Insurers are looking to integrate generative AI into their processes. See Swiss Re (Ladva & Grasso) 2024. See Table 1 for the definitions of *AI system*.

25 See Table 1.

26 See Section 3 (Contributions) and 5 (Method).

2 Key terms

Chapters II to V of the thesis are standalone chapters. Common terms between the chapters may have slightly different meanings. In Table 1, I present a mapping of some key terms, to avoid confusion.

Table 1: Key terms. The precise meaning of some terms varies slightly between chapters, reflecting their specific usage in each context.

Chapt.	Page	Term	Meaning
I	3	Underwriting	Give an activity financial support and take responsibility for paying any costs if it fails.
II	37	Underwriting	Estimating the expected claims cost of a consumer via a risk assessment.
II	39	Data-intensive underwriting	Insurers use and analyse more data for estimating the odds that a consumer files a claim and calculating the price based on that estimation. Insurers could use AI for such underwriting.
II	41	Behaviour-based insurance	Insurers adapt the price based on how a consumer behaves in real-time (for example step counter, car tracker).
II	38	AI System (context: insurance)	The science and engineering of making intelligent machines, especially intelligent computer programs. [...] Intelligence is the computational part of the ability to achieve goals in the world (note: in particular, machine learning).
III	85	AI System (context: sensitive data problem)	Computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.
IV	129	AI System (context: Art. 10(5) AI Act, de-biasing exception))	For a given piece of software to qualify as an AI system, it must have four characteristics: autonomy, adaptiveness, an objective, and the capability to infer.

V	164	AI System (context: AI survey)	The science and engineering of making intelligent machines, especially intelligent computer programs.
III / IV	91	Sensitive data / Special category data	The GDPR, like its predecessor, the Data Protection Directive from 1995, contains an in-principle prohibition (in Art. 9 GDPR) on using certain types of extra sensitive data, called 'special categories' of personal data.
I / V	7	Discriminatio related	Discrimination and Unfair Differentiation
II/III/ IV/V	44	Discrimination	Discrimination occurs only when an insurer differentiates between groups based on legally protected characteristics, such as ethnicity or gender.
II/III/ IV/V	45	Unfair Differentiatio	Differentiation between groups with characteristics that are not legally protected, but the differentiation may nevertheless feel unfair.
II/III/ IV/V	46	Differentiation	Treatment of people or things in different ways.

3 Contributions to existing literature

Non-discrimination by and fairness of AI systems is a 'hot topic': there is much interest in interdisciplinary research into the topic.[27] From the point of view of a non-discrimination lawyer, the insurance sector seems understudied: much work focuses on other sectors, such as the labour market or welfare fraud detection.[28] My PhD thesis contributes to existing work with its interdisciplinary perspective (*where law meets artificial intelligence*), engagement with society, applying different areas of law, and studying norms that go beyond existing non-discrimination laws, such as unfair

27 Apart from the many publications, the topic has large, dedicated conferences and journals. Regulating AI is high on the regulatory agenda worldwide, and a subject of controversy in society.

28 In recent years, labour and welfare fraud detection have seen many scandals involving AI: Welfare scandals at the Dutch government, AI recruitment systems, labour conditions, the Dutch SyRI case. Perhaps these scandals have drawn the interest of many scholars, compared to the insurance sector. Note, however, that the insurance sector also has recent controversial examples, such as the house number example at the beginning of the thesis, and controversies about newer insurance products.

differentiation. While the thesis focuses mostly on European Union Law (see Table 2), the thesis could be useful in other countries, as the majority of the problem statement in Section 1 also applies outside of the EU.[29]

As far as I know, this is the first *law & technology* PhD thesis focusing on the interaction between AI and non-discrimination law in the insurance sector.[30] The closest related PhD thesis is about equality in the insurance sector outside of the AI context, and compares the logic of *actuarial fairness* in insurance with the principle of equality in non-discrimination law.[31] Because this thesis explores new terrain, I present a research agenda in Chapter VI (conclusion), hopefully inspiring future academic work.

Overall, the thesis further works out definitions and insights from previous law & technology research in three ways. First, the thesis contributes to existing work about non-discrimination in insurance. Existing work warned for possible discrimination in insurance (see Section 3.1). In a report for the Council of Europe, in 2019, Zuiderveen Borgesius warned for the effect of *unfair differentiation*: differentiation that escapes current laws, but that nevertheless feels unfair (see also Table 1).[32] This is the first thesis to deepen out unfair differentiation in the insurance sector.

Second, the thesis deepens legal understanding of fundamental rights: part of the problem statement is a possible conflict between the rights to non-discrimination and privacy. Such a conflict was already theorized to be possible, but in this thesis, a concrete conflict came to light (see Section 3.2).

Third and finally, the thesis is a demonstration for how law & tech scholarship could effectively use empirical work such as survey research. For example, the thesis fills in unfair differentiation empirically, because the definition requires a consumer perspective.[33]

29 In countries outside of the EU, insurers may not have a 'special categories of data' problem. On the other hand, in countries such as the US, consumers make strong moral distinctions among types of data, see Section 3.4.

30 For more details about the research method, see Section 5.

31 See Thiery 2010. I mention actuarial fairness briefly in the thesis, in Chapter II.

32 Zuiderveen Borgesius 2019. In later work, Zuiderveen Borgesius adds the 'feels unfair' element. See Zuiderveen Borgesius, *The International Journal of Human Rights* 2020/24.

33 As noted, unfair differentiation is about cases of differentiation that 'feel unfair'. See also Section 5 (Method).

Apart from design choices and the correction of minor spelling mistakes, Chapters II to V are a copy of the content of peer-reviewed journal publications.[34] Most of the chapters are based on papers that I wrote with co-authors.[35] For ease of reading, in the following subsections, I write about the contributions of individual chapters in the first-person singular (*I*) instead of the first-person plural (*we*).[36]

3.1 Chapter II: AI, insurance, discrimination and unfair differentiation

Existing work describes AI use cases in insurance in roughly three ways. First, many authors focus on the data insurers use for underwriting. For example, a report by the European Insurance and Occupational Pensions Authority (EIOPA) describes 'traditional data sources and data sources enabled by digitalisation'.[37] The report describes new data sources for insurers, such as Internet of Things data (for example, data from wearable devices), online media data (for example, web searches), and bank account data (for example, data about a consumer's shopping habits).[38] Such work does not focus on the possible *use* of the data.[39]

Another angle to describe how AI might change insurance is to look at the underwriting process itself, and especially how insurers determine prices. For example, in a report, the Dutch financial authority AFM takes a look at different parts of insurance prices, and which part a particular use case of AI might affect. For instance, if insurers can predict the risk that a consumer might file a claim more accurately and thereby lower their cost, then the cost (which is part of the price) will go down. Another example is if an insurer tracks a consumer's devices to offer them a discount, such as behaviour-

34 I made the citation style of references uniform. The list of author contributions, acknowledgements et cetera is in a table at a separate page of this thesis instead of in footnotes or at the end of the paper, see Author Contributions on page 246 The content of tables, figures and quotes is the same as in the peer-reviewed paper, while I changed the styling to make it uniform in the PhD thesis, like a book. The chapters are not chronological in order of publication.

35 Except for Chapter IV, for which I am the only author.

36 For the attribution of the Chapters' content, see 'Author contributions', page 246.

37 EIOPA 2019.

38 EIOPA 2019, p. 9.

39 EIOPA 2019/6.2.

based car insurance, in which case the consumer gets a discount after the price is already determined.[40]

Third, work by sociologists warns for upcoming types of insurance, such as behaviour-based insurance.[41] Stroobants & Van Schoubroeck discuss some threats for specifically telematics insurance.[42] Insurers themselves also place behaviour-based insurance separately. In an interview by Cevolini & Esposito, an insurer states that behaviour-based insurance:

> [...] is a very different analysis from any other way that insurance is used, right? The rest of insurance is looking at proxy variables ... that are correlated but not causative. What we're trying to do, specifically, is identify those behaviours that are causative of risk.[43]

Because the literature distinguishes between underwriting and behaviour-based insurance, I follow a mixed approach, focusing on both *use cases* of new data, and the data themselves. Building on existing literature, I identify two AI-related trends in insurance: 'data-intensive underwriting' and 'behaviour-based insurance'.

Existing work warns for the effects that AI might have on underwriting. For example, Charpentier & Barry discuss that different types of bias may occur if insurers use AI systems, and these biases could seep into the underwriting process.[44] Bednarz et al. name cherry-picking (insurers pick only low-risk groups of customers) and lemon-dropping (avoiding high-risk groups of customers).[45] In 2014 Swedloff, applying American non-discrimination law, already warned for (roughly speaking) indirect discrimination and non-causal correlations.[46] Other scholars warned that insurers might target

40 Dutch Financial Authority (AFM) 2021/2.
41 Meyers calls behaviour-based insurance a 'not-yet market'. Meyers 2018.
42 Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13.
43 Cevolini & Esposito, *Valuation Studies* 2022/9, p. 129. Cevolini & Esposito studied the actuarial literature from a historical-sociological perspective, interviewed insurers, analyzed insurance reports and followed recent debates.
44 Barry & Charpentier, *Ethics and Information Technology* 2023/25.
45 Bednarz, Lewis & Sadowski, *Computer Law & Security Review* 2025/56.
46 Swedloff 2014/21, p. 360-368.

specific groups[47] or focus on the aspects concerning the data insurers use – for example, whether consumers feel that they have control over a factor.[48]

Chapter II seems the first to clearly compare the two trends, and compare the associated discrimination-related threats systematically, in a single piece of writing. In Chapter II, I apply four factors inspired by previous work to both trends and give some comparative conclusions. The analysis should give a frame of reference from which to discuss to what extent AI is a threat to fairness and non-discrimination in insurance.

3.2 Chapter III: using sensitive data to test AI systems for discrimination. The ban in Art. 9(1) GDPR

Insurers who wish to test whether their AI system discriminates against certain groups run into a problem: the GDPR, in principle, prohibits the insurer from gathering sensitive data.[49] An early paper discussing the problem is from 2016.[50] In a scientific blogpost from 2018, Moerel argues that 'If we want to develop fair algorithms, we must get rid of the taboo of collecting ethnic data'.[51] In 2020, Prince & Schwarz called for 'mandating the collection and disclosure of data about impacted individuals' membership in legally protected classes'.[52]

Chapter III adds the following to the discussion. First, I analyse the ban concerning special categories of personal data in Article 9 GDPR, discussing whether the GDPR could indeed hinder organisations testing whether an AI system discriminates against certain protected groups. Second, I discuss the arguments *for and against* a new exception to the GDPR's ban, allowing

47 Bednarz, Lewis & Sadowski, *Computer Law & Security Review* 2025/56.

48 Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13.

49 For a paper making the point that data protection and non-discrimination law can clash, see Gellert and others/Custers and others 2013, p. 61-89.

50 Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24.

51 L. Moerel, 'Algorithms can reduce discrimination, but only with proper data', https://iapp.org/news/a/algorithms-can-reduce-discrimination-but-only-with-proper-data. The argument of Moerel harmonizes with a broader argument: Van Der Sloot argued that in order to place data in the right context, developers must gather more data than the absolute minimum: data mini*mum*ization. Van Der Sloot/Custers and others 2013, p. 273-287.

52 Prince & Schwarcz, *IOWA LAW REVIEW* 2020/105.

insurers to collect sensitive data for testing AI systems for discrimination. Overall, because the arguments for and against are balanced, I could not take a strong stance for or against an exception.

This chapter seems one of the first to show that the rights to data protection and non-discrimination can conflict in a way that is difficult to balance: within the context discussed in the chapter, it seems difficult to properly respect both rights at the same time, even with safeguards. I created overarching categories of arguments, mapping the literature into arguments for and against. This creates a new overview of the discussion, revealing its complexity. The chapter does not focus specifically on insurance (it focuses on organisations in general), but its conclusions also apply to insurers.[53]

3.3 Chapter IV: using sensitive data to test AI systems for discrimination. The exception in Art. 10(5) GDPR

In Chapter IV, I discuss Article 10(5) AI Act, allowing (at first glance) the developers of AI systems to gather sensitive data for testing for potential biases in AI systems. Apart from providing commentary to a complex legal provision in the AI Act, the chapter lays bare which choices the legislator made in balancing the rights to data protection and non-discrimination.

As far as I know, this is the first article about the balancing act of a conflict between the fundamental rights to data protection and non-discrimination in a specific legal provision. The closest related work is by Hacker, who wrote in general about Article 10 AI Act.[54] More generally, the paper in Chapter IV deeply investigates a complex provision in the new AI Act since its coming into force.[55]

The chapter is useful for scholars and practitioners working on the conflict between two fundamental rights, and investigates a complex legal provision in the long and detailed AI Act. The chapter does not focus specifically on

53 In Chapter VI, I use the insights from Chapter III to answer the central research question, showing that the insights indeed apply to insurers.

54 Hacker wrote a single paragraph about the exception. See Hacker, *Law, Innovation and Technology* 2021/13, p. 297.

55 Because the GDPR was the most relevant law, previous law & technology research that seemed relevant mostly focused on, for example, the GDPR. See Thouvenin and others 2019.

insurance (it focuses on providers/developers), but its conclusions also apply to insurers.[56]

3.4 Chapter V: a survey of AI, insurance, discrimination and unfair differentiation

Little publicly available empirical work exists that analyses the consumer perspective regarding the fairness of AI in insurance.[57] Insurance consumers, in general, seem under-researched: Tanninen writes that 'insurance customers are generally in under-researched, with only a few studies examining the insureds' experiences.'[58] Many existing surveys focus on the expectations of insurers: focusing on whether introducing AI systems changes the insurance sector.[59]

Chapter V is a survey, in which I ask Dutch consumers for their opinions about specific examples of data, inspired by the analysis in Chapter II. The survey seems the first to include examples of behaviour-based insurance. This Chapter contributes to previous work by showing the perspective of the consumer about these data-related trends in relation to discrimination-related threats.

As far as I know, the only work with a similar focus as Chapter V is the work by Kiviat. Kiviat asked American consumers about certain uses of data by insurers. The survey shows how American consumers 'make strong moral distinctions among the types of [...] data used in market transactions'.[60] Kiviat's results show, for example, that consumers care about the relevance (relatedness) of data.[61] I make two distinctions between Kiviat's survey and mine: (1) where Kiviat gathered perspectives of Americans, I surveyed Dutch society, (2) the survey in Chapter V asked about types of unfair differentiation that are not in Kiviat's survey.[62]

56 In Chapter VI, I use the insights from Chapter IV to answer the central research question, showing that the insights indeed apply to insurers.
57 Starke and others, *Big Data & Society* 2022/9.
58 Tanninen, *Big Data & Society* 2020/7, p. 10.
59 E.g. Bank of England & Financial Conduct Authority (FCA) 2022. Cevolini & Esposito, *Valuation Studies* 2022/9.
60 Kiviat, *Sociological Science* 2021/8, p. 27.
61 Kiviat, *Sociological Science* 2021/8, p. 36.
62 See Chapter IV, Section 1 (Introduction).

4 Societal impact

I aim for the thesis to be relevant to society in general.[63] First, the thesis warns for possible threats that certain AI could bring to insurance, and for the societal consequences of those threats. Second, by conducting a survey, the thesis explores the perceptions of people in (Dutch) society regarding the identified discrimination-related threats. The survey reveals that the identified threats are a problem for society: people say they care.[64] Finally, the thesis shows how fighting discrimination and respecting data protection can hinder regulators and developers of AI, creating a deeper understanding of the choices regulators and developers must make. In my opinion, public and democratic debate must decide the outcome of such complex discussions: scholars should not dictate policy.

The thesis's results are also useful for (public and private) organisations. Regulators want organisations to properly test the AI systems they use: the EU regulations attempt to be 'trust-centric'.[65] The thesis creates a better understanding of whether insurers can effectively identify possible discrimination-related threat within the systems they use. By exploring the complex exception in Article 10(5) AI Act, I guide practitioners in understanding one of the complex obligations they have under the long and detailed AI Act. Insurers who wish to introduce AI systems into their practice could use the ideas in the thesis as input. However, I do not aim to criticize insurers and do not give a concrete 'checklist' for preventing the threats mentioned in Chapter I.

Policymakers could use the thesis as input for regulating AI in insurance. European Union legislators and regulators currently do not seem to have a unified approach or answer to how to regulate AI systems in insurance. For example, in the AI Act, the EU regulator chooses to only categorize life and health insurances as 'high-risk'.[66] By contrast, in The Netherlands, the Dutch

63 Other aims I mentioned are: creating a better understanding of the problem in Section 1, and contributing to scholarship in Section 3.

64 As argued in Section 1, taking into account society is important in the context of fairness research.

65 European Commission 2020.

66 AI Act, Annex III.

AFM warns for the impact of certain AI uses for all types of insurance (mainly behaviour-based insurance).[67]

At the time of writing, the PhD project created some value for society through valorisation, in the form of presentations to regulators, spreading the research to legislators and being in contact with Dutch insurers.[68] Some of the papers underlying Chapters II to V have been cited by - or used in the work of - regulatory bodies and legislators across the world.[69]

5 Method

As Section 5.1 will explain, at the abstract level, the research method of the thesis is legal yet interdisciplinary: *where Law meets Artificial Intelligence*. The thesis complements normative research from the field of law with insights from technical fields such as Computing Science (AI, data science, cyber security) and the social sciences. In Section 5.2, I outline the underlying research methods.

5.1 General research approach: law & technology (interdisciplinary)

The thesis's research method fits within the broader field of law and technology. A typical paper in law and technology:

> [Opens] with a description of how a given technological advance has changed social relations and practices, scrutinize[s] the benefits and the dangers the change poses, sometimes look[s] at whether the existing regulations might be sufficient to mitigate the harms, and propose[s] some regulatory intervention.[70]

67 Autoriteit Financiële markten (AFM) 2023.

68 This is part of the general approach of the research, see Section 5.1. For an overview of valorisation of the project, see the table of Author Contributions, page 246.

69 Following their publication, some of the papers in this thesis have been cited and used for (preparatory) legislative work and the work of regulators. See, for example, UNESCO 2023; College voor de Rechten van de Mens [The Netherlands Institute for Human Rights] 2023; Amnesty International 2025; European Parliamentary Research Service 2025.

70 Pałka & Brożek/Brożek, Kanevskaia & Pałka 2023, p. 86. Lightly edited.

The thesis (i) contains many characteristics of law & technology and (ii) while mostly legal, is highly interdisciplinary. Finally, the thesis (iii) is based on a great deal of collaboration with other researchers.

(i) The thesis has two parallels with Law & Technology research. First, the thesis outlines the benefits and dangers of AI in insurance and compares existing regulations. Second, the thesis aims to identify societal problems that fall outside of the scope of current regulations. The thesis includes a social-scientific paper in Chapter V. Complementing legal-normative research with insights from empirical work is still relatively new.[71] The thesis focuses less on proposing regulatory changes and mitigating harms.[72] Instead, I focus on analysing the regulatory problems, which do not have a black-and-white solution, from different perspectives.[73]

(ii) The thesis is inherently interdisciplinary. Interdisciplinary research can be defined as:

> '[A] process of answering a question, solving a problem, or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline, and draws on the disciplines with the goal of integrating their insights to construct a more comprehensive understanding'.[74]

71 Scholars have been arguing for a while that empirical work can enrich (insurance) law and economics theory and is useful for investigating policy implications. See, for example, Van Boom/Faure & Stephen 2008, p. 253-276.

72 Giving recommendations to policymakers had become a common practice in legal research. See Rubin, *Michigan Law Review* 1988/86, p. 1904. In their critical piece, Pałka and Brożek note that sometimes, law & technology scholars warn about a technological trend too quickly: 'the need for regulation is not always self-evident'. Pałka & Brożek/Brożek, Kanevskaia & Pałka 2023, p. 89. For more criticism on such policy-driven research, see Van Gestel & Micklitz, *European Law Journal* 2014/20, p. 10.

73 For example, chapters I and IV raise questions regarding what should and should not be regulated, and chapters II and III raise questions about whether preventing AI discrimination is worth it, when that requires collecting sensitive data.

74 Repko & Szostak 2017, p. 8. In this thesis, 'comprehensive' means 'going beyond a simple sum of its parts'. Utrecht University, 'Multi-, inter-, and transdisciplinarity; what is what?', https://www.uu.nl/en/education/educational-development-training/knowledge-dossiers/interdisciplinary-education-and-cel/multi-inter-and-transdisciplinarity-what-is-what.

Where relevant, the thesis compares different subfields of law in the context of specific topics (for example, data protection and non-discrimination law). I back up technical claims with literature from computing science. Apart from a full education in law, the author of this thesis has a full education in computing science, which likely helped in interpreting some more technical topics and the technical papers cited in the thesis. I do not mean to say that an author's background as such validates results, but the background influences the author's interpretation.[75] In addition, the thesis researches insurance, the literature of which is inherently interdisciplinary.[76]

(iii) While writing the papers underlying the thesis, I collaborated with other researchers: this is part of the research method. Many of the chapters are co-authored with researchers from different backgrounds. I presented the abstract papers at many different venues. I organised several workshops to discuss preliminary results and to hear from academics and stakeholders.[77] Finally, I discussed draft papers with, and received commentary from many well-known and specialized scholars across many disciplines: legal, technical, and social science scholars.[78] Such collaborations are in the spirit of *law and technology* research.[79]

75 See Curriculum Vitae at the end of the thesis, page 237. Some subtopics, such as information regarding synthetic datasets and Secure Multi-Party Computation in Chapter III and IV, required technical computer science literature.

76 See, for example, Chapter II.

77 See Other Contributions, page 251.

78 See the Author Contributions for an overview of the collaborations, page 246.

79 A good way to prevent wrong assumptions in law & technology research is to 'engage in an interdisciplinary collaboration with scholars from other disciplines – engineers, computer scientists, marketing experts, etc [. . .].' Pałka & Brożek/Brożek, Kanevskaia & Pałka 2023, p. 88.

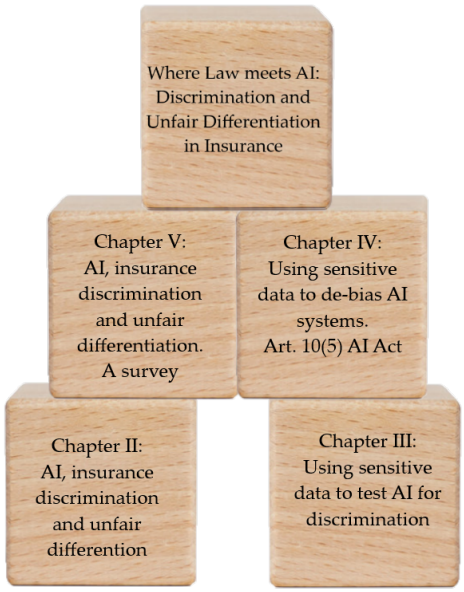


Figure 1: The thesis as a tower of blocks.

Figure 1 sketches the building blocks of the thesis. Chapters II and III could be seen as literature reviews: they explore the problem statement in Section 1 more deeply. Chapter IV (investigating the AI Act Exception) builds on the insights from Chapter III (the ban on special category data). Chapter V is a survey based on the insights from Chapter II (AI, insurance, discrimination and unfair differentiation). As follows from Table 1, the definition of *AI system* is context-dependent for each of the Chapters II to V: a deliberate choice. Using the same definitions for all chapters could hinder the analysis of other chapters. For example, applying the AI Act's definition (Chapter IV) to Chapter II could have led to an overly broad analysis that does not match the insurance sector.

5.2 Underlying research methods (disciplinary)

In this section, I outline the disciplinary research methods I used in writing the Chapters II to V. First, in the thesis, I *analysed literature*. In Chapter II,

I describe AI in insurance mostly by discussing insurance reports and use some more technical sources to support the definitions of AI. In Chapter II, I identify two trends of AI in insurance, which are based on existing empirical work by sociologists and insurance reports: experiments and interviews with insurers. In Chapter III, I discuss arguments for and against. These arguments come from literature from several fields: papers from non-discrimination law, privacy law, from historians, and the 'computing science' fairness community.

Second, the research is *legal-normative:* I interpret the law using various interpretation angles.[80] In the normative part of Chapters II, III and IV, I discuss the EU non-discrimination directives and discuss the most relevant legal cases, the GDPR and the AI Act. The thesis looks mainly into the following European Union directives and regulations:

Table 2: Overview of the most relevant legislation in the thesis

Chapter	Short legislation name	Full legislation name
II & III	Race and Ethnicity Equality Directive	Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Directive 2000/43/EC).
II & III	Gender Equal Access to Goods and Services Directive	Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services (Directive 2004/113/EC).
III	General Data Protection Regulation (GDPR) (Mostly Art. 9)	Regulation (EU) 2016/179 of the European Parliament and the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

80 For a book about the legal research method, see Kestemont 2018. For more about the descriptive (doctrinal) legal research method, see Hutchinson & Duncan, *Deakin Law Review* 2012/17.

IV	Artificial Intelligence Act (AI Act) (Mostly Art. 10)	Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

I did not use some methods that are common in dogmatic legal research. For example, the systemisation and analysis of case law plays a minor role, for three reasons. First, non-discrimination by AI is a relatively new problem in law, case law is scarce. Second, some of the legislation is also too new: they do not compare to the classical fields of law (for example, administrative law). Third, the thesis has a broader scope than non-discrimination law alone (unfair differentiation). The thesis is the first to analyse certain laws and provisions (e.g., the AI Act) and discusses the most relevant case law that exists.

Third, the thesis concerns fairness, in the context of *unfair differentiation* in Chapter II. This part is also normative, but not legal-normative. I base *unfair differentiation* loosely on existing work concerning fair AI in insurance and on some first attempts by philosophers.

Finally, in Chapter V, I present a survey among 1000 Dutch residents about the public perceptions on possible data-driven pricing. The survey should give a first impression of what consumers think about the discrimination-related threats from Chapter II. The impression of what consumers think contributes to answering the central research question by (i) allowing a reflection across chapters and (ii) highlighting another perspective missing in much existing work: the consumer perspective.

The survey complements Chapter II, because the definition of certain threats includes the word 'feeling', requiring such a consumer perspective.[81] With the survey, I aim to inform policy. While the thesis does not aim to reform the

81 See the definition of 'unfair differentiation' in Table 1.

law, surveys such as this one can provide evidence that current laws do not protect consumers.[82]

The analysis in the thesis is based on the literature available and developments occurring up to 25 July 2025. For specific information about the method of each chapter, see Chapters II to V.

6 Structure of the thesis

The thesis is composed as follows. For reference, see also Figure 1. Each of Chapters II-V is standalone: readers can start any one of them without reading the other chapters. Chapter II describes the two AI-related trends in insurance and discusses the discrimination-related risks of using AI for insurance underwriting. Chapter III discusses the GDPR's ban on collecting sensitive data for AI de-biasing in Article 9(1) GDPR, and the arguments for and against an exception to the ban. Chapter IV evaluates the EU AI Act's exception (Article 10(5)) for using the sensitive data for bias detection and correction. Chapter V presents the survey about the public perceptions regarding specific examples. Chapter VI concludes.

82 Salter & Mason write that '[Sociological research that takes the policy dimension of legal topics into account] will (…) be supported by evidence that changes in social patterns, lifestyles, attitudes and economic circumstances now mean that the policy underlying a particular area of legal regulation has become outdated and anachronistic, even if it fully meets the aspirations of the black-letter model. Salter & Mason 2007, p. 162-163.

References of chapter I

Amnesty International 2025
Amnesty International, *Etnisch profileren is overheidsbreed probleem*, Tweede Kamer 2025, https://www.tweedekamer.nl/kamerstukken/detail?id=2025D04274&did=2025D04274.

Autoriteit Financiële markten (AFM) 2023
Autoriteit Financiële markten (AFM), *Technology towards 2033. The future of insurance and supervision [Technologie richting 2033. De toekomst van verzekeren en toezicht]*, AFM 2023, https://www.afm.nl/nl-nl/sector/actueel/2023/april/kansen-risico-digitalisering-verzekeringsmarkt.

Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019
Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB), *Artificial intelligence in the insurance sector. An exploratory study*, AFM & DNB 2019, https://www.afm.nl/~/profmedia/files/rapporten/2019/afm-dnb-verzekeringssector-ai-eng.pdf.

Bank of England & Financial Conduct Authority (FCA) 2022
Bank of England & Financial Conduct Authority (FCA), *Machine Learning in UK Financial Services*, 2022, https://www.bankofengland.co.uk/Report/2022/machine-learning-in-uk-financial-services.

Barry & Charpentier, *Ethics and Information Technology* 2023/25
L. Barry & A. Charpentier, 'Melting contestation: insurance fairness and machine learning', *Ethics and Information Technology* 2023/25, issue 4, p. 49, https://link.springer.com/10.1007/s10676-023-09720-y, doi:10.1007/s10676-023-09720-y.

Bednarz, Lewis & Sadowski, *Computer Law & Security Review* 2025/56
Z. Bednarz, K. Lewis & J. Sadowski, '"It's not personal, it's strictly business": Behavioural insurance and the impacts of non-personal data on individuals, groups and societies', *Computer Law & Security Review* 2025/56, p. 106096, https://linkinghub.elsevier.com/retrieve/pii/S0267364924001614, doi:10.1016/j.clsr.2024.106096.

Bhattacharya and others, *Frontiers in Artificial Intelligence* 2025/8
S. Bhattacharya and others, 'AI revolution in insurance: bridging

research and reality', *Frontiers in Artificial Intelligence* 2025/8, p. 1568266, https://www.frontiersin.org/articles/10.3389/frai.2025.1568266/full, doi:10.3389/frai.2025.1568266.

Cevolini & Esposito, *Valuation Studies* 2022/9
A. Cevolini & E. Esposito, 'From Actuarial to Behavioural Valuation. The impact of telematics on motor insurance', *Valuation Studies* 2022/9, issue 1, p. 109-139, https://valuationstudies.liu.se/article/view/3697, doi:10.3384/VS.2001-5992.2022.9.1.109-139.

Charpentier & Vamparys, *Big Data & Society* 2025/12
A. Charpentier & X. Vamparys, 'Artificial intelligence and personalization of insurance: Failure or delayed ignition?', *Big Data & Society* 2025/12, issue 1, p. 20539517241291817, https://journals.sagepub.com/doi/10.1177/20539517241291817, doi:10.1177/20539517241291817.

College voor de Rechten van de Mens [The Netherlands Institute for Human Rights] 2023
College voor de Rechten van de Mens [The Netherlands Institute for Human Rights], *Dating-app Breeze mag (en moet) algoritme aanpassen om discriminatie te voorkomen [Dating app Breeze may (and must) adjust algorithm to prevent discrimination]*, Ministerie van Algemene Zaken 2023, https://www.mensenrechten.nl/actueel/nieuws/2023/09/06/dating-app-breeze-mag-en-moet-algoritme-aanpassen-om-discriminatie-te-voorkomen.

De Jonge & Hiemstra, *ACM Conference on Fairness, Accountability, and Transparency* 2023
T. De Jonge & D. Hiemstra, 'UNFair: Search Engine Manipulation, Undetectable by Amortized Inequity', *ACM Conference on Fairness, Accountability, and Transparency* 2023, https://dl.acm.org/doi/10.1145/3593013.3594046, doi:10.1145/3593013.3594046.

Dutch Financial Authority (AFM) 2021
Dutch Financial Authority (AFM), *The personalisation of prices and conditions in the Insurance Sector. An exploratory study*, The Netherlands: AFM 2021, https://www.afm.nl/en/sector/actueel/2021/juni/aandachtspunten-gepersonaliseerde-beprijzing.

EIOPA 2019
EIOPA, *Big data analytics in motor and health insurance: a thematic review*, Luxembourg: Publications Office of the European Union 2019.

EIOPA 2024
EIOPA, *Report on the digitalisation of the insurance sector EIOPA-BoS-24/139*, 2024, www.eiopa.europa.eu/publications/eiopas-report-digitalisation-european-insurance-sector_en.

European Commission 2020
European Commission, *COM(2020) 65 White paper On Artificial Intelligence - A European approach to excellence and trust*, EU 2020, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Insurance and Occupational Pensions Authority (EIOPA) 2021
European Insurance and Occupational Pensions Authority (EIOPA), *Artificial Intelligence governance principles, towards ethical and trustworthy Artificial Intelligence in the European insurance sector: a report from EIOPA´s Consultative Expert Group on Digital Ethics in insurance.*, LU: Publications Office 2021, https://data.europa.eu/doi/10.2854/49874.

European Parliamentary Research Service 2025
European Parliamentary Research Service, *Algorithmic discrimination under the AI Act and the GDPR*, European Parliament 2025, https://www.europarl.europa.eu/thinktank/en/document/EPRS_ATA(2025)769509.

Gandy 2016
O.H. Gandy, *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*, Routledge 2016.

Gellert and others/Custers and others 2013
R. Gellert and others/B. Custers and others, 'A Comparative Analysis of Anti-Discrimination and Data Protection Legislations', in: *Discrimination and Privacy in the Information Society* (Studies in Applied Philosophy, Epistemology and Rational Ethics), Berlin, Heidelberg: Springer Berlin Heidelberg 2013, p. 61-89, http://link.springer.com/10.1007/978-3-642-30487-3_4, doi:10.1007/978-3-642-30487-3_4.

Gupta, *The Actuary Magazine* 2024
D. Gupta, 'The History of Actuarial Science', *The Actuary Magazine* 2024, https://www.theactuarymagazine.org/the-history-of-actuarial-science/.

Hacker, *Law, Innovation and Technology* 2021/13
P. Hacker, 'A legal framework for AI training data—from first principles to the Artificial Intelligence Act', *Law, Innovation and Technology* 2021/13, issue 2, p. 257-301, https://www.tandfonline.com/doi/full/10.1080/17579961.2021.1977219, doi:10.1080/17579961.2021.1977219.

Hildebrandt 2015
M. Hildebrandt, *Smart Technologies and the End(s) of Law: novel entanglements of law and technology*, Cheltenham, UK Northampton, MA, USA: Edward Elgar Publishing 2015, https://china.elgaronline.com/monobook/9781849808767.xml.

Hong 2020
S. Hong, *Technologies of speculation: the limits of knowledge in a data-driven society*, New York: New York University Press 2020.

Hutchinson & Duncan, *Deakin Law Review* 2012/17
T. Hutchinson & N. Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research', *Deakin Law Review* 2012/17, issue 1, p. 83, https://ojs.deakin.edu.au/index.php/dlr/article/view/70, doi:10.21153/dlr2012vol17no1art70.

Insurance Europe 2019
Insurance Europe, *Big data and its big benefits for insurance consumers*, 2019, https://www.insuranceeurope.eu/publications/503/insight-briefing-big-data-and-its-big-benefits-for-insurance-consumers/download/Big+data%20Insight%20Briefing-January%202019.pdf.

Kestemont 2018
L. Kestemont, *Handbook on legal methodology: from objective to method*, Cambridge Antwerp Portland: Intersentia 2018.

Kieslich, Keller & Starke, *Big Data & Society* 2022/9
K. Kieslich, B. Keller & C. Starke, 'Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence', *Big Data & Society* 2022/9,

issue 1, p. 20539517221092956, https://doi.org/10.1177/20539517221092956, doi:10.1177/20539517221092956.

Kiviat, *Sociological Science* 2021/8
B. Kiviat, 'Which Data Fairly Differentiate? American Views on the Use of Personal Data in Two Market Settings', *Sociological Science* 2021/8, p. 26-47, https://sociologicalscience.com/articles-v8-2-26/, doi:10.15195/v8.a2.

Meyers 2018
G. Meyers, *Behaviour-based Personalisation in Health Insurance: a Sociology of a not-yet Market*, KU Leuven 2018, https://lirias.kuleuven.be/2087689&lang=en.

Minty 2023
D. Minty, *Revolutionising fairness to enable digital insurance*, Institute and Faculty of Actuaries 2023, https://actuaries.org.uk/media/z5rh5jhl/revolutionising-fairness-to-enable-digital-insurance.pdf.

Pałka & Brożek/Brożek, Kanevskaia & Pałka 2023
P. Pałka & B. Brożek/B. Brożek, O. Kanevskaia & P. Pałka, 'How not to get bored, or some thoughts on the methodology of law and technology', in: *Research Handbook on Law and Technology*, Edward Elgar Publishing 2023, p. 82-98, https://www.elgaronline.com/view/book/9781803921327/chapter6.xml, doi:10.4337/9781803921327.00013.

Prince & Schwarcz, *IOWA LAW REVIEW* 2020/105
A.E.R. Prince & D. Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data', *IOWA LAW REVIEW* 2020/105, p. 62.

Rahwan, *Ethics and Information Technology* 2018/20
I. Rahwan, 'Society-in-the-loop: programming the algorithmic social contract', *Ethics and Information Technology* 2018/20, issue 1, p. 5-14, https://doi.org/10.1007/s10676-017-9430-8, doi:10.1007/s10676-017-9430-8.

Repko & Szostak 2017
A.F. Repko & R. Szostak, *Interdisciplinary Research: Process and Theory*, UK: Sage 2017.

Rubin, *Michigan Law Review* 1988/86
E.L. Rubin, 'The Practice and Discourse of Legal Scholarship', *Michigan*

Law Review 1988/86, issue 8, https://www.jstor.org/stable/1289072?origin=crossref, doi:10.2307/1289072.

Salter & Mason 2007
M. Salter & J. Mason, *Writing law dissertations: an introduction and guide to the conduct of legal research*, Harlow, England ; New York: Pearson/Longman 2007.

Van Der Sloot/Custers and others 2013
B. van der Sloot/B. Custers and others, 'From Data Minimization to Data Minimummization', in: *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Berlin, Heidelberg: Springer Berlin Heidelberg 2013, p. 273-287, https://doi.org/10.1007/978-3-642-30487-3_15, doi:10.1007/978-3-642-30487-3_15.

Starke and others, *Big Data & Society* 2022/9
C. Starke and others, 'Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature', *Big Data & Society* 2022/9, issue 2, p. 205395172211151, http://journals.sagepub.com/doi/10.1177/20539517221115189, doi:10.1177/20539517221115189.

Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13
N. Stroobants & C. Van Schoubroeck, 'Telematics Insurance: Legal Concerns and Challenges in the EU Insurance Market', *European Journal of Commercial Contract Law* 2021/13, issue 3, p. 51-81, https://www.ingentaconnect.com/content/10.7590/187714621X16463138640038, doi:10.7590/187714621X16463138640038.

Swedloff 2014/21
R. Swedloff, 'Risk Classification's Big Data (R)evolution', 2014/21, p. 36.

Swedloff, *61 B.C. L. REV.* 2020/2031
R. Swedloff, 'The new regulatory imperative for insurance', *61 B.C. L. REV.* 2020/2031, p. 55.

Swiss Re (Ladva & Grasso) 2024
Swiss Re (Ladva & Grasso), *The evolution of AI in the insurance industry*, Swiss Re 2024, https://www.swissre.com/risk-knowledge/advancing-societal-benefits-digitalisation/evolution-of-ai-in-insurance-industry.html.

Tanninen, *Big Data & Society* 2020/7
M. Tanninen, 'Contested technology: Social scientific perspectives of behaviour-based insurance', *Big Data & Society* 2020/7, issue 2, p. 205395172094253, http://journals.sagepub.com/doi/10.1177/2053951720942536, doi:10.1177/2053951720942536.

The Geneva Association (Noordhoek) 2023
The Geneva Association (Noordhoek), *Regulation of Artificial Intelligence in Insurance: Balancing consumer protection and innovation*, The Geneva Assocation 2023, https://www.genevaassociation.org/publication/public-policy-and-regulation/regulation-artificial-intelligence-insurance-balancing.

Thiery 2010
Y. Thiery, *Discriminatie en verzekering. Economische efficiëntie en actuariële fairness getoetst aan het juridisch gelijkheidsbeginsel*, KU Leuven 2010, https://lirias.kuleuven.be/1844266&lang=en.

Thouvenin and others 2019
F. Thouvenin and others, 'Big Data in the insurance industry: Leeway and limits for individualising insurance contracts', 2019, https://www.zora.uzh.ch/id/eprint/179171, doi:10.5167/UZH-179171.

UNESCO 2023
UNESCO, *Ethical impact assessment: a tool of the Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library*, 2023, https://unesdoc.unesco.org/ark:/48223/pf0000386276.

Urban Institute 2024
Urban Institute, 'Evidence of Disparities in Access to Mortgage Credit', 2024, https://www.urban.org/sites/default/files/2024-03/Evidence_of_Disparities_in_Access_to_Mortgage_Credit.pdf.

Van Boom/Faure & Stephen 2008
W.H. van Boom/M. Faure & F. Stephen, 'Insurance Law and Economics: An Empirical Perspective', in: *Essays in the Law and Economics of Regulation - in honour Anthony Ogus*, Antwerp: Intersentia 2008, p. 253-276, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1290237.

Van Gestel & Micklitz, *European Law Journal* 2014/20
R. Van Gestel & H. Micklitz, 'Why Methods Matter in European Legal

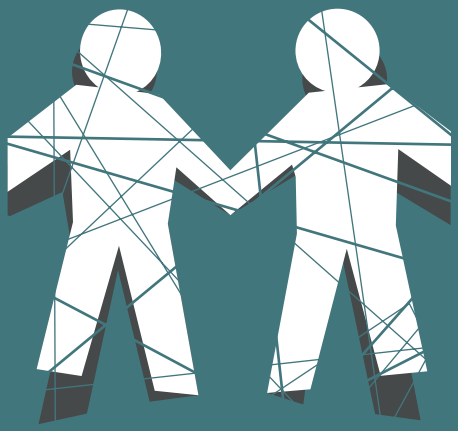
Scholarship', *European Law Journal* 2014/20, issue 3, p. 292-316, https://onlinelibrary.wiley.com/doi/10.1111/eulj.12049, doi:10.1111/eulj.12049.

Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023
H. Weerts and others, 'Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree', *ACM Conference on Fairness, Accountability, and Transparency* 2023, https://dl.acm.org/doi/10.1145/3593013.3594044, doi:10.1145/3593013.3594044.

Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24
I. Žliobaitė & B. Custers, 'Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models', *Artificial Intelligence and Law* 2016/24, issue 2, p. 183-201, http://link.springer.com/10.1007/s10506-016-9182-5, doi:10.1007/s10506-016-9182-5.

Zuiderveen Borgesius 2019
F.J. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making. Report for the European Commission against Racism and Intolerance (ECRI)*, Council of Europe 2019.

Zuiderveen Borgesius, *The International Journal of Human Rights* 2020/24
F.J. Zuiderveen Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence', *The International Journal of Human Rights* 2020/24, issue 10, p. 1572-1593, https://www.tandfonline.com/doi/full/10.1080/13642987.2020.1743976, doi:10.1080/13642987.2020.1743976.



II.

AI, insurance, discrimination and unfair differentiation. An overview and research agenda

Submitted to *Law, Innovation and Technology* as:

M. van Bekkum, F. Zuiderveen Borgesius & T. Heskes, 'AI, insurance, discrimination and unfair differentiation: an overview and research agenda', *Law, Innovation and Technology* 2025, p. 1-28, https://www.tandfonline.com/doi/full/10.1080/17579961.2025.2469348, doi:10.1080/17579961.2025.2469348.

The authors are especially grateful to Bart van der Sloot (Tilburg University), Paola Lopez (University of Vienna) and Balázs Bodó (University of Amsterdam) for their insightful comments during the PLSC-style session of Digital Legal Talks 2023. We also thank Jos Schaffers (Dutch Association of Insurers) and Maaike Harbers (Rotterdam University of Applied Sciences). We also thank the participants of the workshop 'Insurance, algorithmic decision-making, and discrimination' at iHub, Radboud University (2022).

Abstract

Insurers underwrite risks: they calculate risks and decide on the insurance price. Insurers seem captivated by two trends enabled by Artificial Intelligence (AI). First, insurers could use AI for analysing more and new types of data to assess risks more precisely: data-intensive underwriting. Second, insurers could use AI to monitor the behaviour of individual consumers in real-time: behaviour-based insurance. For example, some car insurers offer a discount if the consumer agrees to being tracked by the insurer and drives safely. While the two trends bring many advantages, they may also have discriminatory effects on society. This paper focuses on the following question. Which effects related to discrimination and unfair differentiation may occur if insurers use data-intensive underwriting and behaviour-based insurance?

1 Introduction

Insurers offer important services to modern societies. For example, motor insurance is important for people's mobility, household insurance for protecting property, and life insurance for protecting family members against poverty.[1] Meanwhile, most insurers are companies aiming for profit. Consumers pay a price to an insurer to have their risks covered. Based on, for example, the consumer's car age and weight, insurers calculate the risk that the consumer will file a claim, and base the premium (the price of insurance) on the risk.[2]

We identify two AI-related trends in the insurance sector. (i) The first trend is data-intensive underwriting: insurers experiment with AI to calculate risks and to set insurance prices. For example, insurers could analyse new types of data with the help of AI. (ii) The second trend is behaviour- based insurance: insurers increasingly monitor the behaviour of individual consumers.[3] In this paper, we explore the following question: Which effects related to discrimination and unfair differentiation may occur if insurers use data-intensive underwriting and behaviour-based insurance?

Our main contributions to the literature are as follows. First, as far as we know, we are the first to compare both trends in detail.[4] Our paper makes the point that many threats concerning AI systems are not necessarily about machine learning as a specific technology, but rather about organisations using vast datasets to make decisions about people. Second, in our analysis, we include insights from legal, philosophical, sociological and computer science literature.[5] We aim to avoid unnecessary jargon in our writing, making the paper more accessible to non-specialists.

1 European Insurance and Occupational Pensions Authority (EIOPA) 2021, p. 24.

2 In the rest of the paper, we use the word *price* instead of *premium*. Insurance Europe 2012.

3 For example, some life and health insurers give discounts to consumers wearing fitness trackers. See e.g. Henkel, Heck & Göretz/Meiselwitz 2018, p. 28-49. Krüger & Ní Bhroin, *The Journal of Media Innovations* 2020/6. Martani, Shaw & Elger, *Swiss Medical Weekly* 2019.

4 Most other papers consider one of the two trends, or do not contrast the trends. For examples in other work, see e.g. Infantino, *European Review of Private Law* 2022/4, p. 613. Cevolini & Esposito, *Big Data & Society* 2020/7. Prince & Schwarcz, *IOWA LAW REVIEW* 2019/105. Charpentier 2022.

5 The paper defines AI based on technical definitions and discusses causality in the paragraphs concerning 'seemingly irrelevant characteristics'. For a more technical

Third, we analyse both trends, discussing them from the perspectives of non-discrimination law and fairness.[6] We assess the fairness of the two trends largely based on existing literature about insurance fairness. We use the distinction between discrimination (in the narrow legal sense) and unfair differentiation as we proposed in earlier work.[7] Finally, we provide a research agenda.

A few remarks about the scope of the paper. Where we write 'insurer', we mean any type of company offering insurance. This could include a tech startup just entering the insurance business, but also a centuries-old insurance company. Our paper focuses on insurance underwriting, the insurance industry's core business. Since we focus on discrimination and other unfair differentiation, many questions are outside the scope of the paper. For example, we do not discuss privacy or data protection, and we do not discuss the right to healthcare (relevant for health insurance).[8]

We do not discuss all possible AI applications. For example, AI-driven fraud detection is outside the scope of the paper. We focus mostly on countries within the European Union. However, the paper could be relevant outside Europe.[9] We introduce some aspects of non-discrimination law, but do not discuss other fields of law. The paper focuses on AI and the processing of vast amounts of data in the insurance sector, but can also be useful for research and policy in other sectors working with AI: the insurance sector has decades of experience with statistics, AI and the regulation of AI.[10] Some scholars see

discussion, see Greenland 2002. Our paper also mentions the fair machine learning literature where relevant, such as in Section 4.2.2.

6 Due to length limitations, we will not apply all possible notions of fairness in insurance. For example, we do not explicitly discuss 'solidarity'. We apply certain elements of solidarity, such as exclusion.

7 Zuiderveen Borgesius 2019, p. 7, 67.

8 See about the GDPR, data privacy and insurance e.g. Thouvenin and others 2019. Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13. Truby, Brown & Antoine Ibrahim, *International Review of Law, Computers & Technology* 2023, p. 1-25. Bednarz, Lewis & Sadowski, *Computer Law & Security Review* 2025/56.

9 For example, in the United States, some insurers collaborate with car manufacturers to include behaviour-based insurance in smart cars. See https://www.nytimes.com/2024/03/11/technology/carmakers-driver-tracking-insurance.html.

10 Gandy 2016. Bouk 2015.

insurance as 'an insightful analogue for the social situatedness and impact of machine learning systems'.[11]

The paper is structured as follows. We first summarise (in Section 2) how insurance functions and discuss data-intensive underwriting and behaviour- based insurance. In Section 3, we discuss the difference between non-discrimination and fairness. In Sections 4 and 5, we discuss non-discrimination and fairness aspects of the two trends. In section 6, we present a research agenda. Section 7 concludes.

2 Insurance, AI, and two trends

This section discusses insurance underwriting (2.1), AI for insurance underwriting (2.2), two trends related to AI in insurance (2.3 and 2.4) and the similarities and differences between the two trends (2.5).

2.1 Underwriting

The core business of insurers is underwriting risks: estimating the expected claims cost of a consumer via a risk assessment.[12] Roughly summarised, risk-based underwriting works as follows. First, the insurer collects characteristics about the individual consumer, such as the consumer's car type for car insurance. The insurer assigns the consumer to a *risk pool*, a group of consumers with the same characteristics. The insurer estimates the probability that the consumer might file a future claim and the amount in Euros of the claim: the expected claim cost. The insurer estimates the claim cost on, for example, past data about the severity (size) of claims and the number of claims within the same risk pool.[13]

To illustrate with a simple example, take a consumer Alice who has one characteristic: 'drives in a light car'. A car insurer assigns a low estimated

11 Fröhlich & Williamson 2024/7.

12 The Geneva Association, the global association of insurance companies, describes underwriting as 'a core process of insurance that involves assessing and pricing risks presented by applicants seeking insurance coverage'. The Geneva Association 2020, p. 17.

13 Insurance Europe 2021a. Insurance Europe 2021b. Insurance Europe 2012, p. 12. Some insurers also assign a weight to the rating factors. EIOPA 2019, p. 34.

claim cost to Alice, because (as shown by past data) consumers with a light car cause less damage in an accident. The insurer charges Alice a low price (plus possible extra costs or discounts). Let us now take the same example for another consumer, Bob. Bob has two characteristics: 'drives in a light car' and 'is senior'. The underwriter assumes a higher estimated claim cost for Bob than for Alice, because (as shown by past data) older drivers are more often involved in accidents. The insurer decides that Bob's risk pool represents a medium risk. The insurer charges Bob a medium price (again, plus possible extra costs or discounts).[14]

2.2 AI systems in insurance

Many insurers use statistical regression models to analyse datasets.[15] The insurer fits such a model to the data: for different consumer characteristics, the model should reflect the claims in the dataset as accurately as possible.[16] For example, a car insurance model could express that elderly drivers represent a higher risk, which follows from the data. We will use the term 'traditional' for such underwriting. Another example of traditional underwriting could be an insurer who finds a statistical correlation between postal codes and the number or burglaries. The insurer can then use the postal code for risk pooling in, e.g. home insurance.

Many insurers also use forms of AI to underwrite risks or to analyse past data. AI can be defined as 'the science and engineering of making intelligent machines, especially intelligent computer programs. [...] Intelligence is the computational part of the ability to achieve goals in the world'.[17] According to the Geneva Association, the international association of insurance companies, a promising type of AI for underwriting is machine learning.[18] The promise of such systems is that insurers can more accurately analyse more data. Machine learning systems can 'learn from data and automate the process of analytical model building and solve associated

14 In practice, the risk could be more granular than only 'low, medium or high', such as a number between 0 and 1.

15 Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019, p. 8.

16 A model is a representation of reality.

17 John McCarthy, 'What is AI? Basic Questions', http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.

18 The Geneva Association (Noordhoek) 2023, p. 10, 15, 17-19.

tasks'.[19] Machine learning models could identify new attributes that indicate risk in datasets, without the insurer defining in advance which attributes the machine learning system should consider.[20] The insurer could calculate prices based on the new attributes. The underwriting process itself remains largely the same: insurers charge higher prices to riskier consumers.[21]

In countries such as the United Kingdom, insurers increasingly use machine learning for their underwriting practices.[22] According to the UK Financial Conduct Authority, a UK insurer uses machine learning-based application to pre-approve consumers for life insurance using data already available from bank accounts and credit ratings.[23] In other countries, such as the Netherlands, insurers seem more cautious of using machine learning systems for underwriting.[24] In the following sections, we describe two AI- related trends in the insurance sector.

2.3 Trend 1: data-intensive underwriting

The first AI-related trend is data-intensive underwriting: insurers use and analyse more data for estimating the odds that a consumer files a claim and calculating the price based on that estimation. Insurers could use AI for underwriting in two different ways. First, insurers could use AI systems to analyse the data they already have. Second, insurers could use AI systems to analyse more data and new types of data. For more than a century, insurers gathered data directly from the consumers: for example, demographic data such as a person's age, or the consumer's smoking and drinking habits.[25]

19 Janiesch, Zschech & Heinrich, *Electronic Markets* 2021/31, p. 685.
20 The Geneva Association (Noordhoek) 2023, p. 10-11.
21 See Section 2.1. See also The Geneva Association (Noordhoek) 2023, p. 17. We note that insurers could also automate other processes that affect the acceptance of consumers (non-risk-based pricing). For example, an insurer could use an (machine learning) algorithm to pre-approve customers based on their bank account data. Or an insurer could assess the willingness to pay of the consumer. See Davey/Arvind & Steele 2020, p. 269-294.
22 Piesse 2023. See on a global scale Jesus, Brito & Duarte 2023, p. 19-24.
23 Bank of England & Financial Conduct Authority (FCA) 2022/5.3.
24 Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019, p. 8.
25 EIOPA 2019, p. 9 and further.

Insurers used questionnaires to gather consumer characteristics such as the type of car they drive.[26]

Now, insurers could analyse more data gathered from many different sources. For example, insurers could analyse a consumer's web searches on the internet (online media data), a consumer's shopping habits (bank account data), or selfies to estimate the consumer's age.[27] To illustrate, Dutch insurers expect to use social media data, bank account data, and data gathered from the Internet of Things devices.[28] Furthermore, insurers could analyse data with machine learning to predict risks more accurately.[29]

Many insurers have high hopes for data-intensive underwriting, but the trend is only slowly catching on in Europe. The European Union's data protection laws may slow down the abilities of insurers to gather the new types of data compared to other countries.[30] And there seem to be practical limitations. For instance, Dutch insurers are not yet convinced of the value of social media data.[31]

However, the limitations will not prevent data-intensive underwriting altogether. If insurers find new insights from applying machine learning to large datasets, we can expect more data-intensive underwriting in Europe as well. The Dutch financial authority AFM expects the trend to catch on in the Netherlands and the rest of Europe:

> Digitisation is undeniably going to have an impact on the insurance sector and its supervision. Insurers are gathering more data, applying smarter data analytics, and in doing so influence their entire operations and distribution, from pricing to claims handling. This applies not only to Dutch insurers but to many European parties [...].[32]

26 Barry & Charpentier, *Big Data & Society* 2020/7, p. 3-6.
27 EIOPA 2019, p. 9 and further. More examples are possible, even emotion detection: McStay & Urquhart, *International Review of Law, Computers & Technology* 2022/36, p. 470.
28 Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019.
29 See Section 2.1. EIOPA 2019, p. 12.
30 Infantino, *European Review of Private Law* 2022/4. Thouvenin and others 2019, p. 242.
31 Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019.
32 Translated by authors. Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019, p. 3.

In sum, the first insurance trend enabled by AI is data-intensive underwriting.

2.4 Trend 2: behaviour-based insurance

The second AI-related trend is that insurers adapt the price based on how the consumer behaves in real-time: behaviour-based insurance. For example, a life or health insurer may offer a discount to consumers who, according to their health tracker, walk a certain number of steps.[33] A car insurer may offer a discount to consumers who, according to a device in their car, drive safely.[34] Such car insurances are often called telematics insurance.

These two examples of behaviour-based insurance are a first step towards individualising risk. Scholars started writing about such individualisation since the 80's.[35] In 1981, Walters stated that 'since the insurer assumes the *individual* insured's risk of loss, the price should be fundamentally based upon the expected value of an insured's losses'.[36]

In recent years, new behaviour-based insurance products are being introduced. European insurers are interested in personalising the insurance of their consumers, and reports by insurers highlight behaviour-based insurance as an important trend.[37] Insurance could become more and more behaviour-based in the future: a 'not-yet market'.[38]

The prevalence of behaviour-based insurance varies from country to country in Europe. Italy was an early adopter of behaviour-based insurance. Italian insurance companies successfully introduced behaviour-based insurance by offering extra services and discounts to the originally high price of the Italian motor third-party liability insurance.[39] Another early adopter is the United

33 Meyers 2018, p. 18.
34 The device measures, for instance, how often someone brakes hard, how fast someone drives, and how someone steers. Cevolini & Esposito, *Valuation Studies* 2022/9, p. 109, 111-112. Meyers & Hoyweghen, *Big Data & Society* 2020/7, p. 7.
35 See Barry & Charpentier, *Big Data & Society* 2020/7, p. 5.; Frezal & Barry, *Journal of Business Ethics* 2020/167, p. 127, 129. Walters 1981.
36 Walters 1981, p. 3.
37 The Geneva Association (Noordhoek) 2023.
38 Meyers 2018.
39 Swiss Re 2017, p. 7.

Kingdom. The UK Financial Conduct Authority writes: 'Some [insurers] use telematics risk modelling based on Deep Neural Networks to estimate driver behaviour and, thereby, predict the magnitude of the claim and determine the prices charged to the consumer.'[40] In the Netherlands, behaviour-based car insurance is rarer.[41]

So far, in Europe, insurers have not adopted behaviour-based insurance for all their (car or life) insurance. One factor is cost: it costs too much for insurers to keep track of the behaviour of individual consumers, and creating a behaviour-based infrastructure also comes with high fixed costs for collecting data and administration.[42] While in this section we focus on the existing products in behaviour-based car and life insurance, insurers could use other Internet of Things (IoT) devices to offer more behaviour-based insurance in the future.[43]

The distinction between data-intensive underwriting and behaviour-based insurance is not clear-cut. For example, insurers could combine the two trends.[44] One can think of borderline cases, with characteristics of both practices. For example, insurers can combine real-time behavioural analysis with further data analysis of the driving habits of young drivers over a certain period.[45] Nevertheless, because of their different effects on the pricing of insurance, we discuss the two trends separately in the paper.

40 Bank of England & Financial Conduct Authority (FCA) 2022/5.3.

41 A rare example of a behaviour-based insurance product in The Netherlands is 'ANWB Veilig Rijden': https://www.anwb.nl/verzekeringen/autoverzekering/veilig-rijden.

42 Infantino, *European Review of Private Law* 2022/4, p. 623.

43 In 2014, Tim O'Reilly, founder and CEO of O'Reilly Media, said: 'I think that insurance is going to be the native business model for the Internet of Things'. R. Myslewski, 'The Internet of Things helps insurance firms reward, punish', https://www.theregister.com/2014/05/23/the_internet_of_things_helps_insurance_firms_reward_punish/.

44 In her pop-sci book, O'Neil warns for insurers' feeding behavioural data into AI systems. See O'Neil 2016, p. 139.

45 Henckaerts & Antonio, *Insurance: Mathematics and Economics* 2022/105, p. 79. Behaviour combined with contextual information, see Moosavi & Ramnath, *Accident Analysis & Prevention* 2023/192. Certain practices, such as credit scoring may also fall between the two trends. See e.g. The Geneva Association 2022, p. 7-8.

2.5 The two trends: similarities and differences

Data-intensive underwriting and behaviour-based insurance are both made possible, in part, by developments in AI. We highlight two differences between traditional insurance and the two trends.

i Traditional insurance is group-based. Assuming that insurers use AI to make more accurate predictions by analysing more data, insurers will assign fewer people with the same attributes to the same risk pool. In other words, risk pools become smaller, but the underlying system is still group-based. With behaviour-based insurance, insurers can follow the concrete behaviour of *individuals*, and adapt insurance based on their real-time behaviour.[46]

ii In general, insurers focus on paying compensation for risks that unfold. Some small exceptions exist where insurers aim to prevent risks, such as an insurer giving a discount on home insurance if the consumer installs certified locks. Data-intensive underwriting still follows the traditional idea of insurance: compensating for claims based on group predictions. By contrast, with behaviour-based insurance, insurers focus more on preventing risks. For example, insurers could (at least in theory) prevent accidents by nudging risky drivers to drive safer.

In Table 1, we summarise the main differences between traditional insurance, data-intensive underwriting, and behaviour-based insurance.

46 The discounts the insurer gives are not based on risk pooling, and this can cause tensions in the risk pooling of insurance. See Cevolini & Esposito, *Valuation Studies* 2022/9, p. 130-131.

Table 1: Comparison of traditional insurance, data-intensive underwriting, and behaviour-based insurance.

	Traditional insurance	**Data-intensive underwriting**	**Behaviour-based insurance**
Group / individual	group characteristics	(more) group characteristics	individual behaviour
Focus on compensation / prevention	compensation	compensation	prevention

3 Non-discrimination and (other) unfair differentiation

In this paper, we refer to discrimination in the legal sense, which is narrower than in common English. Discrimination occurs only when an insurer differentiates between groups based on a legally protected characteristics, such as ethnicity or gender.[47] Such discrimination of protected groups can occur accidentally. Many non-discrimination statutes around the world focus on a limited list of protected characteristics. For example, the European Union's non-discrimination directives together prohibit discrimination for six protected characteristics: age; disability; gender; religion or belief; racial or ethnic origin; sexual orientation.[48] But some of the EU non-discrimination

47 In general, ethnicity is the 'most protected' characteristic in the EU non-discrimination directives. Watson writes: '[T]here appears to be a hierarchy of norms with gender and race discrimination given considerably more protection than discrimination on other grounds. This in itself leads to inequalities.' Ellis & Watson 2012, p. 497. Indirect discrimination was introduced by the CJEU: 'the concept of indirect discrimination, and indispensable tool in the fight against discrimination, was developed by the Court in some of its earliest case law on sex equality and has now been written into all the equality directives'. Ellis & Watson 2012, p. 497.

48 Religion or belief, disability, age or sexual orientation: Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. Racial or ethnic origin: Council Directive 2000/43/EC Implementing the Principle of Equal Treatment Between Persons Irrespective of Racial or Ethnic Origin, 2000 OJ L 180/22. Gender: Council Directive 2004/113/EC Implementing the principle of Equal Treatment between Men and Women in the Access to and Supply of Goods and

directives have a narrower scope, focusing on, for instance, the employment context.

For insurance, the EU non-discrimination directives only prohibit discrimination based on gender and ethnicity.[49] Member states of the EU must implement the directives into national law. Many member states extended the scope of the non-discrimination rules to other sectors, including the insurance sector.[50] We limit our discussion to gender and ethnicity, because those characteristics are specifically protected in EU-wide non-discrimination law.

Many forms of differentiation in insurance are not illegal under non-discrimination law, because they do not harm groups with protected characteristics. But differentiation between groups with other characteristics, that are not legally protected, may nevertheless feel unfair. We call that 'other unfair differentiation'.

Services, 2004 OJ L 373/37. Directive 2006/54/EC of the European Parliament and of the Council on the implementation of the Principle of Equal Opportunities and Equal Treatment of Men and Women in Matters of Employment and Occupation (Recast), 2006 OJ L 204/23.

49 Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Directive 2000/43/EC); Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services (Directive 2004/113/EC).

50 European Commission. Directorate General for Justice and Consumers & European network of legal experts in gender equality and non discrimination 2023, p. 67.

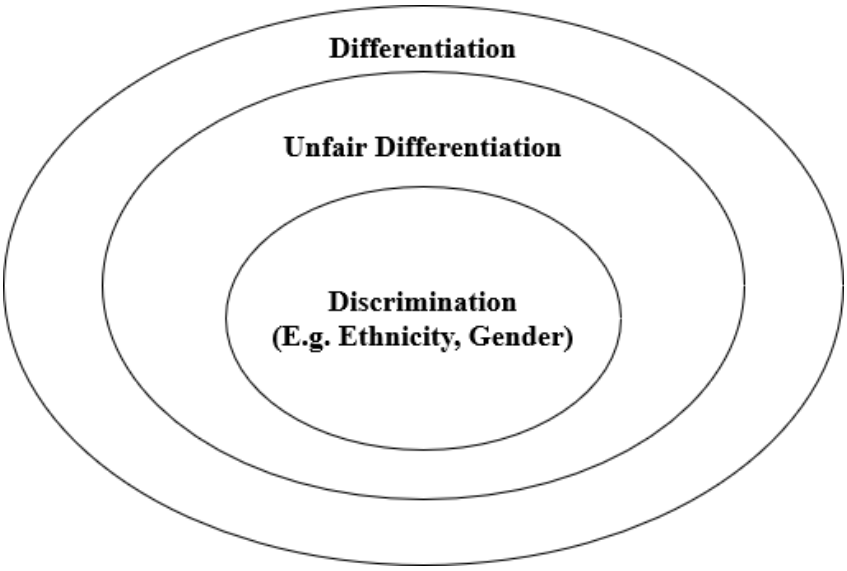


Figure 1: Discrimination and (unfair) differentiation. Discrimination is a type of differentiation that is considered unfair and can occur directly or indirectly. Unfair differentiation is a type of differentiation.

We see the relation between differentiation, other unfair differentiation, and discrimination as follows. With differentiation, we mean 'treatment of people or things in different ways'.[51] In that sense, discrimination is always a form of differentiation. Unfair differentiation is a form of differentiation. And discrimination in the legal sense is one type of unfair differentiation (Figure 1).[52]

51 Cambridge Dictionary, 'Definition of Differentiation' https://dictionary.cambridge.org/dictionary/english/differentiation, accessed 30 January 2025. The Geneva Association defines differentiation as 'any (lawful) differential treatment.' The Geneva Association 2022, p. 9.

52 We acknowledge that sometimes a decision may seem unfair to a consumer, but that does not make the decision as such unfair. Fairness and equality are explicit aims of non-discrimination law. Their aim is to give all individuals an equal and fair chance to access opportunities available in a society. See

4 Data-intensive underwriting

In this section, we discuss possible discriminatory effects of data-intensive underwriting (4.1) and possible unfair differentiation of data-intensive underwriting (4.2).

4.1 Discriminatory effects of data-intensive underwriting

From a legal perspective, insurers may differentiate based on any characteristic, subject to some exceptions, namely legally protected characteristics such as gender and ethnicity.[53] Insurance is largely governed by contractual freedom: in principle, insurers choose with who they want to enter a contract, for what price, and under which conditions.[54] Non-discrimination law limits the contractual freedom of the insurer. To illustrate, because of non-discrimination law, an insurer is not allowed to refuse to deal with non-white people. While insurers are allowed to differentiate, they are not allowed to discriminate in the legal sense.[55] Insurers typically have more leeway in private insurance than in public insurances. For example, for health insurance, contractual freedom typically plays a smaller role. Depending on the country, the law may require health insurers to accept high-risk consumers, such as people with pre-existing diseases.[56]

In EU law, two forms of discrimination are prohibited: *direct* and *indirect* discrimination.[57] US law contains a similar distinction between *disparate treatment* and *disparate impact*. *Direct discrimination* means that organisations discriminate against people on the basis of a protected characteristic, such as ethnicity or gender. An extreme example of direct discrimination is, for example, the differentiation based on ethnicity in life insurance in the US

on the principle of non-discrimination: https://eur-lex.europa.eu/EN/legal-content/glossary/non-discrimination-the-principle-of.html

53 As noted, data protection law is outside the scope of this paper.

54 Davey/Arvind & Steele 2020, p. 269.

55 See Section 3, where we noted that we mean discrimination in the narrow, legal sense. See also Avraham/Lippert-Rasmussen 2018, p. 335.

56 Public health insurance systems such as in, for example, The Netherlands and Germany are (partly) based on solidarity, not on risk calculations. See Van Manen/Wolf Sauter and others 2019, p. 360. See also Von Ulmenstein and others, *Frontiers in Artificial Intelligence* 2022/5, p. 4.

57 For more details, see Zuiderveen Borgesius and others 2024.

before the 1950s. Insurers included ethnicity in their mortality tables, arguing that African Americans died, on average, sooner. From the 1950s, insurers stopped using ethnicity.[58]

In the words of the Racial Equality Directive, 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 ethnic or ethnic origin'.[59] Direct discrimination is forbidden, except for a few specific and narrowly defined legal exceptions.[60] For the insurance sector, there are no exceptions in the non-discrimination directives that allow discriminating directly based on gender or ethnicity.

In the *Test-Achats* case from 2011, the Court of Justice of the European Union prohibited, roughly summarised, life insurers to discriminate directly based on gender. Until the judgment, many EU member states allowed gender-based price differentiation. EU non-discrimination law included a provision that allowed EU member states to adopt an exemption. The court declared that provision and the member states' exemptions invalid.[61] In sum, direct discrimination based on gender and ethnicity is prohibited for insurers.

Even if insurers underwrite increasingly data-intensively, it may seem difficult to imagine an insurer discriminating directly based on ethnicity or gender. Yet perhaps unintentionally, such direct discrimination can occur. In 2013, in the Netherlands, an insurer excluded people living in caravans from home insurance. The non-discrimination authority decided that differentiating based on whether people were living in caravans constitutes direct discrimination, because all Roma were affected by the decision to exclude caravans.[62] The Dutch insurers maybe did not expect that the differentiation based on 'living in a caravan' amounts to direct ethnicity discrimination, because the insurers were using a proxy variable. While in general we do not think that data-intensive underwriting leads to more direct discrimination, it can still occur unintentionally or unexpectedly.[63]

58 Horan 2011, p. 157-158, 186. Bouk 2015, p. 32-35, 199-205.

59 Article 2(2) of the Racial Equality Directive.

60 Zuiderveen Borgesius, *European Business Law Review* 2020/31, p. 408-409.

61 CJEU 1 March 2011, C-236/09 (*Test-Achats*).

62 College voor de Rechten van de Mens 2 September 2018, 2018-14 (*Reaal Schadeverzekeringen N.V.*). See also CJEU 12 December 2013, ECLI:EU:C:2013:823 (*Hay*), para. 44.

63 We note that in Europe, intentionality is not a requirement for direct discrimination.

More likely is that an insurer might discriminate indirectly. Indirect discrimination occurs when a practice is neutral at first glance but ends up discriminating against people with a certain protected characteristic. For example, suppose that a neighbourhood mostly includes people of a certain ethnicity. If an insurer charges higher prices in that neighbourhood, the insurer could accidentally discriminate against that ethnicity. Even if the insurer was unaware of the indirect discrimination, it is in principle prohibited.

However, if the insurer can invoke an 'objective justification', insurers do not indirectly discriminate. According to the non-discrimination directives, insurers can objectively justify indirect discrimination if they have 'a legitimate aim and [if] the means of achieving that aim are appropriate and necessary'.[64] An example of a legitimate aim in practice is if insurers try to lower their costs and reduce the administrative burden on their consumers.[65]

Judges seem to accept many aims as legitimate. But having a legitimate aim is not sufficient; the differentiation must be 'appropriate and necessary'. It depends on all circumstances of a case whether a practice is 'appropriate and necessary'. Generally speaking, the insurer must choose the least intrusive form of differentiation, and must ensure that a practice does not cause disproportionate disadvantages for protected groups.[66]

In data-intensive underwriting, insurers look for more data to further differentiate their risk pools. The threat of indirect discrimination may increase, especially if the insurers are not careful of which proxies they choose.

4.2 Other unfair differentiation of data-intensive underwriting

Which differentiation by insurers is fair and unfair? The line is difficult to draw. There is no uniformly accepted definition of fairness in the insurance

64 Article 2(2)(b) Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin (Directive 2000/43/EC).

65 College voor de Rechten van de Mens 2014.

66 CJEU March 2010, ECLI:EU:C:2010:127 (*ERG and others*), para. 86.

context.[67] The fairness of differentiation depends on all the circumstances of a situation.[68] In the field of AI, many technical definitions of fairness exist, also with sometimes conflicting requirements. Different contexts may require balancing different notions of fairness.[69]

In this section, we analyse data-intensive underwriting according to four main factors that may indicate unfair differentiation:

1. consumers have no influence on the price
2. insurers use characteristics that seem irrelevant for the consumer
3. insurers reinforce financial inequality
4. consumers are excluded from insurance

We base the four factors by combining various strands of literature. We do not argue that each factor always leads to unfairness. And in some situations, it may be unclear which factor carries more weight.[70] An insurance differentiation can be, or feel, unfair for consumers, without any ill intent by the insurer.

4.2.1 No influence on the price

In general, a risk assessment is fairer if a consumer can at least influence the price. It can feel unfair for consumers if they pay a higher price, and they cannot do anything about it.[71] From the perspective of a consumer, taking a risk can be either a gamble or a choice. In a chapter about luck and insurance,

67 Avraham/Lippert-Rasmussen 2018, p. 335. D. Minty, 'Why Equality of Fairness will shape the Future of Insurance', https://www.ethicsandinsurance.info/equality-of-fairness/, 3 March 2021.

68 Frezal & Barry, *Journal of Business Ethics* 2020/167, p. 134. Liukko, *New Genetics and Society* 2010/29, p. 457, 471. Swedloff, *61 B.C. L. REV.* 2020/2031.

69 See e.g. Barocas, Hardt & Narayanan 2023. See also Fröhlich & Williamson 2024, p. 415.

70 Different types of insurance may also require a different weighing of these factors. For example, public insurances versus private insurances.

71 See Barry & Charpentier, *Ethics and Information Technology* 2023/25, p. 49. This is also called 'luck egalitarianism'. See Fröhlich & Williamson 2024, p. 411.

Dworkin distinguishes between two types of luck: 'brute luck' and 'option luck':

> Option luck is a matter of how deliberate and calculated gambles turn out – whether someone gains or loses through accepting an isolated risk he or she should have anticipated and might have declined. Brute luck is a matter of how risks fall out that are not in that sense deliberate gambles. If I buy a stock on the exchange that rises, then my option luck is good. If I am hit by a falling meteorite whose course could not have been predicted, then my bad luck is brute [...].[72]

Data-intensive underwriting may lead to more situations in which insurers base the price on factors that the consumer cannot reasonably influence, leading to a more brute luck for consumers. After all, insurers (or their AI systems) may use more and more types of data to predict claim costs and to set prices. For example, in 2022, a Dutch consumer organisation found that some insurers charge different car insurance prices based on the consumer's house number. A consumer paid more if their house number included a letter: for instance, 11B or 39A.[73] Consumers may feel that they cannot easily influence their house number. Moving to another house to influence one's insurance price is hardly a serious option. A similar argument could be made for postal codes. Consumers can influence other characteristics more easily, such as the type of car they buy.

4.2.2 Seemingly irrelevant characteristics

A strength of machine learning systems is that they can find new correlations in datasets. A weakness is that such systems do not help insurers in understanding the relationships in the data and explaining them. Barocas and others write that 'A major limitation of machine learning is that it only reveals correlations, but [people] often use its predictions as if they reveal causation'.[74]

72 Dworkin 2000, p. 73.

73 Consumentenbond, 'Premies verzekeringen verschillen tot op huisnummer ['Insurance premiums differ by house number']', https://www.consumentenbond.nl/inboedelverzekering/verzekeringspremies-verschillen-tot-op-huisnummer, 2015.

74 Barocas, Hardt & Narayanan 2023, p. 13.

When insurers underwrite using more and more data, they might find many correlations in datasets that look weird. For example, a correlation between the claim cost of car insurance and even house numbers, being born in a certain month, or 'people who spend more than 50% of their days on streets starting with the letter J'.[75]

A possible threat of data-intensive underwriting is that insurers could apply new, opaque relationships for predicting the claim cost, even if they do not (fully) understand the relationship behind the correlation.[76] Some insurance companies may use such predictions if they increase profit.

Other insurers may have less faith in machine learning: They may only apply a certain relationship when they understand the causality behind the relationship.[77] Even if an insurer understands the causal relationship, however, consumers may still be confronted with a price based on characteristics that are – from the consumer's point of view – irrelevant.

We give an example. Suppose that a car insurer finds a correlation between a specific colour of car and the claim cost. It turns out that the colour of the car matches with the colour of many houses in certain cities, and is therefore more difficult for people to see in specific light conditions. The insurer sees people with the specific car colour as a separate risk pool for car insurance, and charges a higher price to consumers who drive a car with that colour.

While the insurer found an explanation for the relationship that is causally true, consumers may still find the characteristic irrelevant (or illogical), because it is difficult to understand how a specific car colour might create more accidents in specific situations. When requesting insurance, a consumer typically enters their car colour, and then sees the new price, without further explanation. We draw the example in Figure 2.

75 O'Neil 2016, p. 138.

76 For more information about causality, see Charpentier 2022, p. 57.

77 See Sections 2.2 and 2.3.

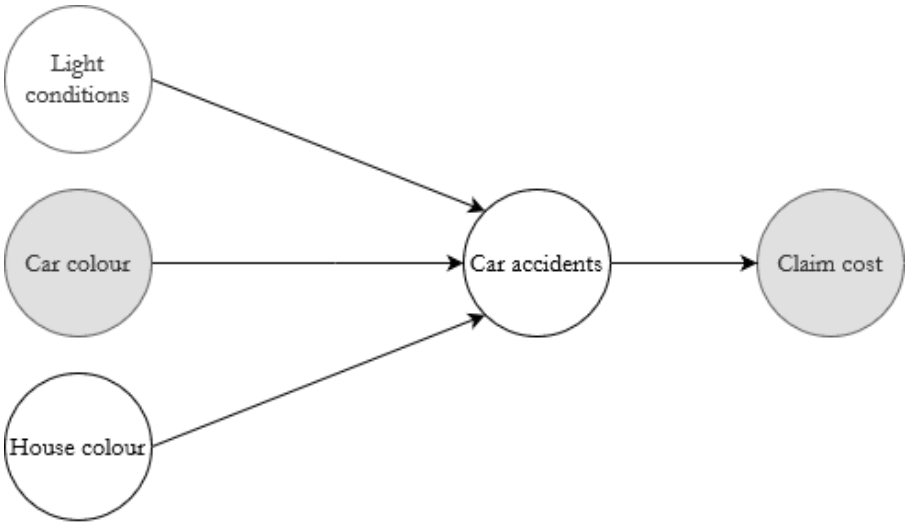


Figure 2: The arrows indicate the direction of the causal relationship. The grey nodes indicate what information about the specific consumer the insurer currently has: 'car colour' and 'claim cost'. The insurer predicts that light conditions, car colour and house colour cause more car accidents, which causes a higher claim cost. The consumer only sees that entering their car colour results in a higher claim cost.

In sum, we see two possible situations in which insurers use characteristics to calculate prices, while the characteristics seem strange or irrelevant for consumers. First, the insurer may not understand the causality behind a correlation, but still use the correlation to set prices. Second, the insurer understands the causal relationship, but the consumer still does not see the logic behind the relationship. As insurers collect and analyse more data for their underwriting, both types may occur.

4.2.3 Reinforcing financial inequalities

Insurance practices could reinforce inequalities that already exist in society, such as the difference between rich and poor. Practices that reinforce financial inequality in society are controversial – some might say unfair. We give a

real-life example. A Dutch insurer offered life insurance over the internet. The insurer wanted to charge higher prices to people likely to die earlier, and therefore charged higher prices to poorer people. In the Netherlands, income correlates strongly with life expectancy: on average, poorer people die years earlier than richer people.[78] Income also correlates with postal code. Therefore, the insurer charged higher life insurance prices in poorer postal codes. Besides the postal code, the insurer also considered the consumer's age, smoking behaviour, and health information for the insurance price. The insurer contacted the national non-discrimination regulator to ask whether the law allowed its practice.

The regulator consulted experts who confirmed that both postal code and income are a good predictor of life expectancy. The regulator ruled that the pricing was allowed because income or financial status are not protected grounds in Dutch non-discrimination law.[79] Therefore, such price differentiation is allowed by law. If the poor pay more, a practice reinforces financial inequality.[80] Many see it as unfair if a practice reinforces financial inequality.

Could data-intensive underwriting reinforce financial inequality in society? As discussed, an insurer could charge, on purpose, higher prices to poor people, because poor people live shorter on average. Similar situations may arise with data-intensive underwriting. First, an insurer's AI systems might find a correlation between being poor and filing more claims on average. The insurer might decide to charge higher prices to poor people. EU non-discrimination law does not prohibit such differentiation, unless the differentiation also disproportionally affects people of, for example, a certain ethnicity.[81]

Second, data-intensive underwriting could also lead to more situations where insurers *unintentionally* differentiate based on income. Take the following

78 See, (in Dutch) Centraal Bureau voor de Statistiek [Dutch Central Bureau of Statistics], 'StatLine - Gezonde levensverwachting; inkomen en welvaart', https://opendata.cbs.nl/#/CBS/nl/dataset/85445NED/table?defaultview&dl=747D6, 2022.

79 College voor de Rechten van de Mens 2014, p. 16.

80 The phrase 'the poor pay more' comes from a book, see David Caplovitz 1963.

81 Some national non-discrimination laws do prohibit discriminating on financial status, but the prohibition is not EU-wide. See, in Dutch: Terlouw, *Nederlands Tijdschrift voor de Mensenrechten* 2022/4. See also Ganty & Benito Sanchez 2021.

hypothetical. An insurer's AI system finds a correlation that improves the prediction of expected claims. Let's say that the insurer finds a new group (risk pool), that is distinguished by a complicated set of characteristics.[82] The insurer does not fully understand the causality behind the AI prediction, but the prediction is accurate in practice. The insurer charges higher prices to that group. Unbeknownst to the insurer, that group is, on average, poor. Hence, an insurer might charge higher prices to a certain group, while the insurer does not realise that this harms poor people.[83]

We do not claim that the poor will always pay more because of data-intensive underwriting. There may be situations where data-intensive underwriting will lead to opposite: the rich paying more. Depending on the situation, data-intensive underwriting could reinforce or mitigate financial inequality.

4.2.4 Exclusion

Insurers could exclude consumers from insurance in several ways, intentionally or not. First, insurers can refuse to insure somebody. For instance, a car insurer can refuse to contract with high-risk drivers. Second, an insurer may charge certain high-risk groups higher prices. Suppose that an insurer divides the total population in increasingly smaller risk pools. At some point, the price could become so high for the riskiest consumers that the insurer excludes those consumers de facto, even if the insurer does not formally refuse to contract with them.[84] Third, an insurer could advertise that it specialises in products for a certain group. For example, the Dutch insurer *Promovendum* says in its advertising that it targets only the higher educated. In reality, the insurer does not refuse applications by lower educated consumers, especially for mandatory insurances.[85] However, due

82 AI systems may identify new groups in society that deserve protection. See e.g. Wachter 2022.

83 Financial status is not an explicit characteristic under EU non-discrimination law. See section 3.1.

84 Avraham/Lippert-Rasmussen 2018, p. 342.

85 See, in Dutch: Promovendum, 'Higher educated ['Hoger opgeleiden']', https://www.promovendum.nl/hoger-opgeleiden.

to targeted advertising, most lower educated consumers will not contact this particular insurer. Hence, de facto the insurer excludes certain groups.[86]

Data-intensive underwriting brings a threat that certain groups in society are excluded from insurance. Insurers may become better at predicting who is a very high-risk consumer. For very high-risk consumers, prices may become unaffordable. Suppose that a care insurer charges a very high price to the riskiest drivers (say 5% of the population). The insurance could become unaffordable for those drivers.

Many insurers realise that underwriting more data-intensively may lead to non-insurability of groups of consumers. To illustrate, the Dutch Association of Insurers developed a 'solidarity monitor' in response to the threat of non-insurability. The Association tries to assess whether groups are excluded by looking at, for example, the spread of the price between groups and the percentage of people who were denied insurance. The Association reports every year on their findings. So far, the Association did not find evidence that the threat materialised in the Netherlands.[87]

Some authors worry that more precise underwriting can undermine the solidarity that was traditionally part of insurance: risks would no longer be cross-subsidised between riskier and less risky people in society.[88] Whether it is unfair if people are excluded from insurance, or large price difference appear, depends on many factors. For instance, the importance of an insurance product in society is relevant. It makes a difference whether somebody is excluded from health insurance for themselves, or from health insurance for their pet. Another factor to consider is whether the consumer has an alternative. If one insurer excludes certain consumers, but they can go to plenty of competitors, the exclusion problem is mitigated. In sum, data-intensive underwriting could lead to certain groups being excluded from insurance.

86 In 2014, a report by several Dutch governmental advisory organs named the insurance as a possible sign of explicit segregation. See in Dutch: Sociaal en Cultureel Planbureau (SCP) & Wetenschappelijke Raad voor het Regeringsbeleid (WRR) 2014, p. 24-25.

87 Verbond van Verzekeraars 2022. EIOPA 2019.

88 Frezal & Barry, *Journal of Business Ethics* 2020/167, p. 134.

4.3 Conclusion

In this section, we discussed the trend of data-intensive underwriting and analysed the trend based on non-discrimination law and fairness considerations. Insurers may (unintentionally) discriminate indirectly more often. For example, an insurer could use a newly found correlation to set prices. Such a practice could harm certain ethnic groups, or other groups with legally protected characteristics, without the insurer realising it.

We found more possible negative effects of data-intensive underwriting. First, consumers may have less influence on the insurance price if insurers use new types of data for calculating prices. Second, insurers are more likely to choose characteristics that consumers find irrelevant, even if insurers find them relevant. Third, data-intensive underwriting may lead to more financial inequality in society. Fourth, insurers could exclude certain groups. Many of these effects could occur by mistake, rather than intentionally.[89] We repeat that we do not claim that all these effects will occur; we merely flag possible effects.

5 Behaviour-based insurance

Next, we discuss whether behaviour-based insurance may lead to more cases of discrimination (section 5.1) or other unfair differentiation (section 5.2).

5.1 Discriminatory effects of behaviour-based insurance

Can behaviour-based insurance lead to indirect discrimination of legally protected groups? If insurers set the price based on the consumer's behaviour, discrimination seems unlikely at first glance. For example, driving behaviour is not a proxy for ethnicity or gender. Therefore, an insurer does not discriminate directly if it sets a price based on driving behaviour.

There is also no reason to assume indirect discrimination: driving behaviour is a neutral criterion, and there is no evidence to suggest that certain ethnicities are disproportionally affected if insurers set prices based on driving behaviour. And even if an insurer's practice affects mostly certain

89 Minty argues that there should be more debate between insurers and consumers, to ensure that insurance is and remains fair. Minty 2023, p. 11, 15-16.

ethnicities, an insurer might objectively justify the indirect discrimination by stating that a consumer's driving behaviour is often the direct cause of an accident, that consumers can easily influence their driving behaviour, and that the behaviour is therefore relevant. If an insurer has an objective justification, the prima facie discrimination is not illegal. Nevertheless, in theory, ethnicity might correlate with driving behaviour. Hence, in theory, behaviour-based car insurance could lead to illegal discrimination.

Second, it is possible that, for instance, driving behaviour correlates with certain genders.[90] In Europe, insurers are not legally allowed to charge different car insurance prices directly based on a person's gender. However, if women generally drive more carefully than men, women will generally pay lower prices with behaviour-based car insurance. Such an indirect price difference seems justifiable. A man could drive more carefully, and thus receive the lower price too. A similar argument could be made for fitness trackers and discounts on insurance prices: irrespective of gender, men and women could walk more often. To sum up, behaviour-based insurance seems unlikely to lead to more cases of illegal discrimination.

5.2 Other unfair differentiation of behaviour-based insurance

In this section, we discuss possible unfair differentiation as a result of behaviour-based insurance. We discuss the same factors as for data-intensive underwriting: no influence on the price (5.2.1), irrelevant characteristics (5.2.2), reinforcing financial inequality (5.2.3) and exclusion (5.2.4).

5.2.1 No influence on the price

Will consumers see behaviour-based insurance as unfair because they cannot influence the price? That seems unlikely. For instance, consumers can influence the price, by driving more carefully or by walking more steps.[91] Consumers may therefore see behaviour-based insurance prices as fairer than the current insurance price, which are often based on characteristics that they have little influence on. On the other hand, consumers may not be able to control behaviour over a longer period. First, changing a habit can take a long

90 See also Infantino, *European Review of Private Law* 2022/4, p. 626.

91 See section 4.2.1 for the distinction between brute and option luck.

time, often months. It is therefore questionable how easy changing behaviour truly is.[92] To answer the question of whether consumers can influence driving behaviour, input from psychology and other disciplines is needed.[93] All in all, consumers probably feel that they can influence the price with behaviour-based insurance.

5.2.2 Seemingly irrelevant characteristics

Behaviour-based insurance is based on the idea that consumers can influence the price by behaving in a certain way. Therefore, it seems unlikely that consumers see the behaviour as an irrelevant characteristic. Consumers receive a discount if they have an active lifestyle (monitored by the insurer through a fitness tracker) or drive carefully (monitored through a device in the car). In these cases, the consumer probably sees the connection between their own behaviour and the odds that they file an insurance claim for accidents.[94]

Another question is whether, for example, the way people drive also affects the claims cost of that group: does more careful driving actually lower the claims cost for the insurer? A survey by Swiss Re from 2017 suggests that while safer drivers generally file fewer claims (–2% to –5%), their average claim cost per accident is higher (3%) than for drivers without behaviour-based insurance.[95] Nevertheless, we expect that most consumers think that their behaviour is relevant for the price of behaviour-based insurance.

5.2.3 Reinforcing financial inequality

With behaviour-based insurance, an insurer monitors a consumer's behaviour, and not their financial status. Hence, at first glance, it seems unlikely that behaviour-based insurance can reinforce financial inequality.

92 Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13, p. 79-80.

93 See also Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13, p. 80. For a review of social scientific perspectives on behaviour-based insurance, see Tanninen, *Big Data & Society* 2020/7.

94 We can see exceptional cases where a fitness tracker is not relevant for life expectancy, such as people in a wheelchair who, because of their disability, cannot walk.

95 Swiss Re 2017, p. 24 & 29.

Nevertheless, some forms of behaviour-based insurance can benefit the rich more than the poor. In such cases, behaviour-based insurance could reinforce financial inequality.

To give an example for health trackers, suppose that many richer and better-educated people have jobs with more freedom, such as the possibility to work from home. For somebody who works at home, it is easier to go for a run or to the gym during the day than, for example, a factory worker. Somebody working from home can catch up on work at night. If a life insurer gives discounts to people with a more active lifestyle, on average, richer people may get more discounts. Such a situation would reinforce financial inequality.

However, the contrary effect is also possible: perhaps office workers have a less active lifestyle than less well-paid people who work in, for example, warehouses, or who deliver the mail. On average, poorer people could have more active lifestyles. Hence, an insurer who gives discounts to people with a more active lifestyle would not reinforce financial inequality.

Could behaviour-based car insurance reinforce financial inequality? At first glance, one's driving behaviour does not seem to correlate with financial status. However, it might be the case that driving style (as measured by a box in a car) correlates with financial status. For example, perhaps people with lower-educated and lower-income jobs work night shifts more often, and are on average more tired than people with high-income jobs. People who are more tired may drive less carefully. In this case, driving less carefully would correlate with being poorer. Thus, behaviour-based insurance could reinforce financial inequality.

But the contrary effect also seems possible. Some well-paid jobs may make people extra tired and thus less careful drivers: think of partners in law firms who work too many hours. There is another unequal effect of behaviour-based insurance: richer people can afford to refuse it. Consumers with more money can forego a discount if they dislike the privacy interference of the insurer's tracking their behaviour. Such differences could reinforce inequality.

In sum, it is impossible to say, in general terms, whether behaviour-based insurance will reinforce or mitigate financial inequality. In some cases,

behaviour-based insurance might reinforce inequality; in other cases it might mitigate inequality.

5.2.4 Exclusion

On the one hand, behaviour-based insurance could lead to more *inclusion*. As the European Insurance and Occupational Pensions Authority (EIOPA), an EU supervisory body, puts it,

> a better understanding of the risks in combination with risk-mitigation services can improve financial inclusion for some high-risk consumers who previously could not access affordable coverage. Examples include young drivers using telematics devices and patients with diabetes using health wearable devices.[96]

Indeed, if behaviour-based car insurance did not exist, perhaps many younger people could not afford car insurance at all.[97] In the United Kingdom, behaviour-based insurance seems necessary for many young drivers to afford car insurance: the prices are unaffordable otherwise.[98]

Behaviour-based insurance may also exclude certain groups. Insurers using behaviour-based insurance might charge such high prices to some people that the insurance becomes unaffordable for them.[99] On average, younger car drivers have more accidents.[100] Suppose that younger people are poorer than older people. Basing insurance prices on monitoring people's driving style may lead to higher, and possibly unaffordable, prices for younger people. This problem is not new, however: without monitoring driving behaviour, many car insurers already charge higher prices to younger drivers.

96 European Insurance and Occupational Pensions Authority (EIOPA) 2021, p. 25 & 10.

97 Dutch Financial Authority (AFM) 2021, p. 21. H Rubin, Aas & Williams, *Journal of Information Technology Teaching Cases* 2023/13, p. 82-83.

98 M. Jenkin, ' "I was quoted £2,000 for annual cover": are young people being priced out of car insurance? - Which? News', https://www.which.co.uk/news/article/are-young-people-being-priced-out-of-car-insurance-aSHpn4t4mz4k, 1 October 2023. H Rubin, Aas & Williams, *Journal of Information Technology Teaching Cases* 2023/13, p. 82-83.

99 Barry & Charpentier, *Ethics and Information Technology* 2023/25, p. 49.

100 See European Commission 2021.

5.3 Conclusion and discussion

We conclude that generally, behaviour-based insurance seems unlikely to cause discrimination against certain ethnicities or genders. But it is possible that behaviour-based insurance could lead to unfair differentiation in some cases.

Consumers can often influence their short-term behaviour. Behaviour seems relevant for behaviour-based car and life insurance. Behaviour-based insurance could reinforce or mitigate financial inequality, depending on the situation. Finally, behaviour-based insurance could lead to more inclusion, but also exclusion: insurers could exclude some people by focusing on behaviour, because insurance prices become too high. But behaviour-based insurance might also enable other consumers to obtain insurance, for whom insurance is too expensive without the option.

Some commenters wonder whether behaviour-based insurance is a 'fair alternative' to traditional insurance.[101] Because behaviour-based insurance focuses on the individual's characteristics (influenceable behaviour) rather than group characteristics that the individual cannot influence, some say that behaviour-based insurance is *actuarially fairer*. Roughly speaking, actuarial fairness means that insurers treat similar risks in similar ways: the price that an individual pays corresponds to the actual risk. But actuarial fairness is not the only type of fairness: there are other aspects to consider.[102] Our analysis of other aspects of fairness (controllable behaviour, relevance, financial inequality, and exclusion) did not show that behaviour-based insurance is generally fairer than traditional insurance.

101 Barry and Charpentier write, in their interpretation: 'The 2011 European Gender Directive, which prohibits the use of gender in pricing can be seen as [...] an invitation from the judge to adopt algorithmic tools and behavioural data instead of broad classes.' Barry & Charpentier, *Ethics and Information Technology* 2023/25, p. 49. See also Rebert & Van Hoyweghen, *Tijdschrift voor Genderstudies* 2015/18, p. 413. See also Barry & Charpentier 2022/4.

102 We will not summarize the whole discussion in the paper. For more details, see e.g. Jha, *Journal of the American College of Radiology* 2012/9, p. 887.; Fröhlich & Williamson 2024, p. 411. D. Minty, 'Behavioural fairness is a serious risk to the future of insurance', https://www.ethicsandinsurance.info/behavioural-fairness/, 29 April 2020. McFall, *Economy and Society* 2019/48, p. 52. The Geneva Association 2020.

6 Research agenda

Based on the findings in our paper, we suggest possible directions for future research into discrimination and unfair differentiation by AI in insurance. We see many possibilities for exciting interdisciplinary research, for instance for actuaries, economists, ethicists, legal scholars, sociologists, and computer scientists.

For analytical purposes, we distinguished data-intensive underwriting and behaviour-based insurance. But the two practices cannot be neatly separated, as the practices can partly overlap. Future research could focus on borderline cases. For instance, insurers could collect data about individuals through behaviour-based insurance, and aggregate those data to build group profiles for data-intensive underwriting. Do insurers do this, and should that influence the normative analysis?

We used four factors to assess unfair differentiation, but we do not claim that they form a complete framework. We think of the factors as a starting point for further discussion and research. Future work could extend our list of factors to build a more comprehensive fairness framework for insurance.

Many of the effects we identified in the paper raise normative questions that could be researched more in-depth. For example: when is it acceptable if insurance practices reinforce financial inequality? Is it fair if people who are born clumsy pay higher insurance prices in behaviour-based car insurance? Is it fair if people bear the costs of traits or behaviour that they cannot control? Is behaviour-based insurance fairer than traditional insurance or data-intensive underwriting? Certain types of insurance call for a specific normative analysis, such as health insurance, where the right to healthcare should play a role. This paper did not focus on the privacy and surveillance aspects of behaviour-based insurance, which also deserve more attention. Some issues that arise in the context of AI may also need more discussion and debate where AI does not play a role, such as the fairness of excluding consumers from insurance.

Many of the topics in our paper deserve more empirical research. For example, will both trends identified in the paper continue? In the paper, we highlighted possible discriminatory and other unfair effects. Which effects

will materialise in practice, if any? And to what extent is behaviour- based insurance based on empirical evidence? Does monitoring driving behaviour actually predict accidents, and are the AI models insurers use to assess driving behaviour correct, or a form of 'snake oil' AI?[103] Further research could investigate whether behaviour-based insurance actually influences consumer behaviour. And what do consumers and insurers think? Consumer surveys and interviews could provide insights about their opinions about AI in insurance. How do insurers use AI in practice? Interviews with insurers could provide valuable insights.

This paper focuses on risk-based pricing, which plays a large role in insurance pricing. But insurers may also adapt prices based on other factors, such as the expected willingness to pay of the consumer or the consumer's credit rating. Insurers could use AI to predict the maximum price a consumer is willing to pay. An insurer could, for instance, charge higher prices to consumers who pay little attention to prices. Such non-risk-based pricing is often called price discrimination in economic literature.[104] Future work could analyse the interplay between the effects of risk-based pricing and price discrimination.

While we discussed on possible effects, we did not discuss how insurers and policymakers could avoid or mitigate the effects. Open legal questions include: to what extent can existing law, such as non-discrimination law, consumer protection law, data protection law, and the AI Act, protect consumers and society against unfair effects of AI in insurance? If legal protection leaves gaps, should the law be improved, and how?

Finally, there are exciting possibilities for AI and computer science research. Could insurers make or use AI systems that are more transparent and explainable? Could sector-specific fairness and non-discrimination norms be built into AI systems?

Some questions that arise in the insurance sector are important in other sectors too. Is it fair if banks refuse to give a consumer credit because of

103 According to Naranyanan, most snake oil AI is 'concentrated in predictive AI'. Narayanan, *Arthur Miller Lecture on Science and Ethics* 2019. Kaltheuner 2021. Narayanan & Kapoor 2024.

104 Zuiderveen Borgesius & Poort, *Journal of Consumer Policy* 2017/40.

seemingly irrelevant characteristics, because the bank's AI systems found some correlations? When is it acceptable if the use of AI increases financial inequality in society?

7 Summary and conclusion

Table 2: Summary of conclusions.

	Effect	***Data-intensive underwriting***	***Behaviour-based insurance***
Discrimination	*Direct or indirect discrimination*	Indirect discrimination could happen more often	Unlikely
Unfair Differentiation	*No influence on the price*	Could happen more often (more characteristics with less influence)	Unlikely (consumers can influence their behaviour)
	Seemingly irrelevant characteristics	Could happen more often (AI can find unexpected correlations)	Unlikely (behaviour seems relevant to consumer)
	Reinforcing financial inequality	Insurers could (unintentionally) reinforce financial inequality	Could reinforce but also mitigate financial inequality
	Exclusion	Could happen more often	Could happen more often

In this paper, we discussed potential effects related to discrimination and unfair differentiation of insurers' use of AI. The paper explored which effects may occur if insurers (i) analyse more and new types of data (data- intensive underwriting), and (ii) adapt the price based on the consumer's real-time

behaviour (behaviour-based insurance). Both trends are enabled by AI. We distinguished discrimination (which is prohibited by non-discrimination law) from other unfair differentiation. We defined discrimination as a form of legally forbidden differentiation consisting of direct and indirect discrimination. Unfair differentiation is a type of differentiation that seems unfair to the consumer.

We described many possible discrimination- or unfair differentiation-related effects of data-intensive underwriting and behaviour-based insurance. Table 2 gives a rough summary of the possible effects. We cannot conclude that data-intensive underwriting is fairer than behaviour-based insurance, or vice versa. Many aspects of data-intensive underwriting and behaviour-based insurance are unclear and deserve more research; we provided an interdisciplinary research agenda. One thing is clear, however: public debate is needed to decide which insurance practices we accept in our societies.

References of chapter II

Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019
Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB), *Artificial intelligence in the insurance sector. An exploratory study*, AFM & DNB 2019, https://www.afm.nl/~/profmedia/files/rapporten/2019/afm-dnb-verzekeringssector-ai-eng.pdf.

Avraham/Lippert-Rasmussen 2018
R. Avraham/K. Lippert-Rasmussen, 'Discrimination and insurance', in: *The Routledge Handbook of the Ethics of Discrimination*, New York 2018, p. 335.

Bank of England & Financial Conduct Authority (FCA) 2022
Bank of England & Financial Conduct Authority (FCA), *Machine Learning in UK Financial Services*, 2022, https://www.bankofengland.co.uk/Report/2022/machine-learning-in-uk-financial-services.

Barocas, Hardt & Narayanan 2023
S. Barocas, M. Hardt & A. Narayanan, *Fairness and Machine Learning*, MIT Press 2023, https://fairmlbook.org/.

Barry & Charpentier, *Big Data & Society* 2020/7
L. Barry & A. Charpentier, 'Personalization as a promise: Can Big Data change the practice of insurance?', *Big Data & Society* 2020/7, issue 1, p. 205395172093514, http://journals.sagepub.com/doi/10.1177/2053951720935143, doi:10.1177/2053951720935143.

Barry & Charpentier 2022
L. Barry & A. Charpentier, 'The Fairness of Machine Learning in Insurance: New Rags for an Old Man?', 2022, http://arxiv.org/abs/2205.08112.

Barry & Charpentier, *Ethics and Information Technology* 2023/25
L. Barry & A. Charpentier, 'Melting contestation: insurance fairness and machine learning', *Ethics and Information Technology* 2023/25, issue 4, p. 49, https://link.springer.com/10.1007/s10676-023-09720-y, doi:10.1007/s10676-023-09720-y.

Bednarz, Lewis & Sadowski, *Computer Law & Security Review* 2025/56
Z. Bednarz, K. Lewis & J. Sadowski, '"It's not personal, it's strictly business": Behavioural insurance and the impacts of non-personal

data on individuals, groups and societies', *Computer Law & Security Review* 2025/56, p. 106096, https://linkinghub.elsevier.com/retrieve/pii/S0267364924001614, doi:10.1016/j.clsr.2024.106096.

Bouk 2015
D. Bouk, *How our days became numbered*, Chicago and London: University of Chicago Press 2015.

Cevolini & Esposito, *Big Data & Society* 2020/7
A. Cevolini & E. Esposito, 'From pool to profile: Social consequences of algorithmic prediction in insurance', *Big Data & Society* 2020/7, issue 2, p. 205395172093922, http://journals.sagepub.com/doi/10.1177/2053951720939228, doi:10.1177/2053951720939228.

Cevolini & Esposito, *Valuation Studies* 2022/9
A. Cevolini & E. Esposito, 'From Actuarial to Behavioural Valuation. The impact of telematics on motor insurance', *Valuation Studies* 2022/9, issue 1, p. 109-139, https://valuationstudies.liu.se/article/view/3697, doi:10.3384/VS.2001-5992.2022.9.1.109-139.

Charpentier 2022
A. Charpentier, *Insurance: Discrimination, Biases & Fairness*, France: Institut Louis Bachelier 2022, https://www.institutlouisbachelier.org/en/pdf-reader/?pid=81252.

College voor de Rechten van de Mens 2014
College voor de Rechten van de Mens, *Advies aan Dazure B.V. over premiedifferentiatie op basis van postcode bij de Finvita overlijdensrisicoverzekering*, 2014, https://publicaties.mensenrechten.nl/publicatie/29c613ac-efab-4d2a-a26c-fe1cb2556407.

Davey/Arvind & Steele 2020
J. Davey/T. Arvind & J. Steele, 'Insurance and Price Regulation in the Digital Era', in: *Contract Law and the Legislature: Autonomy, Expectations, and the Making of Legal Doctrine*, Hart Publishing 2020, p. 269-294, https://eprints.soton.ac.uk/433153/, doi:10.5040/9781509926138.

David Caplovitz 1963
David Caplovitz, *The Poor Pay More: Consumer Practices of Low-Income Families*, New York: Free Press of Glencoe 1963.

Dutch Financial Authority (AFM) 2021
Dutch Financial Authority (AFM), *The personalisation of prices and conditions in the Insurance Sector. An exploratory study*, The Netherlands: AFM 2021, https://www.afm.nl/en/sector/actueel/2021/June/aandachtspunten-gepersonaliseerde-beprijzing.

Dworkin 2000
R. Dworkin, *Sovereign virtue. The theory and practice of equality.*, London, England: Harvard University Press 2000.

EIOPA 2019
EIOPA, *Big data analytics in motor and health insurance: a thematic review*, Luxembourg: Publications Office of the European Union 2019.

Ellis & Watson 2012
E. Ellis & P. Watson, *EU anti-discrimination law*, Oxford: Oxford University Press 2012.

European Commission 2021
European Commission, *Facts and Figures Young People*, European Road Safety Observatory. Brussels, European Commission, Directorate General for Transport 2021, https://road-safety.transport.ec. europa.eu/system/files/2022-01/F%26F_young_people_20211221.pdf.

European Commission. Directorate General for Justice and Consumers & European network of legal experts in gender equality and non discrimination 2023
European Commission. Directorate General for Justice and Consumers & European network of legal experts in gender equality and non discrimination, *A comparative analysis of non-discrimination law in Europe 2022: the 27 EU Member States, Albania, Iceland, Liechtenstein, Montenegro, North Macedonia, Norway, Serbia, Turkey and the United Kingdom compared.*, LU: Publications Office 2023, https://data.europa.eu/doi/10.2838/428042.

European Insurance and Occupational Pensions Authority (EIOPA) 2021
European Insurance and Occupational Pensions Authority (EIOPA), *Artificial Intelligence governance principles, towards ethical and trustworthy Artificial Intelligence in the European insurance sector: a report from EIOPA´s Consultative*

Expert Group on Digital Ethics in insurance., LU: Publications Office 2021, https://data.europa.eu/doi/10.2854/49874.

Frezal & Barry, *Journal of Business Ethics* 2020/167
S. Frezal & L. Barry, 'Fairness in Uncertainty: Some Limits and Misinterpretations of Actuarial Fairness', *Journal of Business Ethics* 2020/167, issue 1, p. 127-136, http://link.springer.com/10.1007/s10551-019-04171-2, doi:10.1007/s10551-019-04171-2.

Fröhlich & Williamson, *FAccT* 2024
C. Fröhlich & R.C. Williamson, 'Insights From Insurance for Fair Machine Learning', *FAccT* 2024, https://dl.acm.org/doi/10.1145/3630106.3658914, doi:10.1145/3630106.3658914.

Gandy 2016
O.H. Gandy, *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*, Routledge 2016.

Ganty & Benito Sanchez 2021
S. Ganty & J.C. Benito Sanchez, *Expanding the List of Protected Grounds within Anti-Discrimination Law in the EU*, Equinet 2021, https://equineteurope.org/publications/expanding-the-list-of-protected-grounds-within-anti-discrimination-law-in-the-eu-an-equinet-report/.

Greenland 2002
S. Greenland, 'Causality theory for policy uses of epidemiological measures', in: A.D. Lopez & Colin D. Mathers, *Summary Measures of Population Health*, 291-301: World Health Organization 2002.

H Rubin, Aas & Williams, *Journal of Information Technology Teaching Cases* 2023/13
T. H Rubin, T.H. Aas & J. Williams, 'Big data and data ownership rights: The case of car insurance', *Journal of Information Technology Teaching Cases* 2023/13, issue 1, p. 82-87, https://doi.org/10.1177/20438869221096859, doi:10.1177/20438869221096859.

Henckaerts & Antonio, *Insurance: Mathematics and Economics* 2022/105
R. Henckaerts & K. Antonio, 'The added value of dynamically updating motor insurance prices with telematics collected driving behavior data', *Insurance: Mathematics and Economics* 2022/105, p. 79-

95, https://linkinghub.elsevier.com/retrieve/pii/S0167668722000385, doi:10.1016/j.insmatheco.2022.03.011.

Henkel, Heck & Göretz/Meiselwitz 2018
M. Henkel, T. Heck & J. Göretz/G. Meiselwitz, 'Rewarding Fitness Tracking—The Communication and Promotion of Health Insurers' Bonus Programs and the Use of Self-tracking Data', in: *Social Computing and Social Media. Technologies and Analytics* (Lecture Notes in Computer Science), Cham: Springer International Publishing 2018, p. 28-49, https://link.springer.com/10.1007/978-3-319-91485-5_3, doi:10.1007/978-3-319-91485-5_3.

Horan 2011
C.D. Horan, *Actuarial age: insurance and the emergence of neoliberalism in the postwar United States*, Minnesota: University of Minnesota 2011.

Infantino, *European Review of Private Law* 2022/4
M. Infantino, 'Big Data Analytics, Insurtech and Consumer Contracts: A European Appraisal', *European Review of Private Law* 2022/4, p. 613-634.

Insurance Europe 2012
Insurance Europe, *How insurance works (booklet)*, Brussels: Insurance Europe 2012, https://insuranceeurope.eu/publications/729/how-insurance-works/.

Insurance Europe 2021a
Insurance Europe, *Risk-based underwriting. How the losses of the few are spread among the many*, 2021, https://www.insuranceeurope.eu/priorities/2473/risk-based-underwriting-incl-ec-beating-cancer-plan.

Insurance Europe 2021b
Insurance Europe, *Risk Pooling information sheet*, 2021, https://www.insuranceeurope.eu/downloads/risk-based-underwriting-pooling/Pooling.pdf.

Janiesch, Zschech & Heinrich, *Electronic Markets* 2021/31
C. Janiesch, P. Zschech & K. Heinrich, 'Machine learning and deep learning', *Electronic Markets* 2021/31, issue 3, p. 685-695, https://link.springer.com/10.1007/s12525-021-00475-2, doi:10.1007/s12525-021-00475-2.

Jesus, Brito & Duarte 2023
R.M. Jesus, M.A. Brito & D.N. Duarte, .

Jha, *Journal of the American College of Radiology* 2012/9
S. Jha, 'Punishing the Lemon: The Ethics of Actuarial Fairness', *Journal of the American College of Radiology* 2012/9, issue 12, p. 887-893, https://linkinghub.elsevier.com/retrieve/pii/S1546144012005418, doi:10.1016/j.jacr.2012.09.012.

Kaltheuner 2021
F. Kaltheuner, *Fake AI*, Manchester: Meatspace Press 2021, https://fakeaibook.com/.

Krüger & Ní Bhroin, *The Journal of Media Innovations* 2020/6
S. Krüger & N. Ní Bhroin, 'Vital signs: Innovations in self-tracking health insurance and social change', *The Journal of Media Innovations* 2020/6, issue 1, p. 93-108, https://journals.uio.no/TJMI/article/view/7836, doi:10.5617/jomi.7836.

Liukko, *New Genetics and Society* 2010/29
J. Liukko, 'Genetic discrimination, insurance, and solidarity: an analysis of the argumentation for fair risk classification', *New Genetics and Society* 2010/29, issue 4, p. 457-475, https://www.tandfonline.com/doi/full/10.1080/14636778.2010.528197, doi:10.1080/14636778.2010.528197.

Van Manen/Wolf Sauter and others 2019
J. van Manen/Wolf Sauter and others, '15. Country report: the Netherlands', in: *The Law and Policy of Healthcare Financing*, Cheltenham: Edward Elgar Publishing Limited 2019, https://doi.org/10.4337/9781788115926.

Martani, Shaw & Elger, *Swiss Medical Weekly* 2019
A. Martani, D. Shaw & B.S. Elger, 'Stay fit or get bit - ethical issues in sharing health data with insurers' apps', *Swiss Medical Weekly* 2019, https://smw.ch/index.php/smw/article/view/2641, doi:10.4414/smw.2019.20089.

McFall, *Economy and Society* 2019/48
L. McFall, 'Personalizing solidarity? The role of self-tracking in health insurance pricing', *Economy and Society* 2019/48, issue 1, p. 52-76, https://www.tandfonline.com/doi/full/10.1080/03085147.2019.1570707, doi:10.1080/03085147.2019.1570707.

McStay & Urquhart, *International Review of Law, Computers & Technology* 2022/36
A. McStay & L. Urquhart, 'In cars (are we really safest of all?): interior sensing and emotional opacity', *International Review of Law, Computers & Technology* 2022/36, issue 3, p. 470-493, https://www.tandfonline.com/doi/full/10.1080/13600869.2021.2009181, doi:10.1080/13600869.2021.2009181.

Meyers 2018
G. Meyers, *Behaviour-based Personalisation in Health Insurance: a Sociology of a not-yet Market*, KU Leuven 2018, https://lirias.kuleuven.be/2087689&lang=en.

Meyers & Hoyweghen, *Big Data & Society* 2020/7
G. Meyers & I.V. Hoyweghen, ' "Happy failures": Experimentation with behaviour-based personalisation in car insurance', *Big Data & Society* 2020/7, issue 1, p. 205395172091465, http://journals.sagepub.com/doi/10.1177/2053951720914650, doi:10.1177/2053951720914650.

Minty 2023
D. Minty, *Revolutionising fairness to enable digital insurance*, Institute and Faculty of Actuaries 2023, https://actuaries.org.uk/media/z5rh5jhl/revolutionising-fairness-to-enable-digital-insurance.pdf.

Moosavi & Ramnath, *Accident Analysis & Prevention* 2023/192
S. Moosavi & R. Ramnath, 'Context-aware driver risk prediction with telematics data', *Accident Analysis & Prevention* 2023/192, p. 107269, https://linkinghub.elsevier.com/retrieve/pii/S0001457523003160, doi:10.1016/j.aap.2023.107269.

Narayanan, *Arthur Miller Lecture on Science and Ethics* 2019
A. Narayanan, 'How to recognize AI snake oil', *Arthur Miller Lecture on Science and Ethics* 2019, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf.

Narayanan & Kapoor 2024
A. Narayanan & S. Kapoor, *AI Snake Oil. What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton University Press 2024, https://www.degruyter.com/document/isbn/9780691249643/html.

O'Neil 2016
C. O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, New York: Penguin Books 2016.

Piesse 2023
D. Piesse, *Embedded Artificial Intelligence (AI) in Financial Services | International Insurance Society*, International Insurance Society 2023, https://www.internationalinsurance.org/insights_cyber_embedded_artificial_intelligence_in_financial_services.

Prince & Schwarcz, *IOWA LAW REVIEW* 2019/105
A.E.R. Prince & D. Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data', *IOWA LAW REVIEW* 2019/105, p. 62.

Rebert & Van Hoyweghen, *Tijdschrift voor Genderstudies* 2015/18
L. Rebert & I. Van Hoyweghen, 'The right to underwrite gender: The Goods & Services Directive and the politics of insurance pricing', *Tijdschrift voor Genderstudies* 2015/18, issue 4, p. 413-431, https://www.aup-online.com/content/journals/10.5117/TVGN2015.4.REBE, doi:10.5117/TVGN2015.4.REBE.

Sociaal en Cultureel Planbureau (SCP) & Wetenschappelijke Raad voor het Regeringsbeleid (WRR) 2014
Sociaal en Cultureel Planbureau (SCP) & Wetenschappelijke Raad voor het Regeringsbeleid (WRR), *Separate worlds? An exploration of socio-cultural divisions in the Netherlands ['Gescheiden werelden? Een verkenning van sociaal-culturele tegenstellingen in Nederland']*, WRR 2014, https://www.wrr.nl/publicaties/publicaties/2014/10/30/gescheiden-werelden-een-verkenning-van-sociaal-culturele-tegenstellingen-in-nederland.

Stroobants & Van Schoubroeck, *European Journal of Commercial Contract Law* 2021/13
N. Stroobants & C. Van Schoubroeck, 'Telematics Insurance: Legal Concerns and Challenges in the EU Insurance Market', *European Journal of Commercial Contract Law* 2021/13, issue 3, p. 51-81, https://www.ingentaconnect.com/content/10.7590/187714621X16463138640038, doi:10.7590/187714621X16463138640038.

Swedloff, *61 B.C. L. REV.* 2020/2031
R. Swedloff, 'The new regulatory imperative for insurance', *61 B.C. L. REV.* 2020/2031, p. 55.

Swiss Re 2017
Swiss Re, *Unveiling the full potential of telematics - How connected insurance brings value to insurers and consumers: An Italian case study*, Swiss Re 2017, https://www.researchgate.net/publication/329775370.

Tanninen, *Big Data & Society* 2020/7
M. Tanninen, 'Contested technology: Social scientific perspectives of behaviour-based insurance', *Big Data & Society* 2020/7, issue 2, p. 205395172094253, http://journals.sagepub.com/doi/10.1177/2053951720942536, doi:10.1177/2053951720942536.

Terlouw, *Nederlands Tijdschrift voor de Mensenrechten* 2022/4
A. Terlouw, 'CLASSISM: DISCRIMINATION BASED ON SOCIAL STATUS ['KLASSISME: DISCRIMINATIE OP GROND VAN SOCIALE STATUS']', *Nederlands Tijdschrift voor de Mensenrechten* 2022/4, https://njcm.nl/wp-content/uploads/2023/01/1.-NTM-47-4_Terlouw.docx.pdf.

The Geneva Association 2020
The Geneva Association, *Research Brief. Promoting Responsible Artificial Intelligence in Insurance*, The Geneva Association—International Association for the Study of Insurance Economics 2020, https://www.genevaassociation.org/sites/default/files/ai_in_insurance_web_0.pdf.

The Geneva Association 2022
The Geneva Association, *Responsible Use of Data in the Digital Age: Customer expectations and insurer responses*, 2022, https://www.genevaassociation.org/publication/new-technologies-and-data/responsible-use-data.

The Geneva Association (Noordhoek) 2023
The Geneva Association (Noordhoek), *Regulation of Artificial Intelligence in Insurance: Balancing consumer protection and innovation*, The Geneva Association 2023, https://www.genevaassociation.org/publication/public-policy-and-regulation/regulation-artificial-intelligence-insurance-balancing.

Thouvenin and others 2019
F. Thouvenin and others, 'Big Data in the insurance industry: Leeway and

limits for individualising insurance contracts', 2019, https://www.zora.uzh.ch/id/eprint/179171, doi:10.5167/UZH-179171.

Truby, Brown & Antoine Ibrahim, *International Review of Law, Computers & Technology* 2023, p. 1-25
J. Truby, R.D. Brown & I. Antoine Ibrahim, 'Regulatory options for vehicle telematics devices: balancing driver safety, data privacy and data security', *International Review of Law, Computers & Technology* 2023, p. 1-25, https://www.tandfonline.com/doi/full/10.1080/13600869.2023.2242671, doi:10.1080/13600869.2023.2242671.

Verbond van Verzekeraars 2022
Verbond van Verzekeraars, *Solidariteitsmonitor*, 2022, https://www.verzekeraars.nl/media/10617/solidariteitsmonitor-2022.pdf.

Von Ulmenstein and others, *Frontiers in Artificial Intelligence* 2022/5
U. Von Ulmenstein and others, 'Limiting medical certainties? Funding challenges for German and comparable public healthcare systems due to AI prediction and how to address them', *Frontiers in Artificial Intelligence* 2022/5, p. 913093, https://www.frontiersin.org/articles/10.3389/frai.2022.913093/full, doi:10.3389/frai.2022.913093.

Wachter 2022
S. Wachter, 'The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law', 2022, https://arxiv.org/abs/2205.01166, doi:10.2139/ssrn.4099100.

Walters 1981
M.A. Walters, *Risk classification standards*, Proceedings of the Casualty Actuarial Society 1981, https://www.casact.org/abstract/risk-classification-standards.

Zuiderveen Borgesius 2019
F. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making. Report for the European Commission against Racism and Intolerance (ECRI)*, Strasbourg: Council of Europe 2019, https://www.coe.int/en/web/artificial-intelligence/-/news-of-the-european-commission-against-racism-and-intolerance-ecri-.

Zuiderveen Borgesius, *European Business Law Review* 2020/31
F. Zuiderveen Borgesius, 'Price Discrimination, Algorithmic Decision-Making, and European Non-Discrimination Law', *European Business Law Review* 2020/31, issue 3, p. 22.

Zuiderveen Borgesius and others 2024
F. Zuiderveen Borgesius and others, *Non-discrimination law in Europe: a primer for non-lawyers [PREPRINT]*, 2024, https://arxiv.org/abs/2404.08519.

Zuiderveen Borgesius & Poort, *Journal of Consumer Policy* 2017/40
F. Zuiderveen Borgesius & J. Poort, 'Online Price Differentiation and EU Data Privacy Law', *Journal of Consumer Policy* 2017/40, issue 3.

CJEU March 2010, ECLI:EU:C:2010:127 (*ERG and others*).

CJEU 1 March 2011, C-236/09 (*Test-Achats*).

CJEU 12 December 2013, ECLI:EU:C:2013:823 (*Hay*).

College voor de Rechten van de Mens 9 February 2018, 2018-14 (*Reaal Schadeverzekeringen N.V.*).

III.

Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?

Submitted to *Computer Law & Security Review* as:

M. van Bekkum & F. Zuiderveen Borgesius, 'Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?', *Computer Law & Security Review 2023/48*, p. 105770, https://linkinghub.elsevier.com/retrieve/pii/S0267364922001133, doi:10.1016/j.clsr.2022.105770.

Both authors contributed equally to the paper. The authors thank Tom Heskes, Michael Veale, Raphaële Xenidis and Steven Bellovin for their insightful comments and discussion. The authors thank all the participants of the Privacy Law Scholars Conference and Digital Legal Talks for their active discussion.

Abstract

Organisations can use artificial intelligence to make decisions about people for a variety of reasons, for instance, to select the best candidates from many job applications. However, AI systems can have discriminatory effects when used for decision-making. To illustrate, an AI system could reject applications of people with a certain ethnicity, while the organisation did not plan such ethnicity discrimination. But in Europe, an organisation runs into a problem when it wants to assess whether its AI system accidentally discriminates based on ethnicity: the organisation may not know the applicants' ethnicity. In principle, the GDPR bans the use of certain 'special categories of data' (sometimes called 'sensitive data'), which include data on ethnicity, religion, and sexual preference. The proposal for an AI Act of the European Commission includes a provision that would enable organisations to use special categories of data for auditing their AI systems. This paper asks whether the GDPR's rules on special categories of personal data hinder the prevention of AI-driven discrimination. We argue that the GDPR does prohibit such use of special category data in many circumstances. We also map out the arguments for and against creating an exception to the GDPR's ban on using special categories of personal data, to enable preventing discrimination by AI systems. The paper discusses European law, but the paper can be relevant outside Europe too, as many policymakers in the world grapple with the tension between privacy and non-discrimination policy.

1 Introduction

Artificial intelligence (AI) can have discriminatory effects. Suppose that an organisation uses an AI system to decide which candidates to select from many job applications. The AI system rejects job applications from people with certain characteristics, say the educational courses taken by the job applicant. The organisation wants to prevent discrimination based on ethnicity or similar protected grounds.

To assess whether its AI system harms people with a certain ethnicity, the organisation needs to know the ethnicity of its job applicants. In Europe, however, the organisation may not know the ethnicity because, in principle, the GDPR prohibits the use of 'special categories of data' (article 9).[1] Special categories of data include data on ethnicity, religion, health, and sexual preferences. Hence it appears that the organisation is not allowed to infer, collect, or use the ethnicity of the applicants.

Our paper focuses on the following questions. (i) Do the GDPR's rules on special categories of personal data hinder the prevention of AI-driven discrimination? (ii) What are the arguments for and against creating an exception to the GDPR's ban on using special categories of personal data, to enable preventing discrimination by AI systems?

We use the word 'discrimination' mostly to refer to objectionable or illegal discrimination. Hence, we do not use 'discrimination' in a neutral sense. With 'preventing' AI-driven discrimination, we mean detecting and mitigating discrimination by AI systems.[2] An exception to the GDPR's ban on using special categories of data could be included in the GDPR, or in another statute, national or EU-wide.

We focus only on AI systems that have discriminatory effects relating to certain protected grounds in EU non-discrimination directives, namely

1 Also called "special category data" for short.

2 For more information about the causes of AI-driven discrimination, see Section 2 of this paper.

ethnicity, religion or belief, disability, and sexual orientation.[3] Hence, many types of unfair or discriminatory AI are outside the scope of this paper.

Our paper makes three contributions to the literature. First, we combine legal insights from non-discrimination scholarship on the one hand, and privacy and data protection scholarship on the other hand. We also consider insights from AI and other computer science scholarship. Most existing literature on using special category data for non-discrimination purposes is fragmented between disciplines and sub-disciplines.

Second, we show that, in most circumstances, the GDPR does not allow an organisation to use special categories of data for de-biasing. The GDPR allows the EU or national lawmakers to adopt an exception under certain conditions, but lawmakers have not adopted such an exception. We thus disagree with some non-discrimination scholars who appear to suggest that data protection law allows such data collection.[4] Third, we map out and analyse the arguments for and against introducing a new exception to the GDPR to enable de-biasing AI systems. We show that there are valid arguments for both viewpoints.

The paper can be relevant for computer scientists, legal scholars, and policymakers. This paper focuses on the law in Europe. However, the paper can be relevant outside Europe too. The problem of discriminatory AI is high on the agenda of policymakers worldwide.[5] In many countries, privacy

3 Racial or ethnic origin: Council Directive 2000/43/EC Implementing the Principle of Equal Treatment Between Persons Irrespective of Racial or Ethnic Origin, 2000 OJ L 180/22. Religion or belief, disability, age, or sexual orientation in the Employment context: Council Directive 2000/78/EC Establishing a General Framework for Equal Treatment in Employment and Occupation, 2000 OJ L 303/16. Gender in the context of the supply of goods and services: Council Directive 2004/113/EC Implementing the principle of Equal Treatment between Men and Women in the Access to and Supply of Goods and Services, 2004 OJ L 373/37. Gender in the employment context: Directive 2006/54/EC of the European Parliament and of the Council on the implementation of the Principle of Equal Opportunities and Equal Treatment of Men and Women in Matters of Employment and Occupation (Recast), 2006 OJ L 204/23. Age and gender are not special categories of data, although they are protected discrimination grounds. See Article 9(1) GDPR.

4 See Section 5.2.

5 See e.g. European Commission 2020. See also the EU Digital Services Act with requirements for certain types of online platforms to assess discrimination risks.

law and non-discrimination policy can be in conflict. We aim to describe European law in such a way that non-specialists can follow the discussion too.

The paper is structured as follows. In Sections 2 and 3 we introduce some key terms, summarise how AI can discriminate based on ethnicity and similar characteristics, and we introduce non-discrimination law. In Sections 4 and 5 we analyse the GDPR's framework for special categories of data and we discuss whether the framework hinders organisations in collecting special categories of data. In Section 6 we map out the arguments in favour of and against introducing a new exception to the collection and use of special category data for auditing AI systems. Section 7 discusses possible safeguards that could accompany a new exception, and Section 8 concludes.

2 AI systems and discrimination

AI systems could be described, in the words of the Oxford Dictionary, as 'computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.'[6] More technical and complicated definitions exist, which we will not discuss in this paper.[7] We focus on AI systems that make decisions that can have serious effects for people. For example, a bank could use an AI system to decide whether a customer gets

Articles 34(1)(b) and 40(4) of the 'Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act)' < https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065. In the US: e.g. the White house is planning an 'AI Bill of Rights'. White house employees announced a national effort to develop a "Bill of Rights for an AI-Powered World" in WIRED. See White House Office of Science and Technology Policy (OSTP), 'ICYMI: WIRED (Opinion): Americans Need a Bill of Rights for an AI-Powered World', https://www.whitehouse.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/, 22 October 2021. A "blueprint" for the AI bill of rights was released: OSTP 2022. See also OECD, 'OECD AI's live repository of over 260 AI strategies & policies', https://oecd.ai/en/dashboards.

6 'Artificial Intelligence' (Oxford Reference) https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095426960> accessed 12 October 2022.

7 See generally on defining AI: High-level expert group on Artificial Intelligence, High-level expert group on Artificial Intelligence 2019.

a mortgage or not. AI systems can help our society in many ways, and can also help make society fairer. In this paper, however, we focus on one specific risk of AI: the risk that AI systems discriminate against people with certain protected characteristics, such as ethnicity.

2.1 Causes of discrimination by AI

Discriminatory input data is one of the main sources of discrimination by AI systems.[8] To illustrate, suppose that an organisation's human resources (HR) personnel has discriminated against women in the past. Let's assume that the organisation does not realise that its HR personnel discriminated in the past. If the organisation uses the historical decisions by humans to train its AI system, the AI system could reproduce that discrimination. Reportedly, a recruitment system developed by Amazon ran into such a problem. Amazon abandoned the project before using it in real recruitment decisions.[9]

AI systems can make discriminatory decisions about job applicants, harming certain ethnicities for instance, even if the system does not have direct access to data about people's ethnicity. Imagine an AI system that considers the postal codes where job applicants live. The postal codes could correlate with someone's ethnicity. Hence, the system might reject all people with a certain ethnicity, even if the organisation has ensured that the system does not consider people's ethnicity. In practice, an AI system might also consider hundreds of variables, in complicated combinations, that turn out to correlate with ethnicity. Variables that correlate with protected attributes such as ethnicity can be called 'proxy attributes'. Such correlations can lead to discrimination by proxy.[10] Because of proxy attributes, AI systems can have discriminatory effects by accident: AI developers or organisations using AI

8 There are other possible causes. See for an overview of ways in which AI systems can have discriminatory effects: Barocas & Selbst, *California Law Review* 2016/104. Zuiderveen Borgesius 2019/III.2. A possible classification of biases was created by TILT. Tilburg Institute for Law, Technology, and Society 2021, p. 5.

9 A recruitment system used by Amazon until October 2018 seemed to display this type of bias. See Reuters, 'Amazon scraps secret AI recruiting tool that showed bias against women', https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 11 October 2018.

10 Barocas & Selbst, *California Law Review* 2016/104, p. 675 & 712. Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24, p. 183, 185.

systems may not realise that the AI system discriminates. More causes of discrimination by AI systems exist that we will not delve into further in this paper, ranging from design decisions to the context in which the system is used.[11]

2.2 Using special categories of data is useful to de-bias AI systems

Suppose that an organisation wants to test whether its AI system unfairly discriminates against job applicants with a certain ethnicity. To test this, the organisation must know the ethnicity of both the people who applied for the job, and of the people the organisation actually hired. Say that half of the people who sent in a job application letter has an immigrant background. The AI system selects the fifty best letters, out of the thousands of letters. The AI system decides based on attributes such as the school of choice, or courses followed. Of the fifty letters selected by the AI system, none is by somebody with an immigrant background. Such numbers suggest that the AI system should be investigated for unfair or illegal bias. Because proxy attributes may hide discrimination, the special categories of data are necessary for a detailed analysis.[12] In sum, collecting special categories of data (such as ethnicity data) is often useful, or even necessary, to audit AI systems for discrimination.[13]

3 Non-discrimination law

The right to non-discrimination is a human right. It is protected, for instance, in the European Convention on Human Rights (1950),[14] the International Convention on the Elimination of all Forms of Racial Discrimination (1965),[15] the International Covenant on Civil and Political Rights (1966),[16] and the Charter of Fundamental Rights of the European Union

11 See in-depth Barocas & Selbst, *California Law Review* 2016/104.
12 See more in-depth Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24, p. 190-193.
13 There are practical hurdles to testing an AI-system for discrimination. See the end of Section 6.3 of this paper.
14 Article 14 of the European Convention on Human Rights.
15 See in particular article 1-7 of the International Convention on the Elimination of all Forms of Racial Discrimination.
16 Article 2 and 26 of the International Covenant on Civil and Political Rights.

(2000).[17]

Here, we focus on European Union law. EU law forbids two forms of discrimination: direct and indirect discrimination. The Racial Equality Directive (about ethnicity) states that 'the principle of equal treatment shall mean that there shall be no direct or indirect discrimination based on racial or ethnic origin.'[18] The Racial Equality Directive describes *direct* discrimination as follows: 'Direct discrimination shall be taken to occur where one person is treated less favourably than another is, has been or would be treated in a comparable situation on grounds of racial or ethnic origin.'[19] A clear example of direct discrimination is the discrimination against Black South Africans by the Apartheid regime in South-Africa in the 20th century.[20]

Direct discrimination is prohibited in EU law. There are some specific, narrowly defined, exceptions to this prohibition. For example, the Racial Equality Directive allows a difference in treatment based on ethnicity if ethnicity 'constitutes a genuine and determining occupational requirement'.[21] A fitting example is the choice of a black actor to play the part of Othello.[22]

In non-discrimination scholarship, the grounds such as ethnicity ('racial or ethnic origin') are called 'protected characteristics' or 'protected attributes'. An AI system that treats individuals differently based on their protected attributes would discriminate directly. A hypothetical example of direct discrimination by a computer system is if the programmer explicitly makes the system reject all women.

Our paper, however, focuses on indirect discrimination. The Racial Equality

17 Article 21 of the Charter of Fundamental Rights of the European Union. See also article 19 of the Treaty on the Functioning of the European Union.

18 See Article 2 Council Directive 2000/43/EC Implementing the Principle of Equal Treatment Between Persons Irrespective of Racial or Ethnic Origin, 2000 OJ L 180/22.

19 Article 2(2)(a) Racial Equality Directive 2000/43/EC.

20 See for example Thomsen 2018, p. 19-29.

21 Article 4 Racial Equality Directive 2000/43/EC.

22 Ellis & Watson 2012, p. 382.

Directive defines indirect discrimination as follows.

> [I]ndirect discrimination shall be taken to occur where (i) an apparently neutral provision, criterion or practice would (ii) put persons of a racial or ethnic origin at a particular disadvantage compared with other persons, (iii) unless that provision, criterion or practice is objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary.[23]

In short, indirect discrimination by an AI system can occur if the system ('practice') is neutral at first glance but ends up discriminating against people with a protected characteristic. For example, even if the protected attributes in an AI system were filtered out, the system can still discriminate based on metrics that are a proxy for the protected attribute. Say that a recruitment system bases its decision on the set of courses, age, and origin university from the job applicant's CV. Such metrics could correlate with ethnicity or another protected attribute.

For both direct and indirect discrimination, it is irrelevant whether the organisation discriminates by accident or on purpose. Hence, an organisation is always liable, even if the organisation did not realise that its AI system was indirectly discriminating.[24]

Unlike for direct discrimination, for indirect discrimination there is an open-ended exception. Indirect discrimination is allowed if there is an objective justification. The possibility for justification is part of the definition of indirect discrimination: 'unless that provision, criterion or practice is objectively justified'.[25] In short, if the organisation (the alleged discriminator) has a

23 Article 2(2)(b) of the Racial Equality Directive 2000/43/EC. Capitalisation and punctuation amended by the authors.

24 For completeness' sake, we add that the difference between direct and indirect discrimination can be somewhat fuzzy. That discussion, however, falls outside the scope of this paper. See e.g. European Union Agency for Fundamental Rights, European Court of Human Rights, & Council of Europe (Strasbourg) 2018, p. 50. Adams-Prassl, Binns & Kelly-Lyth, *The Modern Law Review* 2022, p. 1468.

25 Article 2(2)(b) of the Racial Equality Directive 2000/43/EC. The CJEU says that 'the concept of objective justification must be interpreted strictly'. CJEU (Grand chamber) 16 July 2015, Case C-83/14, ECLI:EU:C:2015:480 (*Chez/Nikolova*), para. 112.

legitimate aim for its neutral practice, and that practice is a proportional way of aiming for that practice, there is no illegal indirect discrimination.[26] If an AI system has discriminatory effects, the general norms from EU non-discrimination law apply.

4 Data protection law

The right to privacy and the right to the protection of personal data are both fundamental rights. Privacy is protected, for instance, in the European Convention on Human Rights (1950),[27] the International Covenant on Civil and Political Rights (1966),[28] and the Charter of Fundamental Rights of the European Union (2000).[29]

Since the 1970s, a new field of law has been developed: data protection law. In the EU, the right to the protection of personal data has the status of a fundamental right.[30] The right to protection of personal data is explicitly protected in the Charter of Fundamental Rights of the European Union.[31] Data protection law grants rights to people whose data are being processed (data subjects), and imposes obligations on parties that process personal data (data controllers). Data protection law aims to protect personal data, and in doing so protects other values and rights. Unlike some seem to assume, data protection law does not only aim to protect privacy, but also aims to protect the right to non-discrimination and other rights.

5 Does the GDPR hinder the prevention of discrimination?

5.1 The GDPR's ban of processing special categories of data

With specific rules, the EU General Data Protection Regulation (GDPR) further works out the right to data protection from the EU Charter. The

26 See in more detail about on applying EU non-discrimination law to AI-driven discrimination: Zuiderveen Borgesius, *European Business Law Review* 2020/31.
27 Article 8 of the European Convention on Human Rights.
28 Article 17 of the International Covenant on Civil and Political Rights.
29 Article 7 of the Charter of Fundamental Rights of the European Union
30 See Granger & Irion 2018, p. 279-302.
31 Article 21 of the Charter of Fundamental Rights of the European Union.

GDPR, like its predecessor, the Data Protection Directive from 1995, contains an in-principle prohibition on using certain types of extra sensitive data, called 'special categories' of data.[32] Some older data protection instruments also included stricter rules for special categories of data. Data protection law's stricter regime for special categories of data can be explained, in part, by the wish to prevent unfair discrimination. In 1972, the Council of Europe said about sensitive personal data: 'In general, information relating to the intimate private life of persons or information which might lead to unfair discrimination should not be recorded or, if recorded, should not be disseminated'.[33] The Guidelines for the Regulation of Computerized Personal Data Files of the United Nations (1990) use the title 'principle of non-discrimination' for its provision on special categories of data.[34]

The GDPR also refers to the risk of discrimination in its preamble. Recital 71 concerns AI and calls upon organisations[35] to 'prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, (…) or that result in measures having such an effect.'[36] Article 9(1) GDPR is phrased as follows:

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

32 Article 8 of the Data Protection Directive.

33 33 Committee of Ministers, Resolution (73)22 on the protection of the privacy of individuals vis-à-vis electronic data banks in the private sector, 26 September 1973, article 1. https://rm.coe.int/1680502830.

34 Principle 5. https://www.refworld.org/docid/3ddcafaac.html.

35 The GDPR puts most responsibilities on the 'data controller', in short, the body which determines the purposes and means of the personal data processing (Article 4(7) GDPR). For ease of reading, we speak of the 'organisation' in this paper.

36 A hypothetical example might be as follows. An organisation uses AI to select the best candidates from a number of job application letters. Recital 71 reminds the organisation that it should prevent that its system discriminates unfairly, for instance on the basis of ethnicity.

Most protected grounds in EU non-discrimination directives are also special categories of data as defined in Article 9(1) GDPR. There are two exceptions. First, 'age' and 'gender' are protected characteristics in non-discrimination law, but are not special categories of data in the sense of the GDPR.[37] Second, 'political opinions', 'trade union membership', 'genetic' and 'biometric' data are special categories of data but are not protected by the European non-discrimination Directives.[38]

We summarize the distinction between the 'special categories of data' and 'protected non-discrimination grounds' in Fig. 1.

37 We add a caveat. In some circumstances, age and gender could be special categories of data if they were broadly interpreted under special categories 'health data', 'biometric data' or 'genetic data'. See Van den Brink & Dunne 2018, p. 9.

38 See European Union Agency for Fundamental Rights, European Court of Human Rights, & Council of Europe (Strasbourg) 2018/5.11. We add a caveat: under circumstances, these special categories of data may strongly correlate with protected characteristics, directly revealing them.

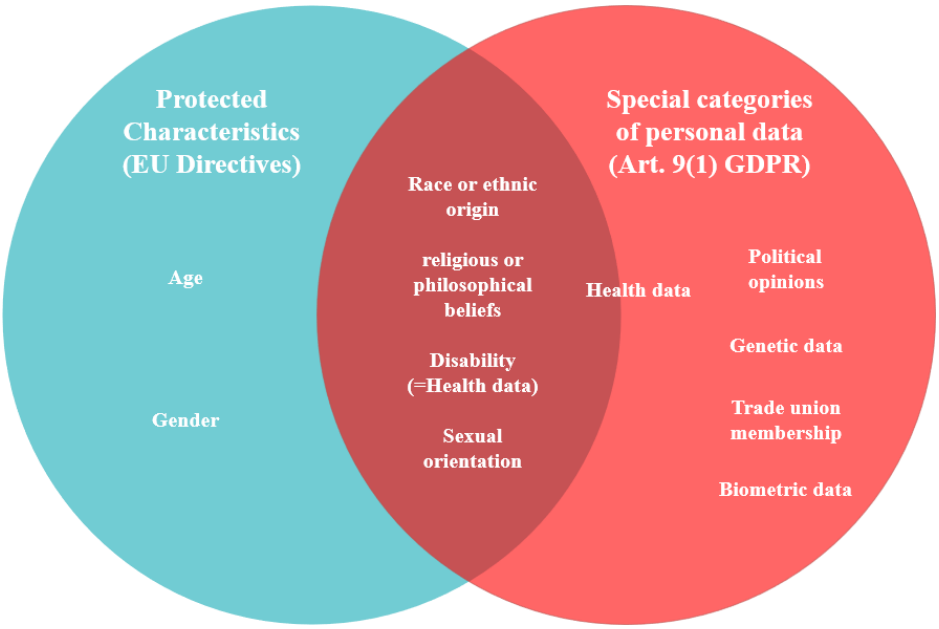


Figure 1: The overlap between protected characteristics and special categories of data.

The GDPR's prohibition to process special category can hinder the prevention of discrimination by AI systems.[39] Think about the following scenario. An organisation uses an AI-driven recruitment system to select the best applicants from many job applications. The organisation wants to check whether its AI system accidentally discriminates against certain ethnic groups. For such an audit, the organisation requires data concerning the ethnicity of the job applicants. Without such data, it's very difficult to do such an audit.

However, Article 9(1) of the GPPR prohibits using such ethnicity data. Article 9(1) not only includes explicit information about a data subject's ethnicity, but also information 'revealing' ethnicity. Hence, the organisation is not allowed

39 Alidadi, *European Equality Law review* 2017/2, p. 21-22. Al-Zubaidi, *European Equality Law review* 2020/2, p. 65.

to infer the ethnicity of its applicants either.[40] The GDPR contains exceptions to the ban; we discuss those in the next section.

5.2 The exceptions to the ban

Article 9(2) GDPR includes a list of exceptions to the general prohibition to process special category data. The subsections (a), (b), (f), (g) and (j) contain possibly relevant exceptions for collecting special categories of data. The exceptions concern (a) explicit consent, and specific exceptions for (b) social security and social protection law, (f) legal claims, (g) reasons of substantial public interest, and (j) research purposes. Georgieva & Kuner say that all these exceptions 'are to be interpreted restrictively'. [41]

Some non-discrimination scholars appear to suggest that data protection law does not hinder collecting or using special categories of data to fight discrimination.[42] One scholar suggests that there is a 'need to "myth bust" the notion that data protection legislation should preclude the processing of equality data.'[43] However, we did not find detailed argumentation in the literature for the view that the GDPR allows using special categories of data for non-discrimination purposes. Below we show that, in most circumstances, the GDPR does not allow the use of special category data to fight discrimination.

5.3 Explicit consent

We discuss each exception to the ban on processing special category data in turn, starting with the data subject's consent. The ban does not apply if '(a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes (...).'[44]

40 Kuner and others 2020/C.1.

41 Kuner and others 2020, p. C.3. 'The list of exceptions is exhaustive and all of them are to be interpreted restrictively'.

42 Farkas 2017, p. 14. Alidadi, *European Equality Law review* 2017/2, p. 21-22.

43 Alidadi, *European Equality Law review* 2017/2. Al-Zubaidi, *European Equality Law review* 2020/2, p. 22. See also Makkonen 2016, p. 27. Farkas 2017, p. 14.

44 Article 9(2)(a) GDPR.

In short, the data subject's explicit consent can lift the ban. However, the requirements for valid consent are very strict.[45] Valid consent requires that consent is 'specific' and 'informed', and requires an 'unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.'[46] Hence, opt-out systems cannot be used to obtain valid consent.

Moreover, Article 4(11) GDPR prescribes that consent must be 'freely given' to be valid. Valid consent thus requires that the consent is voluntary. The GDPR's preamble gives some guidance on interpreting the 'freely given' requirement: 'Consent should not be regarded as freely given if the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment.'[47]

Consent is less likely to be voluntary if there is an imbalance between the data subject and the controller. The preamble says that 'consent should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller'. If there is a 'clear' imbalance, consent is not freely given and thus invalid.'[48]

The European Data Protection Board says that consent from an employee to an employer is usually not valid:[49]

> An imbalance of power also occurs in the employment context. Given the dependency that results from the employer/employee relationship, it is unlikely that the data subject is able to deny his/her employer consent to data processing without experiencing the fear or real risk of detrimental effects as a result of a refusal.[50]

45 See article 4(11) and 7 GDPR.
46 Article 4(11) and 7 GDPR.
47 Recital 43 GDPR.
48 See on an imbalance between a data controller and a data subject also: Kosta 2011, p. 178-183.
49 The EDPB is an advisory body in which all European Data Protection Authorities cooperate. The body's opinions are considered persuasive in data protection law, carrying some weight.

The Board adds:

> [T]he EDPB deems it problematic for employers to process personal data of current or future employees on the basis of consent as it is unlikely to be freely given. For the majority of such data processing at work, the lawful basis cannot and should not be the consent of the employees (Article 6(1)(a)) due to the nature of the relationship between employer and employee.[51]

What does this mean for using special category data for preventing AI-driven discrimination? In many situations this 'freely given' requirement poses a problem. We return to our example: an organisation uses AI to select the best job applicants, and wants to audit its AI system for accidental discrimination. The organisation might consider asking all job applicants for consent to collect data about their ethnicity, to use that information for auditing its AI system. However, as discussed, applicants could fear that they would be rejected because of refusing to share their special category data. The applicant's consent is therefore generally invalid.

Perhaps a system could be designed in which a job applicant can give genuinely 'freely given' and thus valid consent. For instance, an organisation could ask all rejected job applicants for their consent after the position has been filled. In that case, job applicants might not fear anymore that withholding consent diminishes their chances to get the job. However, the organisation might find it awkward to ask people about their ethnicity, religion, or sexual preferences. Moreover, people can refuse their consent. If too many people refuse, the sample will not be representative – and cannot be used to audit AI systems.

We gave one specific example where gathering consent would be non-compliant. In many other contexts, particularly those where no power relationship exists between data subject and data controller, the data subject's

50 European Data Protection Board 2020, p. 9.
51 European Data Protection Board 2020, p. 9.

consent may provide an exception to collect data regarding e.g. ethnicity.[52] All in all, in certain circumstances, organisations could rely on consent to lift the ban. But in many circumstances, the data subject's consent seems not a real possibility.

5.4 Other exceptions

We continue with the other exceptions to the ban on processing special category data, starting with (b), on employment and social security and social protection law. The ban can be lifted if:

> (b) processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law (...) in so far as it is authorised by Union or Member State law or a collective agreement pursuant to Member State law.[53]

Exception (b) only applies to situations relating to the field of employment and social security and social protection law. An example is a provision of the Dutch implementation legislation of the GDPR.[54]

Roughly summarized, that provision allows employers to collect health data of employees if that is necessary for e.g. their re-integration after an illness. In a case about this provision, the Dutch Data Protection Authority stated that an employer must regularly re-consider if gathering the health data is

52 Makonnen writes: 'it is subparagraph a on the consent of the data subject which is likely to become the most frequently used basis for processing sensitive data.' Makkonen 2016, p. 28.

53 Article 9(2)(b) GDPR.

54 Art. 30 Uitvoeringswet Algemene Verordening Gegevensbescherming (UAVG): 'In view of Article 9, Section 2, under b, of [the GDPR], the prohibition on processing data concerning health does not apply if the processing is carried out by administrative bodies, pension funds, employers or institutions working on their behalf, and in so far as the processing is necessary for: (a) the proper execution of statutory provisions, pension schemes or collective agreements providing for entitlements which depend on the state of health of the person concerned; or (b) the reintegration or assistance of employees or beneficiaries in connection with illness or disability.' Translation by the authors of this paper.

still truly necessary in the light of the employee's reintegration duty. This requirement followed from a restrictive interpretation of Art. 9(2)(b) GDPR.[55]

Could this GDPR provision help organisations that want to audit their AI systems for discrimination? No. The main problem is that exception (b) only applies if the EU lawmaker or the national lawmaker has adopted a specific law that enables the use of special category data. To the best of our knowledge, no national lawmaker in the EU, nor the EU, has adopted a specific law that enables the use of special category data for auditing AI systems.[56] We turn to the next possibly relevant exception in the GDPR: the ban does not apply if

> (f) processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity.

Exception (f) applies to 'legal claims'; hence, claims in legal proceedings such as court cases. The exception applies to the use of personal data by courts themselves. An organisation might argue that it needs to audit its AI systems to prevent future lawsuits because of illegal discrimination. However, Kuner & Bygrave note that exception (f) is only valid for concrete court cases. The exception 'does not apply when sensitive data are processed merely in anticipation of a potential dispute without a formal claim having been asserted or filed, or without any indication that a formal claim is imminent.'[57]

It seems implausible that an organisation can use it to collect special categories of data about many people to audit its AI systems preventatively. Could exception (g) and (j) help organisations that want to de-bias their AI systems?

> (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law (...).

55 See in Dutch: Dutch Data Protection Authority 2020, p. 7.

56 A specific duty may be present in Finnish law. We do not know if this duty was intended for auditing AI systems. The duty exists for 'all authorities, all employers and all providers of education' to 'assess the realisation of equality in their functions'. Farkas further notes: 'Obligations to collect racial and ethnic data do not generally seem to be codified in law in the Member States.' See Farkas 2017, p. 15.

57 Kuner and others 2020/3, summation, under 6.

> (j) processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law (. . .).

The two provisions require a legal basis in a law of the EU or of a national lawmaker.[58] Hence, the provisions do not allow processing of special categories of data. Instead, exceptions (g) and (j) give flexibility to the EU or member states to decide whether processing special category data for fighting discrimination is allowed.[59] Current EU law, nor national law, provide such an exception.

To fight discrimination under exceptions (g) and (j), national law must provide a legal ground for processing. To the best of our knowledge, the United Kingdom was the only member state to have ever adopted an exception. (Now the UK is not a EU member anymore). The UK exception is based on Article 9(2)(g), the 'substantial public interest' exception.[60] However, the EU nor the current member states have adopted such a law. (The proposed AI Act contains such a provision; we discuss that in Section 7.1).

During the drafting of the GDPR, the Fundamental Rights Agency (FRA), a European Union agency, realised that the (proposed) GDPR did not allow such data collection. The Agency said that '[the GDPR] could clarify the place of special category categories in anti-discrimination data collection, and make explicit that the collection of special category data is allowed for the purpose of combatting discrimination'[61] But the EU did not follow that suggestion.

58 Personal data collection for equality and anti-discrimination purposes likely constitutes a substantial public interest. See Van Caeneghem 2019, p. 215, 217, 219, 242.

59 Van Caeneghem 2019, p. 218.

60 The UK data protection Act 2018 itself states 'substantial public interest' as the legal ground. See UK Data Protection Act 2018, Schedule 1 Part 2 https://www.legislation.gov.uk/ukpga/2018/12/schedule/1/part/2/crossheading/equality-of-opportunity-or-treatment. See also ICO, 'What are the substantial public interest conditions?', https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/special-category-data/what-are-the-substantial-public-interest-conditions/, 8 June 2021.

61 FRA 2013, p. 21-22.

For completeness' sake, we highlight that an organisation is not out of the woods yet if it has found a way to lift the ban on using special category data. For the processing of personal data, sensitive or not, the GDPR requires a 'legal ground'. Hence, even if the ban on using special category data could be lifted, the organisation must still find a valid legal processing ground as defined in Article 6(1) GDPR. For non-state actors, the legitimate interest ground (Article 6(1)(f) GDPR) seems the most plausible. The organisation must also comply with all the other requirements of the GDPR. But a discussion of all those GDPR requirements falls outside the scope of this paper.

In conclusion, the GDPR indeed hinders organisations who wish to use special category data to prevent discrimination by their AI systems. In some exceptional situations, an organisation might be able to obtain valid consent from data subjects for such use. In other situations, an EU or national law would be needed to enable the use of special categories of data for AI debiasing – at the moment such laws are not in force in the EU. In the next section, we explore whether such an exception is a good idea.

6 A new exception to the ban on using special categories of data?

6.1 Introduction

Policymakers have realised that non-discrimination policy can conflict with data protection law, and some have adopted exceptions. As noted, the UK has adopted an exception to the ban on using special categories of data, for the purpose of fighting discrimination.[62] The Dutch government is considering to create a new national exception for collecting special categories of data.[63]

62 UK Data Protection Act 2018, Schedule 1 Part 2 https://www.legislation.gov.uk/ukpga/2018/12/schedule/1/part/2/crossheading/equality-of-opportunity-or-treatment. See ICO website, ICO, 'What are the substantial public interest conditions?', https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/special-category-data/what-are-the-substantial-public-interest-conditions/, 8 June 2021. See also Open Data Institute 2020, p. 7.

63 See in Dutch answer to question 12 of https://zoek.officielebekendmakingen.nl/kst-26643-727.html and par. 2.2 of https://zoek.officielebekendmakingen.nl/kst-26643-

We are aware of 6 countries outside the EU with an exception in their national data protection act that enables the use of special category data for non-discrimination purposes. These countries are Bahrain, Curaçao, Ghana, Jersey, Sint Maarten, and South Africa.[64] The European Commission presented the AI Act proposal, with a possible new exception in Article 10(5) AI Act (see Section 7.1).

In the following section, we map out arguments in favour and against creating an exception for gathering special categories of data for the purpose of auditing an AI system for discrimination.

6.2 Arguments in favour of an exception

We present two main arguments in favour of creating a new exception that enables the use of special category data to prevent AI-driven discrimination: (i) Several types of organisations could use the data to test whether an AI system discriminates. (ii) The collection of the data has a symbolic function.

A first, rather strong, argument is that, for many types of stakeholders, collecting special category data would make the fight against discrimination easier. Organisations could check, themselves, whether their AI system accidentally discriminates. Organisations may want to ensure that their hiring, firing and other policies and practices comply with non-discrimination laws.

726.html. See p. 3, final paragraph of https://www.rijksoverheid.nl/documenten/kamerstukken/2020/12/04/tk-reactie-op-mededelingen-europese-commissie-over-de-avg

64 Bahrain Personal Data Protection Act of 2018, Article 5 http://www.legalaffairs.gov.bh/146182.aspx?cms=q8FmFJgiscJUAh5wTFxPQnjc67hw%2bcd53dCDU8XkwhyDqZn9xoYKj%2bwKjH8MwskD8zKV4oL8QNchAeJU7Z6zGg%3d%3d#.XAAo1NtKiUl. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Curacao National Ordinance Protection of Personal Data Chapter 2(2), https://media2.mofo.com/documents/Curacao-PRIVACY-ACT.pdf. Ghana Data Protection Act, 2012, https://www.dataprotection.org.gh/data-protection/data-protection-acts-2012. https://www.dataprotection.org.gh/data-protection/data-protection-acts-2012, Schedule 2 Pt. 1, Jersey https://www.jerseylaw.je/laws/enacted/PDFs/L-03-2018.pdf. South Africa Protection of Personal Information Act 2013 Ch. 3, http://www.saflii.org/za/legis/num_act/popia2013380.pdf. Saint Martin Personal Data Protection Act 2010 GT No. 2 Chapter 2 Par. 2, http://www.sintmaartengov.org/government/AZ/laws/AFKONDIGINGSBLAD/AB%2002%20Landsverordening%20bescherming%20persoonsgegevens.pdf.

Organisations may care about fairness and non-discrimination, or may want to protect their reputation.[65]

Regulators, such as equality bodies (non-discrimination authorities) could also benefit from an exception that enables the use of special categories of data for AI de-biasing. Regulators could more easily check an organisation's AI practices if those organisations registered the ethnicity of all their employees, job applicants, etc.[66]

Another group that can benefit from the collection of special category data is researchers.[67] Researchers could use such data to check whether an AI system discriminates. This argument is only valid, however, if an organisation shares its data with the researcher.

A different type of argument in favour of allowing the use of special category data is related to the symbolic function of such use.[68] Auditing an AI system can increase the trust in an organisation's AI practices, if it is publicly known that the organisation checks whether its AI systems discriminate. Potential discrimination victims can see that the organisation takes discrimination by AI seriously.[69] We mention this argument for completeness' sake. However, we do not see this as a particularly strong argument. In sum, there are various arguments in favour of adopting an exception that enables the use of special category data to prevent discrimination by AI.

65 See also Makkonen 2016, p. 21.

66 Makkonen: 'National specialised bodies, such as ombudsmen and equality bodies, and international monitoring bodies, such as the UN treaty bodies and the Council of Europe's European Commission against Racism and Intolerance (ECRI), as well as some other institutions, such as the EU Fundamental Rights Agency, need quantitative and qualitative information in order to perform their functions properly.' Makkonen 2016, p. 20.

67 Makkonen 2016, p. 21.

68 Similar to an argument made by Makonnen: '[...] the compilation of equality statistics can be seen to have more symbolic functions. The mere existence of a data collection system sends a message to actual and potential perpetrators, actual and potential victims and to society in general, signalling that society disapproves of discrimination, takes it seriously and is willing to take the steps necessary to fight it. This can have a preventive effect.' Makkonen 2016, p. 21.

69 See for a similar argument for collecting non-discrimination data in general Alidadi, *European Equality Law review* 2017/2, p. 18.

6.3 Arguments against an exception

There are also strong arguments against adopting an exception that enables the use of special category data for AI de-biasing. Balayn & Gürses warn for the danger of surveillance of protected groups: 'Policy that promotes de-biasing (…) may incentivise increased data collection of exactly those populations who may be vulnerable to surveillance.'[70]

We distinguish three categories of arguments against introducing a new exception to enable the use of special categories of data to mitigate discrimination risks. There are arguments (i) that concern the mere storage of special categories of data, (ii) that concern new uses of those data, and (iii) that show practical hurdles as a reason why an exception is not justified at this time.

We start with arguments that relate to the sole fact that an organisation stores special category data, regardless of how the data are used. People may feel uneasy if their data are collected or stored. People may have that feeling, regardless of whether the data are accessed by anyone. Many people are uncomfortable with organisations storing large amounts of personal data about them, even if no human ever looks at the data. Calo speaks of subjective harm, 'the perception of loss of control that results in fear or discomfort.'[71] Such harms could also be called 'expected harms'.[72]

The Court of Justice of the European Union accepts that the mere storage of personal data can interfere with the rights to privacy and to the protection of personal data.[73] The European Court of Human Rights has also said that storing sensitive personal data can interfere with the right to privacy, regardless of how those data are used: 'even the mere storing of data relating to the private life of an individual amounts to an interference within the

70 Balayn & Gürses 2021, p. 94.
71 Calo, *Indiana Law Journal*, p. 1143.
72 Gurses 2010, p. 87-89.
73 CJEU October 2013, C291/12 (*Schwartz/Stadt Bochum*), para. 25. 'as a general rule, any processing of personal data by a third party may constitute a threat to those rights'. See also the judgment CJEU 8 April 2014, C-293/12 and C-594 (*Digital Rights Ireland Ltd*), para. 29.

meaning of Article 8' of the European Convention on Human Rights, which protects the right to privacy.[74]

In sum, the two most important courts in Europe accept that merely storing personal data can interfere with fundamental rights. Indeed, we think that many people may dislike it when special category data about them (such as their ethnicity) are stored.

i) A second category of arguments concerns risks related to the storage of data. Data that have been collected can be used for many new purposes, that can harm both the individuals involved and society as a whole. These arguments relate to 'experienced harm' or 'objective harm'.[75]

For instance, a data breach can occur. Storing data always brings the risk of data breaches. Employees at an organisation or outsiders may gain unauthorised access to the data. A breached database with personal data about people's ethnicity, religion, or sexual preferences could have very negative effects. From a data security perspective, it is good policy not to develop databases with such sensitive data.

Second, there is a risk that data are used for new, unforeseen, uses. An extreme example of a new use of data about ethnicity and religion concerns the registration of Jewish people in the Netherlands, during the time that the Nazis occupied the Netherlands. The Nazis could easily find Jewish people in the Netherlands, because the citizen registration included the ethnicity and religion for each person. The IBM computers in the citizen registry made it easy to compile lists of, for instance all the Jewish people in Amsterdam.[76] There are many more examples where collecting population data facilitated human rights abuses.[77] Seltzer and Anderson write: 'In many

74 ECHR (Grand Chamber) 25 May 2021, 58170/13, 62322/14 and 24960/15 (*Big Brother Watch and others/United Kingdom*), para. 330. See also: ECHR 26 March 1987, 9248/81 (*Leander/Sweden*), para. 48. ECHR, ECHR 3 April 2007, 62617/00 (*Copland v. United Kingdom*), para. 43-44. ECHR 16 February 2000, 27798/95 (*Amann v. Switzerland*), para. 69. ECHR 4 December 2008, 30562/04 and 30566/04 (*S. and Marper v. United Kingdom*), p. 67, 121.

75 Calo describes objective harm as 'the unanticipated or coerced use of information concerning a person against that person.' Calo, *Indiana Law Journal*, p. 1133. Gurses 2010, p. 87-89.

76 E. Black 2012.

77 For 10 cases, see Table 1 of Seltzer & Anderson 2022.

large-scale human rights abuses, statistical outputs, systems, and methods are a necessary part of the effort to define, find, and attack an initially dispersed target population.'[78]

Third, organisations could misuse the exception to collect large amounts of special categories of data, claiming that they need such data to fight discrimination. An exception that is too wide could open the door for mass data gathering. As Balayn and Gürses note, 'the possibility that de-biasing methods may lead to over-surveillance of marginalised populations should be a very serious concern.'[79]

Fourth, a symbolic argument can be made against an exception that allows the collection and storage of special category data. People want their sensitive data to be handled with care. If people know that an organisation does not collect their special category data, they could trust that organisation more. (As with the symbolic argument in favour of using special category data, we do not think this argument is very strong.) In sum, there are several arguments against adopting an exception that enables using special categories of personal data to prevent discrimination by AI.

Finally, there is a different category of arguments against adopting a new exception to enable collecting and using special categories of data for AI non-discrimination auditing. The world seems not yet ready for such an exception, as auditing AI systems is still very difficult.

Andrus, Spitzer, Brown and Xiang note that 'the algorithmic fairness literature provides a number of techniques for potentially mitigating detected bias, but practitioners noted that there are few applicable, touchstone examples of these techniques in practice.[80] Balayn and Gürses write: '[w]hether applied to data-sets or algorithms, technocentric de-biasing techniques have profound limitations: they address bias in mere statistical terms, instead of accounting for the varied and complex fairness requirements and needs of diverse system stakeholders.'[81]

78 Seltzer & Anderson 2022, p. 507.

79 Balayn & Gürses 2021, p. 94.

80 Andrus and others, *arXiv:2011.02282 [cs]* 2021, p. 257.

81 Balayn & Gürses 2021, p. 12. See also Veale & Binns, *Big Data & Society* 2017/4, p. 13. '[t]o adequately address fairness in the context of machine learning, researchers and practitioners working towards "fairer" machine learning need to recognise that this is not

A 2021 interview study found several practical reasons explaining why industry practitioners *themselves* find it difficult to test if an AI system discriminates. For instance, the collected special category data may not be accurate, because the data may have originally been collected by another party, or for a different purpose. And because data subjects may report their own (self-perceived) ethnicities, the data could be inaccurate or unusable for the test.[82] An interview study from 2019 found that many industry practitioners did not have an infrastructure in place for collecting accurate special categories of data. Furthermore, practitioners mentioned that auditing methods themselves are not holistic enough, and practitioners do not always know which subpopulations they need to consider when auditing an AI system for discrimination.[83]

Since 'best practices' for auditing an AI system for discrimination seem to be in their infancy, it is questionable whether creating a new exception is currently justified. However, the techniques for auditing and de-biasing AI are improving. The more knowledge exists about how to test an AI system for discrimination, the more justified a new exception could become in the future.

All in all, there are various arguments in favour of and against adopting an exception that enables the use of special categories of data to prevent AI-driven discrimination. The balance between the pro and contra arguments is difficult to find. If such an exception were adopted, the exception should also include safeguards to minimise risks. We discuss possible safeguards in the next section.

7 Possible safeguards if an exception were adopted

7.1 Safeguards in the proposed AI Act

A proposal by the EU illustrates some possibilities for safeguards. Early 2021, the European Commission presented a proposal for an AI Act, with an

just an abstract constrained optimisation problem. It is a messy, contextually-embedded and necessarily sociotechnical problem, and needs to be treated as such.'

82 Andrus and others, *arXiv:2011.02282 [cs]* 2021/6.3.

83 Holstein and others, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 2019, p. 1.

exception to the ban on using special category data. The proposed exception is phrased as follows:[84]

> To the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems, the providers of such systems may process special categories of personal data referred to in [Article 9 of the GDPR], subject to appropriate safeguards for the fundamental rights and freedoms of natural persons, including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving measures, such as pseudonymisation, or encryption where anonymisation may significantly affect the purpose pursued.

There are various safeguards in the proposed provision that aim to prevent abuse of the special category data.

7.1.1 Strictly necessary for preventing discrimination

In the AI Act, the exception to the ban on using special category data only applies '[t]o the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems'.

The phrase 'strictly necessary' implies a higher bar than merely 'necessary'. The word 'necessary' is already quite stern. The CJEU said in the *Huber* case that 'the concept of necessity (…) has its own independent meaning in Community law.'[85] And necessity 'must be interpreted in a manner which fully reflects the objective of [the Data Protection] directive'.[86] CJEU case law shows the word 'necessary' must be interpreted narrowly, in favour of the data subject: 'As regards the condition relating to the necessity of processing personal data, it should be borne in mind that *derogations and limitations in*

84 Quotation lightly edited by authors. See Article 10(5) and recital 44 of the European Commission Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM/2021/206 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

85 CJEU 16 December 2008, C-524/06, ECLI:EU:C:2008:724 (*Huber*), para. 52.

86 CJEU 16 December 2008, C-524/06, ECLI:EU:C:2008:724 (*Huber*), para. 52.

relation to the protection of personal data must apply only in so far as is strictly necessary (…).'[87] In sum, organisations should only rely on the exception in the AI Act if using special category data is genuinely necessary.

7.1.2 The AI Act exception only applies to providers of high-risk AI systems

The AI Act's exception only applies to high-risk AI systems. High-risk AI systems can be divided into two types: First, products already covered by certain EU health and safety harmonisation legislation (such as toys, machinery, lifts, or medical devices). Second, AI systems specified in an annex of the AI Act, in eight areas or sectors.[88]

If lawmakers consider creating a new exception to enable the use of special category data, they could limit the scope of the exception in a similar way. Perhaps organisations should not be allowed to rely on the exception if the AI system does not bring serious discrimination risks.

7.1.3 Appropriate safeguards

The exception in the proposed AI Act says that special category data can be used to prevent AI-driven discrimination, 'subject to appropriate safeguards for the fundamental rights and freedoms of natural persons'.[89] The provision gives examples of such safeguards: 'including technical limitations on the re-use and use of state-of-the-art security and privacy-preserving

87 CJEU May 2017, C-13/16 (*Rigas*), para. 30.: 'As regards the condition relating to the necessity of processing personal data, it should be borne in mind that derogations and limitations in relation to the protection of personal data must apply only in so far as is strictly necessary.' See also CJEU 8 April 2014, C-293/12 and C-594 (*Digital Rights Ireland Ltd*), para. 52.: 'according to the Court's settled case-law, (…) derogations and limitations in relation to the protection of personal data must apply only in so far as is strictly necessary .' See also ECHR 25 March 1983, 5947/72; 6205/73; 7052/75; 7061/75; 7107/75; 7113/75; 7136/75 (*Silver and others/Verenigd Koninkrijk*), para. 97. The European Court of Human Rights says that '[t]he adjective 'necessary' is not synonymous with 'indispensable', neither has it the flexibility of such expressions as 'admissible', 'ordinary', 'useful', 'reasonable' or 'desirable' (…).' As stated previously in Section 5.4, the Dutch DPA has also used the phrase 'truly necessary', referring to a higher standard of proportionality.

88 See Veale & Zuiderveen Borgesius, *Computer Law Review International* 2021/4, p. 102.

89 Article 10(3) Draft AI Act.

measures, such as pseudonymisation, or encryption where anonymisation may significantly affect the purpose pursued.'[90] If an exception were adopted to enable the use of special category data for fighting AI-driven discrimination, a similar requirement should be included.

Some elements in the proposed AI Act exception are controversial. For instance, the current exception leaves unclear *who* decides what the appropriate safeguards are. With the current text, that burden seems to rest on the provider of the AI system him or herself. Therefore, the exception seems to leave the exact safeguards up to the provider of the AI system.

Moreover, as Balayn & Gürses note, the 'European Commission proposal to regulate AI enables the use of sensitive attributes for de-biasing, without further consideration of the risks it imposes on exactly the populations that the regulation says it intends to protect.'[91] Indeed, the AI Act does not include measures to limit the risks associated with collecting special category data.

7.2 Other possible safeguards

Are other safeguards, not mentioned in the AI Act, viable? A specific technical safeguard is to create a synthetic, anonymous dataset from the 'real' dataset. The fake dataset represents the same (or a similar) distribution of individuals but can no longer be linked to the individuals. Such data can be safely stored.

The original special categories of data still need to be collected to create a synthetic dataset, but the original data can be stored for less time. It is controversial whether and when using such a dataset is effective for testing AI systems. At the current time of writing, the privacy gain seems to vary greatly, and it is unpredictable how much utility is lost by making the dataset synthetic.[92]

Other possible safeguards are more organisational. For example, a trusted third party could collect the special categories of data, store them and use them for auditing the AI system. The organisation using the AI system then

90 Article 10(3) Draft AI Act.

91 Balayn & Gürses 2021, p. 94.

92 Stadler, Oprisanu & Troncoso 2022. See also on the limits of using synthetic datasets Bellovin, Dutta & Reitinger, *22 Stan. Tech. L. Rev. 1 (2019)*, p. 14 and further.

no longer needs to store the data itself. However, it seems debatable if such a construction is practical, and financially feasible.[93]

Such organisational safeguards raise many questions. For example, which third party can be trusted with special categories of data? One possibility might be the national Statistics Bureaus of member states. Such bureaus could store and manage the special category data safely. In Europe, statistics bureaus have long been responsible for collecting special categories of data for statistical purposes, on a large scale. Or perhaps an independent supervising authority could appoint trustworthy researchers and give them access to the special categories of data, to prevent discrimination by AI systems. Specific privacy-related solutions also exist that might allow a third party to more safely use the data.[94]

8 Conclusion

In this paper, we examined whether the GDPR needs a new exception to the ban on using special categories of data, such that an organisation can mitigate discrimination by artificial intelligence. We mapped out the arguments in favour of and against such a new exception.

We presented the following main arguments in favour of such an exception. (i) Organisations could use the special category data to test AI against discrimination. (ii) AI discrimination testing could increase the trust consumers have in an AI system.

The main arguments against such an exception are as follows. (i) Storing personal data about, for instance, ethnicity can be seen as a privacy interference. (ii) Such data can be abused, or data breaches can occur. (iii) The exception could be abused to collect special categories of data

93 See also Kilbertus and others, *arXiv:1806.03281 [cs, stat]* 2018, p. 1-2.

94 To illustrate, say a statistics bureau audits an organisation's patented AI system. If both organisations use secure multi party computation (MPC), the statistics bureau cannot see the model of the organisation it audits and the organisation cannot see the special category data used for the audit, but they could audit the AI system together. A caveat is that MPC creates some overhead in calculations and is complex to set up. See Kilbertus and others, *arXiv:1806.03281 [cs, stat]* 2018/3.2. See also Agahari, Ofe & De Reuver, *Electronic Markets* 2022.

for other uses than AI discrimination testing. (iv) In addition, merely allowing organisations to collect special category data does not guarantee that organisations can de-bias their AI systems. Auditing and de-biasing AI systems remains difficult.

In the end, it is a political decision how the balance between the different interests must be struck. Ideally, such a decision is made after a thorough debate. We hope to inform that debate.

References of chapter III

Adams-Prassl, Binns & Kelly-Lyth, *The Modern Law Review* 2022
J. Adams-Prassl, R. Binns & A. Kelly-Lyth, 'Directly Discriminatory Algorithms', *The Modern Law Review* 2022, https://onlinelibrary.wiley.com/doi/10.1111/1468-2230.12759, doi:10.1111/1468-2230.12759.

Agahari, Ofe & De Reuver, *Electronic Markets* 2022
W. Agahari, H. Ofe & M. de Reuver, 'It is not (only) about privacy: How multi-party computation redefines control, trust, and risk in data sharing', *Electronic Markets* 2022, https://link.springer.com/10.1007/s12525-022-00572-w, doi:10.1007/s12525-022-00572-w.

Alidadi, *European Equality Law review* 2017/2
K. Alidadi, 'Gauging progress towards equality? Challenges and best practices of equality data collection in the EU', *European Equality Law review* 2017/2.

Al-Zubaidi, *European Equality Law review* 2020/2
Y. Al-Zubaidi, 'Some reflections on racial and ethnic statistics for anti-discrimination purposes in Europe', *European Equality Law review* 2020/2.

Andrus and others, *arXiv:2011.02282 [cs]* 2021
M. Andrus and others, ' "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness', *arXiv:2011.02282 [cs]* 2021, http://arxiv.org/abs/2011.02282.

Balayn & Gürses 2021
A. Balayn & S. Gürses, *Beyond Debiasing. Regulating AI and its inequalities*, European Digital Rights (EDRi) 2021, https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/.

Barocas & Selbst, *California Law Review* 2016/104
S. Barocas & A.D. Selbst, 'Big Data's disparate impact', *California Law Review* 2016/104, p. 671.

Bellovin, Dutta & Reitinger, *22 Stan. Tech. L. Rev. 1 (2019)*
S.M. Bellovin, P.K. Dutta & N. Reitinger, 'Privacy and Synthetic Datasets', *22 Stan. Tech. L. Rev. 1 (2019)*, https://law.stanford.edu/wp-content/uploads/2019/01/Bellovin_20190129.pdf.

Calo, *Indiana Law Journal*
M.R. Calo, 'The Boundaries of Privacy Harm', *Indiana Law Journal*, p. 31, https://www.repository.law.indiana.edu/ilj/vol86/iss3/8/.

Dutch Data Protection Authority 2020
Dutch Data Protection Authority, *Besluit tot het opleggen van een bestuurlijke boete [decision to impose an administrative fine]*, 2020, https://autoriteitpersoonsgegevens.nl/nl/nieuws/boete-voor-cpa-om-privacyschending-zieke-werknemers.

E. Black 2012
E. Black, *IBM and the Holocaust: The Strategic Alliance Between Nazi Germany and America's Most Powerful Corporation*, Dialog press 2012.

Ellis & Watson 2012
E. Ellis & P. Watson, *EU anti-discrimination law*, Oxford: Oxford University Press 2012.

European Commission 2020
European Commission, *COM(2020) 65 White paper On Artificial Intelligence - A European approach to excellence and trust*, EU 2020, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Data Protection Board 2020
European Data Protection Board, *Guidelines 05/2020 on consent under Regulation 2016/679*, EDPB 2020, https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf.

European Union Agency for Fundamental Rights, European Court of Human Rights, & Council of Europe (Strasbourg) 2018
European Union Agency for Fundamental Rights, European Court of Human Rights, & Council of Europe (Strasbourg), *Handbook on European non-discrimination law*, LU: Publications Office 2018, https://fra.europa.eu/en/publication/2018/handbook-european-non-discrimination-law-2018-edition.

Farkas 2017
L. Farkas, *The meaning of racial or ethnic origin in EU law: between stereotypes and identities, European network of legal experts in gender equality*

and non-discrimination (report for European Commission, Directorate-General for Justice and Consumers), Luxembourg: Publications Office of the European Union 2017, https://www.equalitylaw.eu/downloads/4030-the-meaning-of-racial-or-ethinic-origin-in-eu-law-between-stereotypes-and-identities.

FRA 2013
FRA, *FRA Opinion on the situation of equality in the European Union 10 years on from initial implementation of the equality directives*, 2013, https://fra.europa.eu/en/publication/2013/fra-opinion-situation-equality-european-union-10-years-initial-implementation.

Granger & Irion 2018
M. Granger & K. Irion, 'The right to protection of personal data: the new posterchild of European Union citizenship?', in: S. de Vries, H. de Waele & M. Granger, *Civil Rights and EU Citizenship*, Cheltenham: Edward Elgar 2018, p. 279-302, https://www.elgaronline.com/view/edcoll/9781788113434/9781788113434.00019.xml.

Gurses 2010
F.S. Gurses, *Multilateral Privacy Requirements Analysis in Online Social Network Services*, KU Leuven 2010.

High-level expert group on Artificial Intelligence 2019
High-level expert group on Artificial Intelligence, *A definition of AI. Main capabilities and disciplines*, Brussels: European Commission 2019, https://ec.europa.eu/newsroom/dae/redirection/document/56341.

Holstein and others, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 2019, p. 1-16
K. Holstein and others, 'Improving fairness in machine learning systems: What do industry practitioners need?', *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 2019, p. 1-16, http://arxiv.org/abs/1812.05239, doi:10.1145/3290605.3300830.

Kilbertus and others, *arXiv:1806.03281 [cs, stat]* 2018
N. Kilbertus and others, 'Blind Justice: Fairness with Encrypted Sensitive Attributes', *arXiv:1806.03281 [cs, stat]* 2018, http://arxiv.org/abs/1806.03281.

Kosta 2011
E. Kosta, *Unravelling consent in European data protection legislation - a prospective*

study on consent in electronic communications, KU Leuven 2011, https://lirias.kuleuven.be/retrieve/c329678b-dc83-4830-bb1e-01146591c17d.

Kuner and others 2020
C. Kuner and others, *The EU General Data Protection Regulation (GDPR): A Commentary*, New York: Oxford University Press 2020, https://oxford.universitypressscholarship.com/10.1093/oso/9780198826491.001.0001/isbn-9780198826491, doi:10.1093/oso/9780198826491.001.0001.

Makkonen 2016
T. Makkonen, *European handbook on equality data: 2016 revision*, LU: Publications Office 2016, https://data.europa.eu/doi/10.2838/397074.

Open Data Institute 2020
Open Data Institute, *Monitoring equality in digital public services*, 2020, https://theodi.org/wp-content/uploads/2020/01/OPEN-ODI-2020-01_Monitoring-Equality-in-Digital-Public-Services-report.pdf.

OSTP 2022
OSTP, *Blueprint for an AI Bill of Rights. Making automated systems work for the American people*, The White house | OSTP 2022, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Seltzer & Anderson 2022
W. Seltzer & M. Anderson, 'The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses', 2022, https://www.jstor.org/stable/40971467.

Stadler, Oprisanu & Troncoso, *USENIX* 2022
T. Stadler, B. Oprisanu & C. Troncoso, 'Synthetic Data – Anonymisation Groundhog Day', *31st USENIX security symposium* 2022, https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

Thomsen 2018
F.K. Thomsen, 'Direct Discrimination', in: K. Lippert-Rasmussen, *The Routledge Handbook of the Ethics of Discrimination*, London and New York: Routledge Taylor & Francis Group 2018, p. 19-29.

Tilburg Institute for Law, Technology, and Society 2021
Tilburg Institute for Law, Technology, and Society, *Handbook on*

non-discriminating algorithms. Summary research report, 2021, https://www.tilburguniversity.edu/about/schools/law/departments/tilt/research/handbook.

Van Caeneghem 2019
J. Van Caeneghem, 'Ethnic Data Collection: Key Elements, Rules and Principles', in: J. Van Caeneghem, *Legal Aspects of Ethnic Data Collection and Positive Action*, Cham: Springer International Publishing 2019, p. 155-257, http://link.springer.com/10.1007/978-3-030-23668-7_3, doi:10.1007/978-3-030-23668-7_3.

Van den Brink & Dunne 2018
M. Van den Brink & P. Dunne, *Trans and intersex equality rights in Europe: a comparative analysis.*, LU: Publications Office 2018, https://data.europa.eu/doi/10.2838/75428.

Veale & Binns, *Big Data & Society* 2017/4
M. Veale & R. Binns, 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', *Big Data & Society* 2017/4, issue 2, p. 205395171774353, http://journals.sagepub.com/doi/10.1177/2053951717743530, doi:10.1177/2053951717743530.

Veale & Zuiderveen Borgesius, *Computer Law Review International* 2021/4
M. Veale & F.J. Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act', *Computer Law Review International* 2021/4.

Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24
I. Žliobaitė & B. Custers, 'Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models', *Artificial Intelligence and Law* 2016/24, issue 2, p. 183-201, http://link.springer.com/10.1007/s10506-016-9182-5, doi:10.1007/s10506-016-9182-5.

Zuiderveen Borgesius 2019
F.J. Zuiderveen Borgesius, *Discrimination, artificial intelligence, and algorithmic decision-making. Report for the European Commission against Racism and Intolerance (ECRI)*, Council of Europe 2019.

Zuiderveen Borgesius, *European Business Law Review* 2020/31
F. Zuiderveen Borgesius, 'Price Discrimination, Algorithmic Decision-

Making, and European Non-Discrimination Law', *European Business Law Review* 2020/31, issue 3, p. 22.

CJEU 16 December 2008, C-524/06, ECLI:EU:C:2008:724 (*Huber*).

CJEU October 2013, C291/12 (*Schwartz/Stadt Bochum*).

CJEU 8 April 2014, C-293/12 and C-594 (*Digital Rights Ireland Ltd*).

CJEU May 2017, C-13/16 (*Rigas*).

CJEU (Grand chamber) 16 July 2015, Case C-83/14, ECLI:EU:C:2015:480 (*Chez/Nikolova*).

ECHR 25 March 1983, 5947/72; 6205/73; 7052/75; 7061/75; 7107/75; 7113/75; 7136/75 (*Silver and others/Verenigd Koninkrijk*).

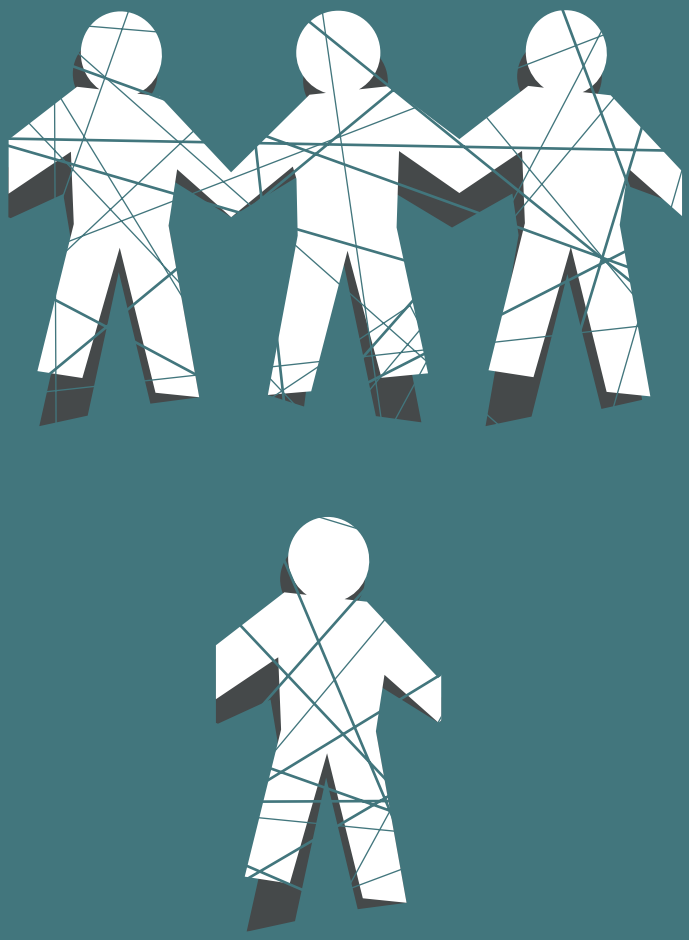
ECHR 26 March 1987, 9248/81 (*Leander/Sweden*).

ECHR 16 February 2000, 27798/95 (*Amann v. Switzerland*).

ECHR 3 April 2007, 62617/00 (*Copland v. United Kingdom*).

ECHR 4 December 2008, 30562/04 and 30566/04 (*S. and Marper v. United Kingdom*).

ECHR (Grand Chamber) 25 May 2021, 58170/13, 62322/14 and 24960/15 (*Big Brother Watch and others/United Kingdom*).



IV.

Using sensitive data to de-bias AI systems: Article 10(5) of the EU AI act

Submitted to *Computer Law & Security Review* as:

M. van Bekkum, 'Using sensitive data to de-bias AI systems: Article 10(5) of the EU AI act', *Computer Law & Security Review* 2025/56, issue 106115, https://linkinghub.elsevier.com/retrieve/pii/S026736492500010X, doi:10.1016/j.clsr.2025.106115.

The author thanks Ylja Remmits (AlgorithmAudit), Tom Heskes (Radboud University), Frederik Zuiderveen Borgesius (Radboud University), Jaap-Henk Hoepman (Radboud University), Francien Dechesne (Leiden University) and the participants of the Leiden eLaw Conference 2024 for their useful commentary.

Abstract

In June 2024, the EU AI Act came into force. The AI Act includes obligations for the provider of an AI system. Article 10 of the AI Act includes a new obligation for providers to evaluate whether their training, validation and testing datasets meet certain quality criteria, including an appropriate examination of biases in the datasets and correction measures. With the obligation comes a new provision in Article 10(5) AI Act, allowing providers to collect sensitive data to fulfil the obligation. Article 10(5) AI Act aims to prevent discrimination. In this paper, I investigate the scope and implications of Article 10(5) AI Act. The paper primarily concerns European Union law, but may be relevant in other parts of the world, as policymakers aim to regulate biases in AI systems.

1 Introduction

In June 2024, the EU AI Act came into force.[1] The European Union legislator aims for, on the one hand, innovation within the European Union: introducing AI could have many competitive advantages in many areas of life, from healthcare to culture, public services, justice and climate change mitigation and adaptation.[2] The AI Act aims to 'improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI)'.[3] On the other hand, the AI Act aims to protect people against risks to health, safety, fundamental rights, and several other public interests.[4]

The AI Act includes obligations for the provider of an AI system: the 'natural or legal person that develops an AI system [...] or that has an AI system [...] developed and places it on the market or puts the AI system into service'.[5] Article 10 of the AI Act includes a new obligation concerning the de-biasing of AI systems: in short, the providers of AI systems must evaluate whether their training, validation and testing datasets meet certain quality criteria. Those criteria include an 'examination in view of possible biases' and 'appropriate measures to detect, prevent and mitigate possible biases.'[6] With the obligation comes a new provision in Article 10(5) AI Act, allowing providers to collect sensitive data to fulfil the obligation. The provision aims to prevent discrimination.[7] Before the AI Act, the ban on collecting sensitive

1 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689.

2 Recital 4 AI Act.

3 Article 1(1) AI Act. The term 'human-centric AI' originates from earlier European Commission documents from 2019. See European Commission 2019.

4 Recital 5 AI Act and Article 1(1) AI act.

5 Definition of 'provider', Article 3(3) AI Act.

6 Article 10(2)(f) and (g) AI Act.

7 Recital 70 AI Act.

data in the GDPR meant that developers were not allowed to collect such data for AI de- biasing.[8]

In this paper, I investigate the scope and implications of Article 10(5) AI Act. Practitioners implementing the AI Act, or legal scholars who wish to understand the AI Act's exception and definitions better, may benefit from the paper's analysis. The paper focuses primarily on European Union law, but may be useful in other parts of the world as well.[9] In the analysis, I will include insights from law and computer science that further elaborate the AI Act and the de-biasing of AI systems.

The contributions of the paper are as follows. A paper by Hacker discusses AI training data and Article 10 AI Act in general.[10] I focus on a different aspect of Article 10 AI Act: the exception allowing sensitive data and de-biasing. As far as I know, this is the first paper analysing the exception in-depth. The paper discusses the AI Act's definitions of AI system, provider, and deployer within the specific context of AI de- biasing. I only discuss the definitions in the AI Act as far as is necessary to understand the exception.[11] Finally, the practical difficulties providers could have to collect sensitive data fall outside of the paper's topic.[12]

The paper first discusses why providers cannot de-bias without sensitive data and an exception, in Section 2. Section 3 presents the text of the exception. Section 4 analyses the text of Article 10(5) AI Act, element by element. In Section 5, I discuss whether the aim of the exception – fighting discrimination by AI systems - is practicable for providers. Section 6 concludes.

8 Article 9 GDPR. See also Van Bekkum & Zuiderveen Borgesius, *Computer Law & Security Review* 2023/48.

9 For example, the blueprint for the US AI Bill of Rights refers to 'proactive equity assessments as part of the system design'. Article 10 AI Act is an example of such a proactive assessment. See OSTP 2022, p. 5, 23, 26.

10 Hacker, *Law, Innovation and Technology* 2021/13.

11 For a full paper about the definition of the term AI system, see e.g. Ruschemeier, *ERA Forum* 2023/23. See also Boone, *Journal of Data Protection & Privacy* 2023/6.

12 It is unclear if, in practice, providers will succeed in collecting the sensitive data. Organisations have several options, such as data donation campaigns, asking the data subject, or inferring from proxy attributes. For a list, see for example UK Government (Department for Science, Innovation and Technology), 'Improving responsible access to demographic data to address bias', https://rtau.blog.gov.uk/2023/06/14/improving-responsible-access-to-demographic-data-to-address-bias/, 14 June 2023.

2 Providers cannot de-bias without sensitive data and an exception

Before discussing the exception itself, I discuss why providers need to process sensitive data to de-bias AI systems. Next, I aim to show why providers may require an exception to the GDPR that allows AI de- biasing, based on Van Bekkum & Zuiderveen Borgesius (2023). Readers who are familiar with the topic can skip this section.[13]

2.1 Without sensitive data, providers cannot test proxies in data

Organisations can use AI systems to make decisions about people. For instance, a bank could use an AI system to assess whether a customer is creditworthy enough to apply for a mortgage. But using AI systems to make such decisions can lead to discrimination. For example, the bank's AI system could disproportionally deny mortgages to people with a certain ethnicity, even if the bank did not plan to discriminate.

Suppose that the developer of the bank's AI system wishes to test whether the decisions made by the AI system have effects that result in indirect discrimination of people with a certain ethnicity.[14] The developer must then first find possible biases in the AI's decisions. For such a bias test, the developer needs to know the ethnicity of the people about whom its AI system makes decisions.[15]

The developer of an AI system cannot prevent the system from having discriminatory effects against a certain ethnicity by removing explicit references to ethnicity from the AI system's design. Even if an AI system does not use explicit ethnicity data, other attributes such as a postal code can act as a proxy, or placeholder, for ethnicity. To find out if an attribute is a proxy for ethnicity with respect to an AI system's decision, the provider must collect ethnicity data. If a provider builds an AI system based on proxies that correlate strongly with ethnicity, then the user of the AI system may indirectly discriminate based on ethnicity. Roughly summarised, indirect

13 Van Bekkum & Zuiderveen Borgesius, *Computer Law & Security Review* 2023/48.

14 The developer is not necessarily the user of the AI system.

15 Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24.

discrimination occurs when a seemingly neutral practices ends up harming people with a certain ethnicity or other legally protected characteristic.[16]

As a fictive example, take a government organisation offering student grants to students.[17] The government organisation must check whether students live at home or not, because students who live at home receive a lower monthly grant. Students could commit fraud registering a different home address. The government organisation uses an AI system to find potential fraudsters. Based on three factors, the AI system identifies potential fraudsters: 1. type of education, 2. age and 3. distance from parent(s) house. The organisation wants to test whether the AI system, built based on these three factors, discriminates against students with a certain ethnicity.

The factors *age, education or distance* do not reveal information about the ethnicity of applicants: at first glance, the factors are neutral. To test whether the AI system discriminates against certain ethnicities, the government organisation must therefore test if the factors correlate with the ethnicities of the students.[18] Based on the collected ethnicity data, the government organisation can, for example, calculate the correlation between each of the three factors and ethnicity. Furthermore, the organisation can try to predict ethnicity using the three factors.[19] If the government organisation can use the factor (or a combination) to accurately predict ethnicity, then the factor

16 Zuiderveen Borgesius and others 2024.

17 The example is based on a real bias audit of the AI system of the Dutch Education Executive Agency, and the paper by Žliobaitė & Custers. See AlgorithmAudit 2024. Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24, p. 183.

18 In this case, the organisation could not use existing publicly available, aggregated data about ethnicity by postal code, such as the data collected by bureaus of statistics, because such information is probably not representative: the public data is not only about students who were given a grant, but about the general population. See AlgorithmAudit 2024, p. 23.

19 A factor can (individually) be non-correlated to ethnicity, but still discriminate in combination with other factors. See McLoughney and others, *IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* 2023. Instead, predicting ethnicity is often necessary. See Keegan Hines, 'Mining for Proxies in Machine Learning Systems', https://www.arthur.ai/blog/mining-for-proxies-in-machine-learning-systems, 18 April 2022. Zhang, Lemoine & Mitchell 2018. Chaudhary and others claim to not require sensitive data by using proxies, but probably require sensitive data to validate whether the generated proxies are correct (Compare with footnote 22 of this paper). Chaudhary and others 2023.

(or combination) is a proxy, or a placeholder, for ethnicity. The government organisation can decide, based on the circumstances of the case (and perhaps fairness metrics), if the AI system discriminates against a certain ethnicity.[20]

If the government organisation had used an AI system (such as machine learning) built not upon three, but thousands of factors, the problem remains the same: the provider would still need sensitive data to decide if using such an AI system has discriminatory effects on certain ethnicities.[21]

In some cases, an audit without sensitive data may be enough. The government agency could have tested whether non-protected vulnerable groups suffer a disadvantage (for example, migrants in general).[22] But for finding discriminatory effects regarding specific ethnicities, the organisation would still need sensitive data.

2.2 The GDPR, in principle, prohibits providers from collecting sensitive data

In Europe, developers used to run into a problem: In principle, the GDPR prohibits the use of certain categories of personal data (also called 'sensitive data'). The GDPR's special categories of data overlap with protected characteristics in non-discrimination law. The EU non- discrimination directives prohibit discrimination based on six protected characteristics: age; disability; gender; religion or belief; racial or ethnic origin; sexual orientation. Of those six characteristics, the following four are also special categories of

20 In their assessment, the organisation could also weigh so-called fairness metrics, (roughly speaking) numbers which express how biased an AI system is toward different groups. Currently, it is controversial how heavy providers must weigh a fairness metric in assessing legal non-discrimination risks. See Barocas, Hardt & Narayanan 2023.

21 Perhaps the government organisation could try to automate or support some of the assessment with an algorithm, such as in discrimination-aware data mining (note: building such an algorithm is a challenge in itself). However, in that case, the organisation still needs sensitive data to make the algorithm and verify its output.

22 In the real-life case this example is based on, the auditing organization tested for biases against non-protected categories, due to lack of sensitive data. Such a test could give an (initial) impression of whether the system is biased. See AlgorithmAudit 2024. Some automated tools offer a 'bias scan' attempting to find (new) vulnerable groups in data, see S. Muhammad, 'Auditing Algorithmic Fairness with Unsupervised Bias Discovery', http://amsterdamintelligence.com/posts/bias-discovery, 14 September 2021. See in more detail Misztal-Radecka & Indurkhya, *Information Processing & Management* 2021/58.

data in the GDPR: (1) disability, (2) religion or belief, (3) racial or ethnic origin and finally (4) sexual orientation.[23]

The GDPR itself includes one exception to the ban that is sometimes suitable for de-biasing AI systems. Organisations could ask the individual's explicit consent for collecting and using the sensitive data. But in many situations, an organisation cannot obtain valid consent of an individual. In such situations, the GDPR effectively bans using special category data for AI de-biasing.[24] Other possible exceptions in Article 9 (2) GDPR require a 'union or member state law'. As far as I know, national laws in Europe do not provide for an exception enabling the processing of sensitive data for AI de-biasing.[25]

The AI Act's exception is a European Union law within the meaning of the GDPR. The AI Act's exception is in the public interest, building on Article 9(2)(g) GDPR.[26] I use the term exception, even if dogmatically one could call the provision in Article 10 an implementation.[27] Article 9(2)(g) GDPR allows an exception 'for reasons of substantial public interest' such as fighting discrimination, based on 'union or member state law' such as the AI Act, under three conditions. First, the law much be *proportionate to the aim pursued*. Second, the law must *respect the essence of the right to data protection*. Third, the law must *provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject*.[28]

23 Article 9(1) GDPR. See also Van Bekkum & Zuiderveen Borgesius, *Computer Law & Security Review* 2023/48, p. 5-6.

24 Van Bekkum & Zuiderveen Borgesius, *Computer Law & Security Review* 2023/48.

25 Van Bekkum & Zuiderveen Borgesius, *Computer Law & Security Review* 2023/48.

26 According to Recital 70 of the AI Act, the AI Act's exception is 'a matter of substantial public interest within the meaning of Article 9(2)(g) of [the GDPR]'.

27 In the strict legal sense, the provision in Article 10(5) AI Act implements an exception in the GDPR. However, Article 9(2)(g) GDPR would not have a legal effect for providers without the provision in Article 10(5) AI Act, so I call it an exception in this paper.

28 I discuss whether some of these conditions hold in Section 5 of the paper.

3 Relevant text of the exception

Find below, for ease of reading, the text of Article 10(5) AI Act.

Article 10(5)

5. To the extent that it is strictly necessary for the purpose of ensuring bias detection and correction in relation to the high-risk AI systems in accordance with paragraph (2), points (f) and (g) of this Article, the providers of such systems may exceptionally process special categories of personal data, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons. In addition to the provisions set out in [the GDPR] and [the GDPR for EU institutions] and [the law enforcement directive], all the following conditions must be met in order for such processing to occur:

(a) the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data;

(b) the special categories of personal data are subject to technical limitations on the re-use of the personal data, and state-of-the-art security and privacy-preserving measures, including pseudonymisation;

(c) the special categories of personal data are subject to measures to ensure that the personal data processed are secured, protected, subject to suitable safeguards, including strict controls and documentation of the access, to avoid misuse and ensure that only authorised persons have access to those personal data with appropriate confidentiality obligations;

(d) the special categories of personal data are not to be transmitted, transferred or otherwise accessed by other parties;

(e) the special categories of personal data are deleted once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first;

(f) the records of processing activities pursuant to [the GDPR] and [the GDPR for EU institutions] and [the law enforcement directive] include the reasons why the processing of special categories of personal data was strictly necessary to detect and correct biases, and why that objective could not be achieved by processing other data.

4 Analysis of the exception

The introduction of paragraph 5 of Article 10 AI Act introduces several terms that narrow the scope of the exception. I discuss these terms in the following sections: *high-risk AI system* (4.1), *provider* (4.2), *bias detection and correction* and *the de-biasing obligation* (4.3) and *strictly necessary* (4.4). The introduction states that the processing of the sensitive data must be *subject to appropriate safeguards for the fundamental rights and freedoms of individuals* (4.5). The exception applies 'in addition to the provisions set out in the GDPR' (4.6).

4.1 Exception only for high-risk AI

The exception applies to high-risk AI systems. The definition of the term AI system in Article 3(1) AI Act is:

a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments

For a given piece of software to qualify as an AI system, it must have four characteristics: autonomy, adaptiveness, an objective, and the capability to

infer.[29] Any software could, in principle, show these four characteristics. The EU legislator added the characteristic 'infer' later in the legislative process, to keep the definition technologically neutral. At the very least, the definition includes machine learning and logic- and knowledge-based approaches.[30]

According to Recital 12 AI Act, the EU legislator intended to exclude 'simpler traditional software systems or programming approaches', such as Microsoft Excel's auto-sum function, from the definition. On the other hand, the definition of the term *AI system* is open-ended, where context determines if a simple system qualifies as AI or not.[31] The most decisive of the four characteristics to distinguish between 'simple software' and AI is the ability to infer, which means grossly speaking, software that obtains output from input or that derives models from input.[32] As a result, in principle, traditional rule-based (programmed) systems seem excluded from the definition, but depending on the context, the simpler systems qualify. The burden of assessing which systems are AI rests on the provider.[33] Legal certainty for providers would increase if the AI Office or AI Board give examples of specific use cases where simpler software qualifies as AI.[34]

The exception in Article 10(5) AI Act only applies for AI systems that the AI Act considers *high-risk*. According to Article 6 AI Act, in two cases, the AI Act considers an AI system high-risk.[35] First, the AI Act considers systems that

29 See Article 3 AI Act. In Recital 12 AI Act, the legislator further describes these characteristics.

30 In a previous version of the AI Act, the Commission included a list with explicit AI techniques. In the final version, the EU legislator has removed the list and replaced it with the term 'infer', in order to keep the definition technologically neutral. The explicit list included machine learning and/or logic- and knowledge-based approaches. See Fernández-Llorca and others, *Artificial Intelligence and Law* 2024.

31 See also the explanatory memorandum by the OECD 2023/2.

32 See Recital 12 AI Act.

33 Especially governments use mainly rule-based AI. See for more details about implications Enqvist, *Information & Communications Technology Law* 2024, p. 1.

34 See Articles 64 (1) and 66(e), (g) and (i) AI Act. Article 10 AI Act comes into force on 2 August 2026, Article 113 AI Act. See also Hacker, *SSRN Electronic Journal* 2024/II. Fernández-Llorca and others, *Artificial Intelligence and Law* 2024. M. Grobelnik, K. Perset & S. Russell, 'What is AI? Can you make a clear distinction between AI and non-AI systems?', https://oecd.ai/en/wonk/definition, March 2024.

35 Note: the AI Act's high-risk categories do not guarantee that the system is also high-risk in other laws, such as the GDPR.

fall in the scope of certain existing EU regulations high-risk. Second, the AI Act includes a list of high-risk AI systems in Annex III. Grossly speaking, the term high-risk encompasses AI systems in: (i) biometrics, (ii) critical infrastructures, (iii) education, (iv) employment, (v) essential private and public services, (vi) law enforcement, (vii) migration, (viii) administration of justice and democracy. The categories mention specific applications of such systems, spanning from job recruitment in the employment sector to AI systems helping judicial authorities decide on legal matters.[36] Some AI systems escape the scope of Annex III, such as dating apps.[37] There are three limited exceptions to *high-risk* in the AI Act, but the exceptions only apply if the AI system does not *profile* natural persons, which will often be the case.[38]

In sum, the definition of the term *AI system* is complex, but at the very least covers AI systems that are not too simple. Many AI systems could be high-risk in practice, especially because they profile natural persons and fall under the definition in Annex III AI Act.

4.2 Exception applies to provider

The AI Act's exception applies to providers of AI systems. According to Article 3 AI Act, a provider is 'a natural or legal person, public authority, agency or other body that develops an AI system or a general purpose AI model or that has an AI system or a general purpose AI model developed and places them on the market or puts the system into service under its own name or trademark, whether for payment or free of charge.' Several examples

36 Annex III AI Act.

37 In The Netherlands, a Dutch dating app developer lacks the exception necessary to de-bias its AI system, even after being required to de-bias the AI by national authorities. See (in Dutch): College voor de Rechten van de Mens [The Netherlands Institute for Human Rights], 'Dating-app Breeze mag (en moet) algoritme aanpassen om discriminatie te voorkomen [Dating app Breeze may (and must) adjust algorithm to prevent discrimination]', https://www.mensenrechten.nl/actueel/nieuws/2023/09/06/dating-app-breeze-mag-en-moet-algoritme-aanpassen-om-discriminatie-te-voorkomen, 6 September 2023.

38 See Article 6 AI Act. Because of the broad scope of the term profiling, practitioners may want to first assess whether the AI system profiles natural persons, before looking at the exceptions.

fall outside of the scope of the definition.[39] The provider (who places the system on the market) and developer do not have to be the same entity.[40] The exception does not apply to the deployer, the user of the AI system.[41]

What does it mean to develop an AI system? The OECD's definition distinguishes between different stages in an AI system's lifecycle.[42]

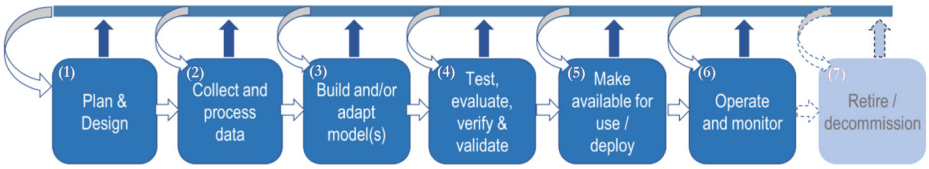


Figure 1: An AI System's lifecycle according to the OECD definition. The lifecycle has seven (cyclical) stage, numbered from 1 to 7 in the figure.

See the diagram depicted in Fig. 1.[43] As the diagram shows, an AI system's development consists of several stages (from the left): (1) planning and designing, (2) collecting and processing data, (3) building and/or adapting AI models, and (4) testing, evaluating, verifying and validating the AI model. (5) using/deploying and (6) operating and monitoring. It is possible to match the AI lifecycle with the definitions of provider and deployer: the provider is mostly responsible for the first four steps, while the deployer is responsible

39 The definition seems to cover all legal personhoods. An organisation renting an AI system from an AI developer is not a provider, unless the organisation also develops the AI system further. A student developing an AI system for personal use, without a name or trademark falls outside of the scope of 'provider'. This sounds logical from a product safety perspective: the moment an AI system is put on the market, its rules apply.

40 The AI Act includes an obligation for provider and supplier to co-operate. See Article 25(4) AI Act.

41 The deployer is 'a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity', Article 3(4) AI Act.

42 The definition of the term AI system in the AI act is based on the definition created by the international organisation OECD. See Organisation for Economic Co-operation and Development (OECD) 2023.

43 Organisation for Economic Co-operation and Development (OECD) 2024, fig. 3.2.

for steps 5 and 6. Where the provider *develops* the AI system and places the AI system on the market or into service, the deployer *uses* the AI system.[44]

The terms provider and deployer are not mutually exclusive: a deployer can (also) become a provider, and a provider can (also) become a deployer. For high-risk AI systems in particular, the EU legislator intended a flexible definition of deployer. Article 25(1) AI Act states that if any deployer or third party makes a *substantial modification* or *uses the AI system for another purpose*, the AI Act considers them a provider. In such a case, the original provider becomes a third-party supplier that must co-operate with the new provider.[45] The AI Act does not define what a substantial modification or modified purpose is.[46]

4.3 Bias detection and correction and the de-biasing obligation

According to Article 10(5) AI Act, providers may only use the exception for 'ensuring bias detection and correction [...] in accordance with paragraph (2), points (f) and (g)'. Let me cite the relevant part of Article 10 AI Act:[47]

> Article 10
>
> 2. Training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: [...]
> (f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations;
> (g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f); [...]

44 Article 3(4) AI Act.

45 Article 25 AI Act and Recital 60 AI Act. Article 25 AI Act does not provide for possible power imbalances between provider and supplier. See also Montinaro, *EJLPT* 2024/1.

46 Section 5 of the paper discusses what the lack of a clear distinction between provider and deployer means for the de-biasing obligation.

47 Lightly edited by author.

The AI Act's exception speaks of 'detecting and correcting' biases. Comparing these terms to points (f) and (g), detecting biases means an 'examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations'. Correcting biases means taking 'appropriate measures to detect, prevent and mitigate possible biases identified according to point (f)'. Both definitions seem fairly open-ended.[48]

The focus is on *possible biases* in points (f) and (g). Neither the AI Act nor its recitals define what biases are. Dictionaries give many possible definitions for the term *bias*, such as for example: (i) 'the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment.' (ii) 'the fact of preferring a particular subject or thing.' (iii) 'an unfair personal opinion that influences your judgment.'[49] In scholarship, the term bias does not have a fixed definition, although many taxonomies exist.[50]

Developers may come across many types of biases during the development and use of an AI system. I name a few possible scenarios (which are not necessarily bias in the language of AI developers):

- An error made in a calculation, which leads to different results than intended. As an example: the results of the German election in Saxony had to be corrected because of a calculation error. After a correction of the error, the party AfD lost the elections in Saxony.[51]

48 See the example in Section 2.1.

49 Cambridge dictionary, definition of ' bias', https://dictionary.cambridge.org/dictionary/english/bias. (accessed 4 February 2025).

50 See, for example, Barocas & Selbst, *Calif Law Rev* 2016/104. For a taxonomy, see Tilburg Institute for Law, Technology, and Society 2021.

51 See (in German) Correctiv, 'Softwarefehler bei der Landtagswahl in Sachsen: Was hinter der Korrektur der Sitzverteilung steckt [Software error in the state election in Saxony: What is behind the correction of the seat distribution]', https://correctiv.org/faktencheck/2024/09/06/softwarefehler-bei-der-landtagswahl-in-sachsen-was-hinter-der-korrektur-der-sitzverteilung-steckt/, 6 September 2024.

- The results of the AI system are biased. For example, men may be overrepresented compared to women in the datasets used to train the AI system.[52]
- The AI system itself seems neutral, but the user base is not. For example, a dating app may have fewer users with specific ethnicities, which results in a (possibly reinforced) underrepresentation of those users in the app.[53]

Typically, an AI developer will refer to 'bias' as a prejudice with respect to certain groups or individuals (the second example above).[54] The AI Act does not define what bias means: currently, all three of the above examples could be 'bias' in the sense of Article 10(2) AI Act.[55] In the context of this paper, which focuses on Article 10(5) AI Act, the second example (overrepresentation) makes the most sense.[56]

It follows from the introduction of Article 10(2) AI Act that detecting and correcting biases in datasets is not only possible for the provider, it is *mandatory*.[57] In Article 10(5) AI Act, the EU legislator explicitly refers to the bias detection and correction in accordance with Article 10(2)(f) and (g). Consequently, providers may only process sensitive data as far as strictly necessary to examine their *training, validation and testing datasets* for the bias test required by Article 10(2) AI Act. Currently, because there is no exact guidance on how to correct biases, the obligation could be broad or narrow: further guidance seems necessary.[58] The provider may not collect the data

52 For example, in 2018, Amazon scrapped an AI system that was apparently biased 'because Amazon's computer models were trained [...] by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry'. See Reuters, 'Amazon scraps secret AI recruiting tool that showed bias against women', https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G, 11 October 2018.

53 I discuss this example more in Section 5 (Discussion).

54 Barocas, Hardt & Narayanan 2023.

55 Article 10(2)(f) AI Act refers not only to bias that possibly lead to discrimination, but also bias that could affect the 'health and safety of persons, negative impact on fundamental rights or lead to discrimination', which could be much broader.

56 Article 10(5) AI Act aims to prevent discrimination. See recital 70 AI Act.

57 See also, for a paper focusing on the obligation itself, Hacker, *Law, Innovation and Technology* 2021/13.

58 See also Section 5 (Discussion).

to look for biases beyond the datasets used for development, such as biases encoded in the AI model or biases that originate from applying an AI system across a different context than the one for which it was originally designed.[59]

In sum, the meaning of the term 'bias' in the AI Act seems open to interpretation, but the exception focuses on biases in the datasets used to develop the AI system, a further limitation.

4.4 Strictly necessary

The AI Act's exception states that providers may only process personal data for as long as 'strictly necessary'. The CJEU uses the phrase 'strict necessity' in case law about the right to the protection of personal data. The case law by the CJEU is relevant for interpreting strict necessity in Article 10 AI Act: Article 10 AI Act implements Article 9 GDPR, and the GDPR implements the right to the protection of personal data.[60]

For a given measure to be a *strict necessity*, first, the measure must be limited to what is *necessary*. Necessity means that the measure is the least intrusive measure amongst the suitable measures for the achievement of the objective it pursues. In the *Schecke* case, the CJEU ruled that a measure was unnecessary because it was 'possible to envisage measures which affect less adversely the fundamental rights of natural persons and which still contribute effectively to the objectives [...] in question [...].'[61]

The CJEU sets a high threshold for the necessity test: the measure must be limited to what is *strictly* necessary in light of its goals. The word strict means that the measure must be carefully drafted and its goals must be well-defined. According to CJEU case law regarding data retention, this requires a high level of nuance and granularity. To illustrate, in the case *Digital Rights Ireland*, the CJEU ruled that the Data Retention Directive should have been '*precisely circumscribed* by provisions to ensure that it is actually limited to what is

59 Biases may also come from the context of use of the AI system. For possible sources of bias in AI systems, see Barocas & Selbst, *Calif Law Rev* 2016/104, p. 671.

60 Recital 1 GDPR.

61 CJEU 9 November 2010, C-92/09 and C-93/09 (*Volker und Markus Schecke GbR (C-92/09) and Hartmut Eifert (C-93/09) v Land Hessen*), para. 86.

strictly necessary.'[62] The strict necessity test is granular: the more precise and targeted a measure is, the higher the chance that the measure becomes the least restrictive option.[63]

The strict necessity test has three possible implications for providers applying the exception in Article 10(5) AI Act. First, providers will have to be precise in defining how they intend to use the sensitive data to remove biases from the dataset. This may be tricky for the purpose of AI de-biasing. While the AI Act states in Recital 70 that the exception aims to prevent discrimination, the AI Act itself does not clarify the link between bias and non-discrimination.[64] Preventing discrimination is not an exact science, unlike measuring bias. Methods for preventing discrimination are still not mature at the time of writing.[65] Providers may therefore have a tough time with the strict necessity test: it sets a high bar for a goal such as AI de-biasing. Providers will have to think carefully about which biases they aim to examine their datasets for, how this can prevent discrimination and how to carry out the examination in the least intrusive way. The state of the art in AI de-biasing, guidance by supervisory authorities and future standards to conduct the examination are also relevant in that assessment.[66]

Second, if suitable measures exist to remove the biases without using sensitive data, then those measures may be less intrusive. And it is preferable if a provider can process less data for removing the biases, for example by only using a sub-sample of the full dataset for testing and deleting the original dataset. Some techniques for measuring bias require less sensitive data than others.[67]

62 Emphasis added. CJEU 8 April 2014, C-293/12 and C-594 (*Digital Rights Ireland Ltd*), para. 65.

63 Dalla Corte, *International Data Privacy Law* 2022/12, p. 259, 270.

64 Non-discrimination law also does not yet play well with algorithmic discrimination. See e.g. Xenidis, *Maastricht Journal of European and Comparative Law* 2020/27.

65 See Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023.

66 See also Section 4.3 and for further discussion Section 5 (Discussion).

67 For example, methods using proxy variables, unsupervised learning models, causal fairness, or synthetic data. See also section 4.6 of the paper. See also Sebastiaan Berendsen & Emma Beauxis-Aussalet, 'Fairness versus Privacy: sensitive data is needed for bias detection', https://ucds.cs.vu.nl/fairness-versus-privacy-sensitive-data-is-needed-for-bias-detection/, 14 March 2024.

Finally, the provider must take into account the risk of unlawful access to the data. The longer the provider keeps the data for de-biasing, the bigger the risk of a data breach.

Overall, the strict necessity test seems to set a high bar for AI de- biasing. The future will tell how the test will be applied in AI de- biasing exactly.

4.5 Appropriate safeguards

While processing the sensitive data, providers must implement 'appropriate safeguards for the fundamental rights and freedoms of natural persons', which is an open norm.[68] Apart from this open norm, Article 10(5) AI Act includes a list of several conditions under Article 10 (5)(a) – (f) AI Act. For a discussion of Article 10(5)(a) AI Act, see Section 4.6.

The conditions in Article 10(5)(b)-(f) AI Act include security, privacy and accountability safeguards that the provider must implement. *Condition (b)* requires technical measures on the re-use of the data, and state- of-the-art security and privacy-preserving measures. *Condition (c)* requires confidentiality obligations with strict access controls. *Condition (d)* forbids that the data are '[. . .] accessed by other parties'. Presumably, the access controls and confidentiality organisations in condition (c) and other safeguards must guarantee that other parties cannot access the sensitive data. *Condition (e*) concerns data retention. *Condition (f)* concerns accountability in light of the GDPR: providers must keep a record of their processing activities, explain why processing the sensitive data is strictly necessary, and why de-biasing could not be achieved by processing other data. All of the conditions in Article 10(5)(b)-(f) overlap with the GDPR's data protection principles.[69]

The list of safeguards shows how the legislator seeks to balance data protection and non-discrimination law: the provider may process the data, but only under strict safeguards.[70] While the safeguards make sense, providers could consider more safeguards to make the exception the least intrusive and therefore *strictly necessary*. A trusted third party that already stores the data could perhaps share the data with provider, which means that

68 The open norm can be filled in with the GDPR, see Section 4.6.1.
69 See Section 4.6.1 for a comparison.
70 See Section 2.

the provider does not need to collect the data directly from data subjects. Condition (d) states that the sensitive data are not to be accessed by other parties: The exception does not allow providers to create, for example, entire data sharing platforms. However, it seems that the provider may still receive (anonymised) sensitive data *from* other parties that legitimately collected it, such as, for example, the national statistics bureaus.[71]

4.6 The GDPR still applies

According to Article 10(5) AI Act, the exception applies 'in addition to the provisions set out in [the GDPR].' We can also see this in one of the specific conditions in the exception: Article 10(5)(a) AI Act states that the exception does not apply if the provider can also fulfil the bias examination and detection effectively by processing other data, such as synthetic or anonymised data.[72] Anonymised data is non-personal, so the GDPR does not apply in that case. As the word *synthetic* suggests, synthetic data is essentially 'fake', or artificially generated data. In the context of AI de-biasing, synthetic data aims to strike a balance between privacy and usability: a dataset that is on the one hand untraceable to individuals, and on the other hand representative of the real world, so that it is still useful for detecting and correcting biases.[73] Data scientists can create representative synthetic data (either manually or by using AI) from a smaller sample of real sensitive data.[74]

Depending on whether synthetic data can be linked to an individual, synthetic data can be anonymous, and the GDPR does not apply.[75] In writing the AI Act, the EU legislator made a few 'slips of the pen' regarding synthetic data: Article 59(1)(b) AI Act explicitly names 'anonymised, synthetic or other non-personal data'. Article 10(5)(a) AI Act states that the exception

71 See Sebastiaan Berendsen & Emma Beauxis-Aussalet, 'Fairness versus Privacy: sensitive data is needed for bias detection', https://ucds.cs.vu.nl/fairness-versus-privacy-sensitive-data-is-needed-for-bias-detection/, 14 March 2024.

72 See also Ashurst & Weller, *Equity and Access in Algorithms, Mechanisms, and Optimization* 2023.

73 See Section 4.4. See also Bellovin, Dutta & Reitinger, *22 Stan. Tech. L. Rev. 1 (2019)*, p. 53.

74 Providers could create synthetic data with generative adversarial networks (GANs), for example. See, for a survey of methods, Figueira & Vaz, *Mathematics* 2022/10.

75 See the definition of personal data in Article 4(1) GDPR.

is unnecessary if 'synthetic or anonymised data' is effective. Because these provisions equate synthetic and anonymous data, it seems as if the AI Act grants synthetic data the same status as anonymous data in law.[76] However, apart from these breadcrumbs, the AI Act is silent on the matter: further exploration of the topic seems necessary, but goes beyond the scope of the paper.[77]

Even if a provider uses (non-personal) synthetic data, the provider may still need the exception to *generate* synthetic data in the first place. Providers can only create representative synthetic data if they have a (smaller) representative sample of real synthetic data. The exception also seems useful for providers creating synthetic data.[78]

4.6.1 Data protection principles

The data protection principles in Article 5 GDPR also apply to the processing of sensitive data. The principles include lawfulness (Article 5 (a) GDPR), purpose limitation (Article 5(b) GDPR), data minimisation (Article 5(c) GDPR), accuracy (Article 5(d) GDPR), storage limitation (Article 5(e) GDPR), integrity and confidentiality (Article 5(f) GDPR) and finally accountability (Article 5(2) GDPR). Section 4.6 discusses the principle of lawfulness.

The principle of *purpose limitation* requires that personal data must be 'collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes'. Providers must clearly specify the exact purpose of the processing: which de-biasing techniques will they use, and why is sensitive data necessary for the technique? We can interpret the purpose limitation principle together with the term *strictly necessary*, which also requires a specific definition of the purposes of the processing and reflection on whether they are legitimate.[79]

76 For example, at first glance, 'the intention of the EU Legislator to equalize anonymous and synthetic data [...] appears undeniable.' See (informally) Lorenzo Christofaro - Legal status of Synthetic Data – 2023, https://www.linkedin.com/pulse/legal-status-synthetic-data-lorenzo-cristofaro.

77 See e.g. López, Augusto & Elbi 2022. Gat & Lynskey, *Iowa Law Review* 2024/109, p. 1087 and further.

78 For recommendations to organisations working with synthetic data, see e.g. Payal Arora and others 2024.

79 See also Section 4.4 on strict necessity.

The *data minimisation* principle implies that the provider must use as little data as possible for the de-biasing. In many cases, the provider could use a *subsample* of the full dataset for the de-biasing, rather than the data of all data subjects. Such a smaller sample of data has similar statistical properties as the full dataset: both must be representative.[80]

According to the *data accuracy* principle, personal data the provider gathers must not be incorrect. At first glance, for the purpose of detecting and correcting biases, the data accuracy principle seems inherently necessary: without accurate data, the bias detection and correction process would not be effective. The data must be representative. It is unclear what kind of accuracy is necessary for AI de-biasing with non-discrimination as the goal. Scholars have long debated whether non-discrimination attributes, such as ethnicity, must be self-reported by the data subject (sometimes called self-identification) or based on a predefined set of criteria.[81] There is something to be said for both approaches: ethnicity, for example, is an ambiguous term with different meanings across countries or even different contexts.[82] Depending on the definition, the ethnicity could be more or less accurate for the data subject or the AI de-biasing. As far as I know, best practices on this issue do not exist at the time of writing, but both approaches seem compatible with human rights.[83]

The *storage limitation* principle states that the provider must keep the sensitive data 'in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed'.[84] When the provider has complied with the de-biasing obligation, the provider must delete the data, unless the provider has another compatible

80 The statistical technique for finding the minimal sample is called 'power analysis'. For an example, see e.g. Yu, Ai & Ye, *Statistical Papers* 2024/65, p. 467.

81 See e.g. Ringelheim, *SSRN Electronic Journal* 2007. AI practitioners who wish to de-bias their AI also struggle with this issue. See Andrus and others, *arXiv:2011.02282 [cs]* 2021/6.1.

82 The CJEU has stated that 'Ethnic origin cannot be determined on the basis of a single criterion but, on the contrary, is based on a whole number of factors, some objective and others subjective.' CJEU 6 April 2017, C-668/15, ECLI:EU:C:2017:278 (*Jyske Finance*), para. 19. See also Jaime & Kern, *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 237-253.

83 Ringelheim, *SSRN Electronic Journal* 2007.

84 Article 5(1)(e) GDPR.

purpose. The provider must choose a retention period that is not longer than strictly necessary for the purpose of the processing.[85] Article 10(5)(e) AI Act rephrases the storage limitation principle quite literally: the special categories of personal data must be 'deleted once the bias has been corrected or the personal data has reached the end of its retention period, whichever comes first'.

The principle of *integrity and confidentiality* requires appropriate security of the personal data, in the form of technical and organisational measures. Article 10(5) AI Act under (b), (c), (d) and (e) give many examples of measures. Depending on the context in which the de-biasing takes place, more safeguards, such as a trusted third party, could be appropriate.[86]

Finally, the principle of *accountability* states that the providers must demonstrate compliance with all the other data protection principles. Read together with the phrase *strictly necessary*, the principle of accountability requires the provider to keep a record stating the exact purpose of the processing, explaining why the processing is truly necessary, and why the chosen measure is the *least intrusive*. Article 10 (5)(f) AI Act explicitly states a similar reading.[87]

The data protection principles can serve as a safety net in specific contexts where some of the AI Act's safeguards are not enough. The data protection principles are a flexible guide for the provider taking appropriate measures, on top of the list in Article 10(5) AI Act, which the context of the de-biasing may require.

4.6.2 Lawful processing under the GDPR

The principle of lawfulness in Article 5(a) GDPR also applies. Even if a provider (here, the data controller) complies with the requirements in Article 10(5) AI Act, the provider is not out of the woods yet. To process the data, the provider must have a lawful basis for the processing under Article 6 GDPR. As I mentioned before, gathering valid consent (sub a) may be difficult in

85 This also overlaps with the meaning of strictly necessary, see Section 4.4. The provider may still create synthetic or anonymised data, as mentioned in the introduction to this section.

86 See Section 4.5.

87 Compare with Section 4.5.

many situations, because of the power asymmetries between data subject and data controller.[88]

A second candidate for a processing ground is Article 6(1)(c) GDPR juncto Article 10(2) AI Act. The AI Act states that providers must examine their datasets for biases. That is a concrete legal obligation. The legislator created Article 6(1)(c) GDPR for both the private and the public sector. An example from practice: banks regularly apply Article 6 (1)(c) GDPR for fraud detection (as mandated by the law). The legal obligation must be directly present in a Union or member state law, which seems the case for Article 10 AI Act.[89]

Another ground for processing may be the public interest, Article 6 (1)(e) GDPR, because de-biasing AI systems aims to protect people's health and safety, non-discrimination, fundamental rights. However, it seems the EU legislator originally intended that ground mainly for the public sector, not the private sector. It is unclear if a private organisation can fulfil a public interest in the sense of Article 6(1)(e) GDPR, because such organisations also have a private interest.[90]

Finally, providers could invoke ground (f): the legitimate interest ground, which the GDPR legislator originally intended only for private sector controllers.[91] In a recent case, the CJEU repeated and clarified the test for what constitutes a legitimate interest.[92] Protecting fundamental rights (such as non-discrimination) is a legitimate interest.[93] A legitimate interest is invalid if the data subject could not reasonably have expected the processing: data subjects must be informed and given a chance to object.[94] For AI de-biasing, it may be difficult for the provider to inform the data subject about the (intended) de-biasing. And if too many data subjects object, the sensitive data may become non- representative and less useful for AI de-biasing. In

88 See Section 2.

89 Kuner and others 2020/6.A Rationale and Policy underpinnings.

90 Kuner and others 2020/6.A Rationale and Policy underpinnings.

91 Kuner and others 2020/6.A Rationale and Policy underpinnings.

92 CJEU October 2024, C-621/22, ECLI:EU:C:2024:857 (*Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens (KNLT v AP)*).

93 CJEU October 2024, C-621/22, ECLI:EU:C:2024:857 (*Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens (KNLT v AP)*), para. 38-40.

94 See Recital 47 GDPR and CJEU October 2024, C-621/22, ECLI:EU:C:2024:857 (*Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens (KNLT v AP)*), para. 55.

sum, the most plausible ground (for both sectors) seems to be Article 6(1)(c) GDPR, the legal obligation.

5 Discussion: is the exception effective?

The de-biasing exception and obligation target discrimination. Biases present in the data used to develop AI systems can (ultimately) have discriminatory effects on society, once the AI system has been trained and the system takes a decision. Preventing discriminatory effects of AI systems seems possible.[95] Non-discrimination law could play an important role here, even if the link between measuring bias and weighing whether a case constitutes discrimination is currently often not clear.[96]

The exception is not perfect, however. Three limitations of the exception require further discussion: (i) only providers may process sensitive data, (ii) for biases in the training, validation and testing datasets, (iii) the AI act does not describe which biases developers must remove.

(i) The exception only applies for the provider of the AI system, and the developer may not share the sensitive data with deployers, which has consequences for the effectiveness of the exception. Non- discrimination law, is a highly contextual field.[97] I distinguish between three main cases of AI development and deployment: *AI as a service*, in-house development of AI, and further development of an existing AI system.[98] First, the provider

95 Biases may ultimately also come from the context in which the AI system is deployed, rather than the data used to train the AI. However, removing biases from datasets can be a good start.

96 See Wachter, Mittelstadt & Russell, *Computer Law & Security Review* 2021/41. In the context of fairness metrics, it is often unclear how measuring bias and weighing discrimination interrelate. See e.g. Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023. The problem also occurs for data scientists who try to build discrimination-aware models. See for Discrimination-aware data mining: Berendt & Preibusch, *IEEE 12th International Conference on Data Mining Workshops* 2012. Fairness-aware machine learning: Pastaltzidis and others, *2022 ACM Conference on Fairness, Accountability, and Transparency* 2022, p. 2302-2314.

97 Veale, Van Kleek & Binns, *CHI Conference on Human Factors in Computing Systems* 2018, p. 10.

98 In the EU, some sectors seem to have more in-house development of AI than others. See Hoffreumon, Forman & Van Zeebroeck, *Journal of Economics & Management Strategy* 2024/33, p. 452.

and the deployer of the AI system may be entirely separate parties: the deployer rents the AI, or in other words, uses *AI as a service (AIaaS)*. Multiple deployers could use an AI system created by one provider in different regions of the world, and across contexts. For example, an employment system developed by a provider in Greece could be rented to an employer (the deployer) in The Netherlands. In such cases, without further context about the setting and target group of the system, the developer cannot take away potential discriminatory effects effectively.[99] The developer and deployers must then collaborate. It seems questionable whether a single AI system can facilitate, for example, hiring in multiple countries: the developer may have to fine-tune the AI system for each individual country, in collaboration with deployers.[100] The AI Act does not require deployers of AI systems to collaborate with providers.

Second, the provider and deployer could be the same organisation. The provider then has the necessary information for a bias examination: the sensitive data, and the information about how the *in-house* AI system is or will be deployed.

Third, sometimes, there is a grey area where a deployer becomes a provider. The deployer can become a provider when making a substantial modification, but the AI Act leaves open what substantial means. Take, as an example, a small insurance company that uses an AI system originally developed by a large AI firm, but the insurance company fine-tunes the system to target its own user-base and datasets. It is currently unclear if the insurer becomes a provider from such fine-tuning.[101] If the deployer becomes a provider, the deployer may have to remove possible biases from the data used for fine-tuning, and apply the AI Act's exception. And perhaps the provider of the original model is a 'supplier' of a high-risk AI system, and must co-operate with the fine-tuner.[102] From the point of view of de-biasing,

99 Data scientists have made this point before. See Lewicki and others, *CHI Conference on Human Factors in Computing Systems* 2023/6.

100 For example, the meaning of 'ethnicity' varies by country. See Jaime & Kern, *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 237-253.

101 See Section 4.2, definition of provider.

102 See Article 25(4) AI Act.

it makes sense for the new provider to also de-bias the AI system: changes the new provider made may have introduced discriminatory effects.[103]

(ii) A second limitation of the exception is that providers must only de-bias their training, validation and testing datasets. However, biases that have discriminatory effects may originate not only from the data used to develop the AI system, but also from choices the developer makes in creating the algorithm itself and the context in which the algorithm is deployed. I give an example. A Dutch dating app ran into a problem when correcting biases. The app used a technique to recommend people based on likes from similar people (called *collaborative filtering*). It is unclear how the developer could mitigate the bias or where the bias originates from exactly: fixing the dataset may not truly prevent discrimination, because the biases may come from the user base itself or the algorithm. Depending on the source of the bias, an advertising campaign to attract more users from certain ethnicities could be effective. Second, the developer could change the way in which the app shows users to each other (the algorithm). Both solutions do not mitigate biases resulting from the *dataset* used to develop the algorithm: they seem to be out of the scope of the exception.[104] I add that, because dating apps are not high-risk, the exception does not apply in this example anyway.[105] Limiting the bias examination to datasets seems like a conscious choice by the EU legislator: the legislator wanted to prevent bias monitoring, preventing providers from continuously processing of sensitive data to de-bias AI through its lifecycle.[106]

(iii) A final limitation is that Article 10 AI Act does not clearly state which biases providers must address and how. A recent competition involving interdisciplinary teams attempted to implement the AI Act's obligations for real-world AI systems. There is something to be said for not requiring AI providers to check for specific biases in all datasets: on the one hand, de-biasing is highly contextual, so a general obligation makes sense. On the other hand, a too generic approach 'poses the risk of missing critical

103 Biases can occur both upstream and downstream in the development of AI. See for an example Steed and others, *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2022.

104 See for a discussion of this example De Jonge & Borgesius 2024..

105 See Section 4.1 of the paper.

106 See Section 4.2 and 4.4.

vulnerabilities; domain-specific knowledge is required to identify, measure, and tackle specific biases.'[107] The European Commission and standardization organisations will have to find a middle ground between these two views. Without further guidance, especially smaller providers may lack the knowledge necessary to effectively remove discriminatory effects from their AI systems.

Overall, Article 10 AI Act could contribute to less discriminatory AI systems. Supervisory authorities could clarify which biases developers must remove to comply with the de-biasing obligation. Or perhaps a harmonised standard could clarify the de-biasing obligation: the European Commission has already requested a standard for 'specifications for appropriate data governance and data management procedures [...] with specific focus on [...] procedures for detecting and addressing biases and potential for proxy discrimination [...].'[108] For situations in which provider and deployer are entirely separate parties, cooperation between provider and deployer seems essential to make Article 10 AI Act more effective. I think such a cooperation could be an appropriate safeguard.

6 Conclusion

This paper discussed the new exception allowing providers to process sensitive data for AI de-biasing (enshrined in Article 10(5) AI Act), and highlighted some of its implications.

In two influential ways, the legislator limited the scope of Article 10 (5) AI Act: (i) by writing the exception for *bias detection and correction* and leaving out *bias monitoring*, the legislator prevents the provider from monitoring biases once the AI system is finished. The focus on *bias detection and correction* prevents sensitive data from being used continuously, without data retention, and preserves necessity. (ii) The exception only applies to providers of AI

107 Scantamburlo and others, *2nd International Workshop on Responsible AI Engineering* 2024.

108 See Annex 1 of European Commission, C(2023)3215 – Standardisation Request M/593. COMMISSION IMPLEMENTING DECISION of 22.5.2023 on a Standardisation Request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in Support of Union Policy on Artificial Intelligence (2023) https://ec.europa.eu/transparency/documents-register/detail?ref=C(2023)3215&lang=en.

systems, rather than third parties, limiting the number of parties that can store and have access to the data.

The legislator's choices result in an exception that is not a silver bullet for practitioners. Especially the two limitations in scope restrict the exception's usefulness: the exception essentially cannot be used for bias monitoring during actual use. Especially for providers of *AI as a service*, the exception seems less useful, because deployers may not collect and use sensitive data for de-biasing. Still, for providers, these limitations do not make the exception useless: providers who wish to test their datasets during the development of the AI system can still apply the exception.

Perhaps many past scandals concerning AI could have been prevented if developers had the sensitive data necessary to carefully examine their datasets. Critics of the exception should not forget that the exception essentially balances the fundamental rights to data protection and non-discrimination. The exception is one of the first attempts of any legislator worldwide to balance these two rights in the context AI de- biasing, which should be applauded.

The AI Act is long and detailed. In the coming years, different professionals must work with the AI Act's rules: government employees, consultants, auditors, researchers, lawyers, developers. . . the list goes ever on. Because the AI Act is complex, many practitioners struggle to determine the implications of the law's obligations. That is no surprise, because even for legal specialists, the AI Act is a complex law. Hopefully, this paper can give some support to practitioners working under the burden of the AI Act's de-biasing obligation and exception.

References of chapter IV

AlgorithmAudit 2024
AlgorithmAudit, *Preventing prejudice. Recommendations for risk profiling in the College Grant Control process: a quantitative and qualitative analysis*, Education Executive Agency of The Netherlands (DUO) 2024, https://algorithmaudit.eu/pdf-files/algoprudence/TA_AA202401/TA_AA202401_Preventing_prejudice.pdf.

Andrus and others, *arXiv:2011.02282 [cs]* 2021
M. Andrus and others, ' "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness', *arXiv:2011.02282 [cs]* 2021, http://arxiv.org/abs/2011.02282.

Ashurst & Weller, *Equity and Access in Algorithms, Mechanisms, and Optimization* 2023
C. Ashurst & A. Weller, 'Fairness Without Demographic Data: A Survey of Approaches', *Equity and Access in Algorithms, Mechanisms, and Optimization* 2023, https://dl.acm.org/doi/10.1145/3617694.3623234, doi:10.1145/3617694.3623234.

Barocas, Hardt & Narayanan 2023
S. Barocas, M. Hardt & A. Narayanan, *Fairness and Machine Learning*, MIT Press 2023, https://fairmlbook.org/.

Barocas & Selbst, *Calif Law Rev* 2016/104
S. Barocas & A. Selbst, 'Big Data's disparate impact', *Calif Law Rev* 2016/104, issue 671, https://www.jstor.org/stable/24758720.

Bellovin, Dutta & Reitinger, *22 Stan. Tech. L. Rev. 1 (2019)*
S.M. Bellovin, P.K. Dutta & N. Reitinger, 'Privacy and Synthetic Datasets', *22 Stan. Tech. L. Rev. 1 (2019)*, https://law.stanford.edu/wp-content/uploads/2019/01/Bellovin_20190129.pdf.

Berendt & Preibusch, *IEEE 12th International Conference on Data Mining Workshops* 2012
B. Berendt & S. Preibusch, 'Exploring Discrimination: A User-centric Evaluation of Discrimination-Aware Data Mining', *IEEE 12th International*

Conference on Data Mining Workshops 2012, http://ieeexplore.ieee.org/document/6406461/, doi:10.1109/ICDMW.2012.109.

Boone, *Journal of Data Protection & Privacy* 2023/6
T.S. Boone, 'The challenge of defining artificial intelligence in the EU AI Act', *Journal of Data Protection & Privacy* 2023/6, issue 2, p. 180-195, https://ideas.repec.org/a/aza/jdpp00/y2023v6i2p180-195.html.

Chaudhary and others 2023
B. Chaudhary and others, 'Practical Bias Mitigation through Proxy Sensitive Attribute Label Generation', 2023, http://arxiv.org/abs/2312.15994, doi:10.48550/arXiv.2312.15994.

Dalla Corte, *International Data Privacy Law* 2022/12
L. Dalla Corte, 'On proportionality in the data protection jurisprudence of the CJEU', *International Data Privacy Law* 2022/12, issue 4, p. 259-275, https://academic.oup.com/idpl/article/12/4/259/6647961, doi:10.1093/idpl/ipac014.

Enqvist, *Information & Communications Technology Law* 2024, p. 1-25
L. Enqvist, 'Rule-based versus AI-driven benefits allocation: GDPR and AIA legal implications and challenges for automation in public social security administration', *Information & Communications Technology Law* 2024, p. 1-25, https://www.tandfonline.com/doi/full/10.1080/13600834.2024.2349835, doi:10.1080/13600834.2024.2349835.

European Commission 2019
European Commission, *COM 2019/168 Building Trust in Human-Centric Artificial Intelligence*, 2019, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0168.

Fernández-Llorca and others, *Artificial Intelligence and Law* 2024
D. Fernández-Llorca and others, 'An interdisciplinary account of the terminological choices by EU policymakers ahead of the final agreement on the AI Act: AI system, general purpose AI system, foundation model, and generative AI', *Artificial Intelligence and Law* 2024, https://link.springer.com/10.1007/s10506-024-09412-y, doi:10.1007/s10506-024-09412-y.

Figueira & Vaz, *Mathematics* 2022/10
A. Figueira & B. Vaz, 'Survey on Synthetic Data Generation, Evaluation

Methods and GANs', *Mathematics* 2022/10, issue 15, https://www.mdpi.com/2227-7390/10/15/2733, doi:10.3390/math10152733.

Gat & Lynskey, *Iowa Law Review* 2024/109
M.S. Gat & O. Lynskey, 'Synthetic Data: Legal Implications of the Data-Generation Revolution', *Iowa Law Review* 2024/109, p. 1087 and further.

Hacker, *Law, Innovation and Technology* 2021/13
P. Hacker, 'A legal framework for AI training data—from first principles to the Artificial Intelligence Act', *Law, Innovation and Technology* 2021/13, issue 2, p. 257-301, https://www.tandfonline.com/doi/full/10.1080/17579961.2021.1977219, doi:10.1080/17579961.2021.1977219.

Hacker, *SSRN Electronic Journal* 2024
P. Hacker, 'Comments on the Final Trilogue Version of the AI Act [Preprint]', *SSRN Electronic Journal* 2024, https://www.ssrn.com/abstract=4757603, doi:10.2139/ssrn.4757603.

Hoffreumon, Forman & Van Zeebroeck, *Journal of Economics & Management Strategy* 2024/33
C. Hoffreumon, C. Forman & N. Van Zeebroeck, 'Make or buy your artificial intelligence? Complementarities in technology sourcing', *Journal of Economics & Management Strategy* 2024/33, issue 2, p. 452-479, https://onlinelibrary.wiley.com/doi/10.1111/jems.12586, doi:10.1111/jems.12586.

Jaime & Kern, *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 237-253
S. Jaime & C. Kern, 'Ethnic Classifications in Algorithmic Fairness: Concepts, Measures and Implications in Practice', *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 237-253, https://dl.acm.org/doi/10.1145/3630106.3658902, doi:10.1145/3630106.3658902.

De Jonge & Borgesius 2024
T. de Jonge & F.Z. Borgesius, 'Mitigating Digital Discrimination in Dating Apps – The Dutch Breeze case [PREPRINT]', 2024, https://arxiv.org/abs/2409.15828.

Kuner and others 2020
C. Kuner and others, *The EU General Data Protection Regulation (GDPR): A Commentary*, New York: Oxford University Press 2020, https:

//oxford.universitypressscholarship.com/10.1093/oso/9780198826491.001.0001/isbn-9780198826491, doi:10.1093/oso/9780198826491.001.0001.

Lewicki and others, *CHI Conference on Human Factors in Computing Systems* 2023
K. Lewicki and others, 'Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service"', *CHI Conference on Human Factors in Computing Systems* 2023, https://dl.acm.org/doi/10.1145/3544548.3581463, doi:10.1145/3544548.3581463.

López, Augusto & Elbi 2022
F. López, C. Augusto & A. Elbi, *On synthetic data: a brief introduction for data protection law dummies*, KU Leuven 2022, https://lirias.kuleuven.be/3838501&lang=en.

McLoughney and others, *IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* 2023
A.J. McLoughney and others, '"Emerging proxies" in information-rich machine learning: a threat to fairness?', *IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* 2023, https://ieeexplore.ieee.org/document/10155045/, doi:10.1109/ETHICS57328.2023.10155045.

Misztal-Radecka & Indurkhya, *Information Processing & Management* 2021/58
J. Misztal-Radecka & B. Indurkhya, 'Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems', *Information Processing & Management* 2021/58, issue 3, p. 102519, https://www.sciencedirect.com/science/article/pii/S0306457321000285, doi:10.1016/j.ipm.2021.102519.

Montinaro, *EJLPT* 2024/1
R. Montinaro, 'Responsible data sharing for AI: a test bench for EU Data Law', *EJLPT* 2024/1, https://universitypress.unisob.na.it/ojs/index.php/ejplt/article/view/1979.

OECD 2023
OECD, *Explanatory memorandum on the updated OECD definition of an AI system* (OECD Artificial Intelligence Papers, OECD Artificial Intelligence Papers), 2023, https://www.oecd-ilibrary.org/science-and-technology/explanatory-

memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en, doi:10.1787/623da898-en.

Organisation for Economic Co-operation and Development (OECD) 2023
Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, 2023, https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449.

Organisation for Economic Co-operation and Development (OECD) 2024
Organisation for Economic Co-operation and Development (OECD), *Meeting of the Council at Ministerial Level, 2-3 May 2024. Report on the implementation of the OECD recommendation on Artificial Intelligence, C/MIN(2024)17*, 2024, https://one.oecd.org/document/C/MIN(2024)17/en/pdf.

OSTP 2022
OSTP, *Blueprint for an AI Bill of Rights. Making automated systems work for the American people*, The White house | OSTP 2022, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Pastaltzidis and others, *2022 ACM Conference on Fairness, Accountability, and Transparency* 2022, p. 2302-2314
I. Pastaltzidis and others, 'Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems', *2022 ACM Conference on Fairness, Accountability, and Transparency* 2022, p. 2302-2314, https://dl.acm.org/doi/10.1145/3531146.3534644, doi:10.1145/3531146.3534644.

Payal Arora and others 2024
Payal Arora and others, *Recommendations on the Use of Synthetic Data to Train AI Models*, NU Centre, UNU-CPR, UNU Macau 2024, https://unu.edu/publication/recommendations-use-synthetic-data-train-ai-models.

Ringelheim, *SSRN Electronic Journal* 2007
J. Ringelheim, 'Processing Data on Racial or Ethnic Origin for Antidiscrimination Policies: How to Reconcile the Promotion of Equality with the Right to Privacy?', *SSRN Electronic Journal* 2007, http://www.ssrn.com/abstract=983685, doi:10.2139/ssrn.983685.

Ruschemeier, *ERA Forum* 2023/23
H. Ruschemeier, 'AI as a challenge for legal regulation – the scope of

application of the artificial intelligence act proposal', *ERA Forum* 2023/23, issue 3, p. 361-376, https://link.springer.com/10.1007/s12027-022-00725-6, doi:10.1007/s12027-022-00725-6.

Scantamburlo and others, *2nd International Workshop on Responsible AI Engineering* 2024
T. Scantamburlo and others, 'Software Systems Compliance with the AI Act: Lessons Learned from an International Challenge', *2nd International Workshop on Responsible AI Engineering* 2024, https://dl.acm.org/doi/10.1145/3643691.3648589, doi:10.1145/3643691.3648589.

Steed and others, *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2022
R. Steed and others, 'Upstream Mitigation Is Not All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models', *60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 2022, https://aclanthology.org/2022.acl-long.247, doi:10.18653/v1/2022.acl-long.247.

Tilburg Institute for Law, Technology, and Society 2021
Tilburg Institute for Law, Technology, and Society, *Handbook on non-discriminating algorithms. Summary research report*, 2021, https://www.tilburguniversity.edu/about/schools/law/departments/tilt/research/handbook.

Van Bekkum & Zuiderveen Borgesius, *Computer Law & Security Review* 2023/48
M. Van Bekkum & F. Zuiderveen Borgesius, 'Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?', *Computer Law & Security Review* 2023/48, p. 105770, https://linkinghub.elsevier.com/retrieve/pii/S0267364922001133, doi:10.1016/j.clsr.2022.105770.

Veale, Van Kleek & Binns, *CHI Conference on Human Factors in Computing Systems* 2018
M. Veale, M. Van Kleek & R. Binns, 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making', *CHI Conference on Human Factors in Computing Systems* 2018,https://dl.acm.org/doi/10.1145/3173574.3174014, doi:10.1145/3173574.3174014.

Wachter, Mittelstadt & Russell, *Computer Law & Security Review* 2021/41
S. Wachter, B. Mittelstadt & C. Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI', *Computer Law & Security Review* 2021/41, p. 105567, https://linkinghub.elsevier.com/retrieve/pii/S0267364921000406, doi:10.1016/j.clsr.2021.105567.

Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023
H. Weerts and others, 'Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree', *ACM Conference on Fairness, Accountability, and Transparency* 2023, https://dl.acm.org/doi/10.1145/3593013.3594044, doi:10.1145/3593013.3594044.

Xenidis, *Maastricht Journal of European and Comparative Law* 2020/27
R. Xenidis, 'Tuning EU equality law to algorithmic discrimination: Three pathways to resilience', *Maastricht Journal of European and Comparative Law* 2020/27, issue 6, p. 736-758, https://journals.sagepub.com/doi/10.1177/1023263X20982173, doi:10.1177/1023263X20982173.

Yu, Ai & Ye, *Statistical Papers* 2024/65
J. Yu, M. Ai & Z. Ye, 'A review on design inspired subsampling for big data', *Statistical Papers* 2024/65, issue 2, p. 467-510, https://link.springer.com/10.1007/s00362-022-01386-w, doi:10.1007/s00362-022-01386-w.

Zhang, Lemoine & Mitchell 2018
B.H. Zhang, B. Lemoine & M. Mitchell, 'Mitigating Unwanted Biases with Adversarial Learning', 2018, http://arxiv.org/abs/1801.07593, doi:10.48550/arXiv.1801.07593.

Žliobaitė & Custers, *Artificial Intelligence and Law* 2016/24
I. Žliobaitė & B. Custers, 'Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models', *Artificial Intelligence and Law* 2016/24, issue 2, p. 183-201, http://link.springer.com/10.1007/s10506-016-9182-5, doi:10.1007/s10506-016-9182-5.

Zuiderveen Borgesius and others 2024
F. Zuiderveen Borgesius and others, *Non-discrimination law in Europe: a primer for non-lawyers [PREPRINT]*, 2024, https://arxiv.org/abs/2404.08519.

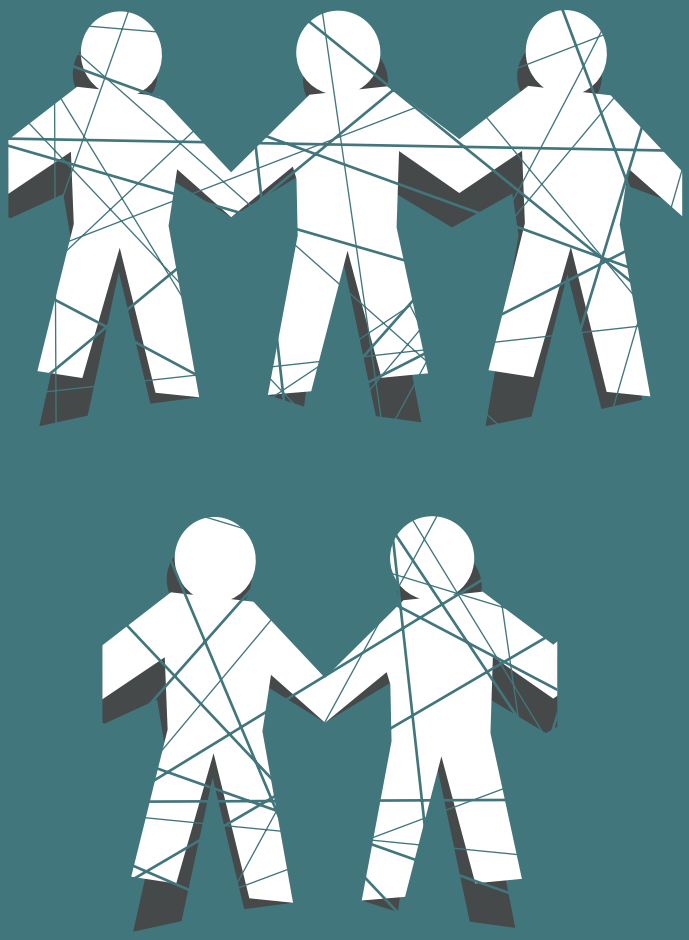
CJEU 9 November 2010, C-92/09 and C-93/09 (*Volker und Markus Schecke GbR (C-92/09) and Hartmut Eifert (C-93/09) v Land Hessen*).

CJEU 8 April 2014, C-293/12 and C-594 (*Digital Rights Ireland Ltd*).

CJEU 6 April 2017, C-668/15, ECLI:EU:C:2017:278 (*Jyske Finance*).

CJEU October 2024, C-621/22, ECLI:EU:C:2024:857 (*Koninklijke Nederlandse Lawn Tennisbond v Autoriteit Persoonsgegevens (KNLT v AP)*).

'Softwarefehler bei der Landtagswahl in Sachsen: Was hinter der Korrektur der Sitzverteilung steckt [Software error in the state election in Saxony: What is behind the correction of the seat distribution]'
Correctiv, 'Softwarefehler bei der Landtagswahl in Sachsen: Was hinter der Korrektur der Sitzverteilung steckt [Software error in the state election in Saxony: What is behind the correction of the seat distribution]', https://correctiv.org/faktencheck/2024/09/06/softwarefehler-bei-der-landtagswahl-in-sachsen-was-hinter-der-korrektur-der-sitzverteilung-steckt/, 6 September 2024.



V.

Discrimination and AI in insurance: what do people find fair? Results from a survey

Submitted to the *European Journal of Risk Regulation* as:

F. Zuiderveen Borgesius, M. van Bekkum, G. Schaap & I. van Ooijen, 'Discrimination and AI in insurance: what do people find fair? Results from a survey', *European Journal of Risk Regulation* 2026/17, 289–311, https://doi.org/10.1017/err.2025.10057.

Abstract

Two modern trends in insurance are data-intensive underwriting and behaviour-based insurance. Data-intensive underwriting means that insurers analyse more data for estimating the claim cost of a consumer and for determining the premium based on that estimation. Insurers also offer behaviour-based insurance. For example, some car insurers use artificial intelligence (AI) to follow the driving behaviour of an individual consumer in real-time and decide whether to offer that consumer a discount. In this paper, we report on a survey of the Dutch population (N=999) in which we asked people's opinions about examples of data-intensive underwriting and behaviour-based insurance. The main results include: (i) If survey respondents find an insurance practice unfair, they also find the practice unacceptable. (ii) Respondents find almost all modern insurance practices that we described unfair. (iii) Respondents find practices for which they can influence the premium fairer. (iv) If respondents find a certain consumer characteristic illogical for basing the premium on, then respondents find using the characteristic unfair. (v) Respondents find it unfair if an insurer offers an insurance product only to a specific group. (vi) Respondents find it unfair if an insurance practice leads to the poor paying more. We also reflect on the policy implications of the findings.

1 Introduction

Insurers underwrite risks by analysing data. Insurers increasingly use Artificial Intelligence (AI): based on correlations found in data, the insurer estimates the expected claims cost of the consumer and assigns the consumer to a risk class. Two modern trends in insurance are (i) data-intensive underwriting and (ii) behaviour-based insurance. (i) Data-intensive underwriting means that insurers use and analyse more data for risk underwriting. Insurers could extend their analyses with machine learning. For example, insurers could collect more data about consumers and use machine learning to find new correlations in data to underwrite risks with. Some of those correlations could look odd to the consumer, such as the correlation between a consumer's house number and the probability that they file an insurance claim.

(ii) A second trend is behaviour-based insurance: insurers can analyse the behaviour of an individual consumer to offer that consumer a discount. For example, some insurers offer a discount to the consumer if that consumer allows the insurer to monitor their driving behaviour (behaviour-based car insurance). Other insurers offer discounts to health or life insurance when consumers show, with a health tracker, that they have an active lifestyle.[1]

Both modern insurance trends can have advantages. For instance, careful consumers or consumers who run less risk may be able to insure themselves for a lower premium.[2] But both trends may also have negative effects. For example, the insurer could, unintentionally, illegally discriminate certain groups. Other effects could be controversial or unfair, even if the insurer does not discriminate illegally. This paper focuses on such possibly unfair – but not necessarily illegal – effects.

While there is much research discussing these trends, less is known about what the public thinks about data-intensive and behaviour-based decisions in the insurance sector. We performed a survey of the Dutch population (N=999) in which we asked respondents for their opinions about examples of (i) data-intensive underwriting and (ii) behaviour-based insurance. In general, we

1 See section 2.

2 However, section 2 shows that the trends may also lead to more expensive insurance for some, and to exclusion.

wanted to know whether, and under which circumstances, the general public finds the practices fair and acceptable. The paper explores the following questions.

1. To what extent do people find data-intensive underwriting and behaviour-based insurance fair and acceptable?
2. If people feel they have more influence on the premium, do they find an insurance practice fairer and more acceptable?
3. If people find an insurance practice more logical, do they find the practice fairer and more acceptable?
4. If an insurer only accepts certain parts of the population, do people find such practices less fair and acceptable?
5. If an insurance practice leads to higher prices for poorer consumers, do people find the practice less fair and less acceptable?

The paper makes the following contributions to the literature. First, this is one of the first academic papers to research what the public finds fair in the context of modern applications of AI in insurance.[3] This is an important addition to the literature, as fairness is always context-specific.[4] Second, we are the first to ask respondents about specific types of insurance fairness. For example, we explore what respondents think about insurers basing the premium on consumer characteristics that may seem illogical, or irrelevant, to the respondent.

Our results can be useful for several groups of readers. For example, the results could inspire scholars or others interested in perceived fairness of AI and data use, as the survey gives an impression of the attitudes of the average citizen for real-life examples. Such surveys are scarce: the perspectives of

3 See for a rare example of another survey about people's opinions on modern insurance practices (in the US, in that paper): Kiviat, *Sociological Science* 2021/8. Binns et al. investigate in experimental studies "people's perceptions of justice in algorithmic decision-making under different scenarios and explanation styles', including in the context of insurance. Binns and others 2018. For an overview of empirical literature until 2022 on people's fairness perceptions regarding algorithmic decision-making, see Starke and others, *Big Data & Society* 2022/9.

4 Starke and others, *Big Data & Society* 2022/9, p. 10.

insurance consumers are "generally an under-researched area".[5] Second, insurers who use AI in their business, or plan to do so, may want to know what respondents think of modern insurance trends.[6] Third, policymakers who consider regulating AI in insurance could use the survey as a first impression of what the public thinks.[7] We do not claim that policymakers should always prohibit practices that seem unfair according to surveys.[8] The results could be of use in other sectors outside of insurance too, as insurance is an example of a sector that has relied on statistics and AI for a relatively long period of time.[9]

We aim to avoid unnecessary jargon, so that the paper is understandable for readers without a background in insurance or actuarial science. We conducted the survey in Europe, and we discuss some EU law in the discussion section. The paper's results can be useful outside the EU too, for example, by providing ideas for further research in other countries, allowing transnational applications of AI, or by enabling more (public) discussion about the topic.[10]

5 Tanninen, *Big Data & Society* 2020/7, p. 10.

6 Charpentier & Vamparys, *Big Data & Society* 2025/12, p. 7-10.

7 Examples of recent attempts at regulating AI are the US bill of rights, *Blueprint for an AI Bill of Rights. Making Automated Systems Work for the American People*, The White house | OSTP 2022, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf and the EU AI Act Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) 2024.

8 See the discussion section.

9 Scholars in other sectors seem to see data-driven decision making as possibly disruptive. See Clarissa Valli Buttow, 'Data-Driven Policy Making and Its Impacts on Regulation: A Study of the OECD Vision in the Light of Data Critical Studies' (2025) 16 European Journal of Risk Regulation 114 <https://www.cambridge.org/core/product/identifier/S1867299X24000734/type/journal_article. Insurers have decades (some maybe centuries) of experience with using large amounts of data. Insurers have decades (some maybe centuries) of experience with using large amounts of data. Some scholars see insurance as "an insightful analogon for the social situatedness and impact of machine learning systems" Fröhlich & Williamson, *FAccT* 2024. See also the work by Barry and Charpentier Barry & Charpentier, *Ethics and Information Technology* 2023/25.

10 See also the suggestions for further research at the end of the paper.

In Section 2, building on earlier work, we provide background to our survey questions. Section 3 discusses the method for the survey. Section 4 presents the results, with each subsection discussing a different research question. In section 5 we reflect on the results and place them in broader context, such as whether the law and fairness metrics can help against the perceived unfairness. Section 6 discusses limitations of the research and provides suggestions for further research. Section 7 concludes.

2 Background: AI in insurance and possibly unfair differentiation

The core business of insurers is underwriting risks: estimating the expected claims cost of a consumer via a risk assessment.[11] Insurers always discriminate, in the neutral sense, between groups of people. Many insurers use a form of AI to underwrite risks or to analyse past data. AI can be described as "the science and engineering of making intelligent machines, especially intelligent computer programs".[12] For more than a century, insurers have gathered data directly from the consumers, such as data about a consumer's age, smoking, or drinking habits. Insurers used questionnaires to gather consumer characteristics such as the type of car they drive.[13]

Now, insurers could analyse more data gathered from many different sources. Insurers seem captivated by two trends (i) First, insurers could analyse more and new types of data to assess risks more precisely: data-intensive underwriting. (ii) Second, insurers could the behaviour of individual consumers in real-time: behaviour-based insurance. For example, some car insurers offer a discount if the consumer agrees to being tracked by the insurer and drives safely. AI makes both types of practices easier to implement for insurers.[14]

11 The Geneva Association (Noordhoek) 2023.
12 John McCarthy, 'What is AI? Basic Questions', http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.
13 Barry & Charpentier, *Big Data & Society* 2020/7, p. 3-6.
14 See for more details on the two trends: Chapter II.

2.1 Data-intensive underwriting

The first trend is data-intensive underwriting: insurers use and analyse more data for underwriting. With machine learning, insurers can find new correlations in data with which the insurer could more accurately predict the expected claim costs of certain consumer groups. Insurers could analyse new datasets to find many new correlations in data.[15]

Data-intensive underwriting comes with potential discrimination-related risks. A legal risk of data-intensive underwriting is that insurers could, accidentally, discriminate indirectly in an illegal way (disparate impact). For example, if an insurer uses a newly found correlation to set premiums, the insurer could inadvertently cause harm to certain ethnic groups, or other groups with legally protected characteristics.[16] There are other possibly unfair, or at least controversial, effects of data-intensive underwriting. Our paper focuses on those effects.

First, consumers may be confronted with premiums based on characteristics that the consumer has little influence on. For example, in the Netherlands, some insurers charged different car insurance premiums based on the consumer's house number. A consumer paid more if their house number included a letter: for instance, 10A or 40B.[17] Consumers may feel that they cannot reasonably influence their house number. Moving to another house to lower one's car insurance premium is hardly a serious option. Consumers can influence other characteristics more easily, such as the type of car they buy.

Second, insurers could use certain consumer characteristics to set premiums, while the consumer does not see the logic of using the characteristic.[18] In other words, the insurer could use seemingly illogical or irrelevant characteristics to set the premium. AI systems are good at finding correlations

15 Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019.

16 European Commission. Directorate General for Justice and Consumers. and others 2022. McLoughney and others, *IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* 2023.

17 Consumentenbond, 'Premies verzekeringen verschillen tot op huisnummer ['Insurance premiums differ by house number']', https://www.consumentenbond.nl/inboedelverzekering/verzekeringspremies-verschillen-tot-op-huisnummer, 2015.

18 Barry & Charpentier, *Ethics and Information Technology* 2023/25, p. 49.

in large datasets. When insurers underwrite using more data, they might find many correlations in datasets that look strange to consumers. In theory an insurer could find a correlation between on the one hand the chance that a consumer files a claim, and on the other hand whether a consumer has an address with an even number, is born in a certain month, or "people who spend more than 50% of their days on streets starting with the letter J".[19]

Third, insurers could introduce insurance products that are only intended for specific groups. Suppose, for example, that according to an insurer's analysis, people with a higher education submit fewer claims to their car insurer. The insurer could decide to offer car insurance only to a highly educated people. To illustrate, Promovendum, a Dutch insurer, markets its insurances as being only for the high-educated.[20] The insurer explains its reasoning: "Based on our years of experience and our claims statistics, we know that this target group has fewer claims. This enables us to offer our customers insurance products with low premiums and excellent terms and conditions".[21] Another insurer offers a car insurance only to medical doctors.[22] Such insurance practices can be controversial because the insurer targets specific groups of the population and thereby excludes the rest.

Fourth, insurance practices could reinforce inequalities that already exist in society, such as the difference between rich and poor. Such an effect could occur if an insurer (on purpose or by accident) charges higher prices to poorer people. In sum, data-intensive underwriting could have several negative effects.

2.2 Behaviour-based insurance

A second trend in insurance is behaviour-based insurance: i*nsurers can adapt the premium* to individual consumers in real-time, *based on how th*e *consumer*

19 O'Neil 2016, p. 138. See also Krippner & Hirschman, *Theory and Society* 2022/51.

20 The insurer does not directly refuse low-educated consumers requesting insurance. However, in advertisements and on the website, the insurer presents itself as being only for the higher-educated.

21 Dutch text translated to English by the authors. Promovendum, 'Higher educated ['Hoger opgeleiden']', https://www.promovendum.nl/hoger-opgeleiden.

22 Axxellence, 'De Artsen-AutoVerzekering van Axxellence nu extra voordelig! ['Axxellence's Physician Car Insurance now extra cheap!']', https://www.axxellence.nl/auto.html, 2025.

behaves. For example, some car insurers offer a discount if the consumer agrees to being tracked by the insurer (with a device in the car) and drives safely.[23] A life insurer could give discounts to people who show, with a health tracker or smart watch, that they lead an active life.[24] Insurers can improve their behaviour-based insurance with AI. For example, with a linear modelling approach, an insurer can consider characteristics such as weather and light conditions, mileage, time, and driving habits such as speeding to assess the consumer's driving behaviour.[25]

The most important difference between data-intensive underwriting and behaviour-based insurance is that with behaviour-based insurance, the insurer bases the premium on the behaviour of an *individual* consumer.[26] The insurer monitors, for instance, how the consumer drives, or how active a consumer is every day, measured by a health tracker or step counter. Another difference is that with behaviour-based insurance, the insurer focuses more on preventing risks. For example, insurers could (at least in theory) prevent accidents by nudging risky drivers to drive safer.

behaviour-based insurance comes with partly similar risks as data-intensive underwriting. For example, some people may be excluded from behaviour-based insurance. For bad drivers, behaviour-based car insurance might be so expensive that, practically speaking, they are excluded from car insurance. In a hypothetical market where all car insurers offer behaviour-based insurance,

23 Meyers & Hoyweghen, *Big Data & Society* 2020/7.

24 For a practical example, see The Vitality Group, 'The Vitality Difference', http://www.vitalitygroup.com/why-vitality/ accessed 23 July 2025. For a discussion of the case study, see Bednarz, Lewis & Sadowski, ' 'It's not personal, it's strictly business': Behavioural insurance and the impacts of non-personal data on individuals, groups and societies', *Computer Law & Security Review* 2025/56, s. 2. For a table with examples, see A Spender and others, 'Wearables and the Internet of Things: Considerations for the Life and Health Insurance Industry' (2019) 24 British Actuarial Journal e22 <https://www.cambridge.org/core/product/identifier/S1357321719000072/type/journal_article, Table 4. See on behaviour-based health insurance also also Meyers 2018.

25 Tselentis, Yannis & Vlahogianni, *Accident Analysis & Prevention* 2017/98, p. 141. For a laymen explanation of linear models in AI, see for example K. Gross, 'Machine Learning and Linear Models: How They Work (In Plain English)', https://blog.dataiku.com/top-machine-learning-algorithms-how-they-work-in-plain-english-1, 2020.

26 Insurers do not merely look at individuals when pricing behaviour-based insurance; they also look at groups. See for a discussion: Moor & Lury, *Journal of Cultural Economy* 2018/11.

bad drivers may not have the possibility to insure themselves for an affordable price.[27]

Second, behaviour-based insurance could reinforce financial inequality. Suppose for example, that on average, people with higher salaries have more time for sports or an active lifestyle. If richer people receive more discounts (based on their health trackers), they pay less. If richer people pay less, financial inequality is reinforced.

In conclusion, both data-intensive underwriting and behaviour-based insurance could have several possible negative effects. We do not claim that all these effects will occur; we merely highlight possible effects. We also note that insurers could combine both trends: insurers can aggregate individual behavioural data, and use the data for more data-intensive underwriting.[28] Insurers could then use that profit for more behaviour-based insurance.[29] Nevertheless, because of their partly different effects, we present the two trends separately.

3 Method

3.1 Survey design and sample

We conducted an online survey about data-intensive underwriting and behaviour-based insurance among 1,000 Dutch citizens of 18 years and older, using a sample provided by panel service Panelclix in the Netherlands. In total, 1,102 respondents completed the questionnaire. We excluded respondents who completed the questionnaire in under 4 minutes (n=96), and respondents failing both attention checks (n=6).[30] One respondent

27 Dutch Financial Authority (AFM) 2021. H Rubin, Aas & Williams, *Journal of Information Technology Teaching Cases* 2023/13.

28 Henckaerts & Antonio, *Insurance: Mathematics and Economics* 2022/105. Moosavi & Ramnath, *Accident Analysis & Prevention* 2023/192. In her book, O'Neil warns that "oceans of behavioural data will feed straight into artificial intelligence systems" O'Neil 2016, p. 139.

29 See, for example, recent work by Minty about the structural change of insurance and the chosen wording in recent debates D. Minty, 'Insurance and the Importance of Words and Language', https://www.ethicsandinsurance.info/insurance-and-the-importance-of-words-and-language/, 10 April 2025.

30 Muszyński, *Ask: Research and Methods* 2023/32.

requested to have their data removed after data collection; therefore, our final sample was N=999.

The final sample consisted of 51.6% female, 46.6% male, and 1.7% other/do not want to say. Mean age was 49.82 (SD=17.13). Of the sample 30.5% was highly educated, 51.7% received mid-level education, and 17.7% received primary or lower-level secondary education (Dutch population: 36.4, 37.0, and 26% respectively; CBS, 2024). Net monthly income was as follows: 6.9% earned less than €1,000, 35.6% between €1,000 and €2,500, 35.4% between €2,500 and €5,000, and 4,8% over €5,000 (net modal monthly income in the Netherlands in 2024 was €2.692; CPB, 2024). Hence, the sample consisted of a diverse demographic in terms of age, gender, education, and income.

Respondents filled out a questionnaire on "practices in the insurance sector". The median completion time was around 8 minutes. Respondents were compensated for their participation. The study complied with the criteria of Ethics Committee of the authors' institution (reference no. ECSW-LT-2025-1-10-82717). All data and materials are available at OSF.[31]

3.2 Procedure

Before respondents started the questionnaire, we asked them to give their informed consent. Subsequently, we asked respondents for their age. Next, respondents indicated their attitudes towards several short scenarios regarding data-intensive underwriting and behaviour-based insurance, related to several insurance domains (e.g., car, life, or burglary insurance).

We included two attention checks in the questionnaire by means of instructed response items. The first check was an embedded item, instructing the respondents to "click fully agree here". A second check instructed respondents to click "green" in a list of colours.[32] Finally, respondents answered questions on socio-demographics. After completion, respondents were thanked and redirected to Panelclix collect their compensation. Respondents were paid 0,75 Euro for their participation.

31 https://osf.io/x2fdy/?view_only=344f176daa404ec6929a21d576c4e159 accessed 25 July 2025.

32 Muszyński, *Ask: Research and Methods* 2023/32.

3.3 Questionnaire

We presented respondents with various scenarios about different data-intensive insurance practices. Respondents read a short introduction, after which they were presented with several one-sentence examples of practices. For example, in one section ("block") of the survey we told respondents:

> *"This block is about how insurance companies determine the insurance premium (how much you pay each month). We want to know your opinion on this.*
>
> *An insurance company (insurer) offers car insurance. The insurer calculates the monthly premium based on predictions of the number of claims for damages a consumer will make. The insurer makes these predictions based on information it has collected previously about the consumer, such as age, address, and type and year of the car. Below we present several examples of how an insurer can determine the premium. We ask for your opinion on each example.*
>
> *Example 1: "The insurer determines the premium based on the postal code of the consumer."*

We wanted to know whether respondents found the practice that we described fair and acceptable. In other words, the two main outcome variables were Fairness and Acceptance. We measured *Fairness* by the statement: "I think this practice is fair". Respondents indicated their agreement on a 5-point Likert scale: (1) completely disagree; (2) disagree; (3) neutral; (4) agree; (5) completely agree. Likewise, for *Acceptance*, respondents indicated their agreement with the statement: "I think this practice is acceptable".

We also wanted to know whether respondents felt that a consumer could influence the premium, and whether they saw the logic behind the insurer using that characteristic to calculate the premium. Hence, additional variables were Perceived Influence and Perceived Logic. For *Influence* respondents indicated on a 5-point scale to what extent they agreed with the statement: "This way, the policyholder can influence the insurance premium". For *Logic*, respondents indicated to what extent they agreed with the statement "I find it logical that the insurer determines the premium

like this". We asked the respondents the questions in Dutch. If we quote a question in this paper, we translate the question to English.

4 Results

In this section, we present the results from the survey. Each subsection discusses a different research question of the paper. We present the results in text and figures. We include the full data in the tables in the Appendices.

4.1 Do people find (i) data-intensive underwriting and (ii) behaviour-based insurance fair and acceptable?

The first question is: do respondents find data-intensive underwriting and behaviour-based insurance fair and acceptable? A simple t-test compared the mean difference in fairness and acceptance between behaviour-based insurance practices vs. practices based on data-intensive data types. Here, two data types classified as sources for behaviour-based insurance, namely car tracker data ("car tracker") and data based on a step counter ("steps"), whereas the other data types classified as sources for data-intensive underwriting. The t-test indicated that behaviour-based insurance practices overall score higher than data-intensive data types for insurance in terms of fairness ($t(998) = 21.58$, $p < .001$ (Mean Difference = .58)), and acceptance ($t(998) = 18.26$, $p < .001$ (Mean Difference = .51)).

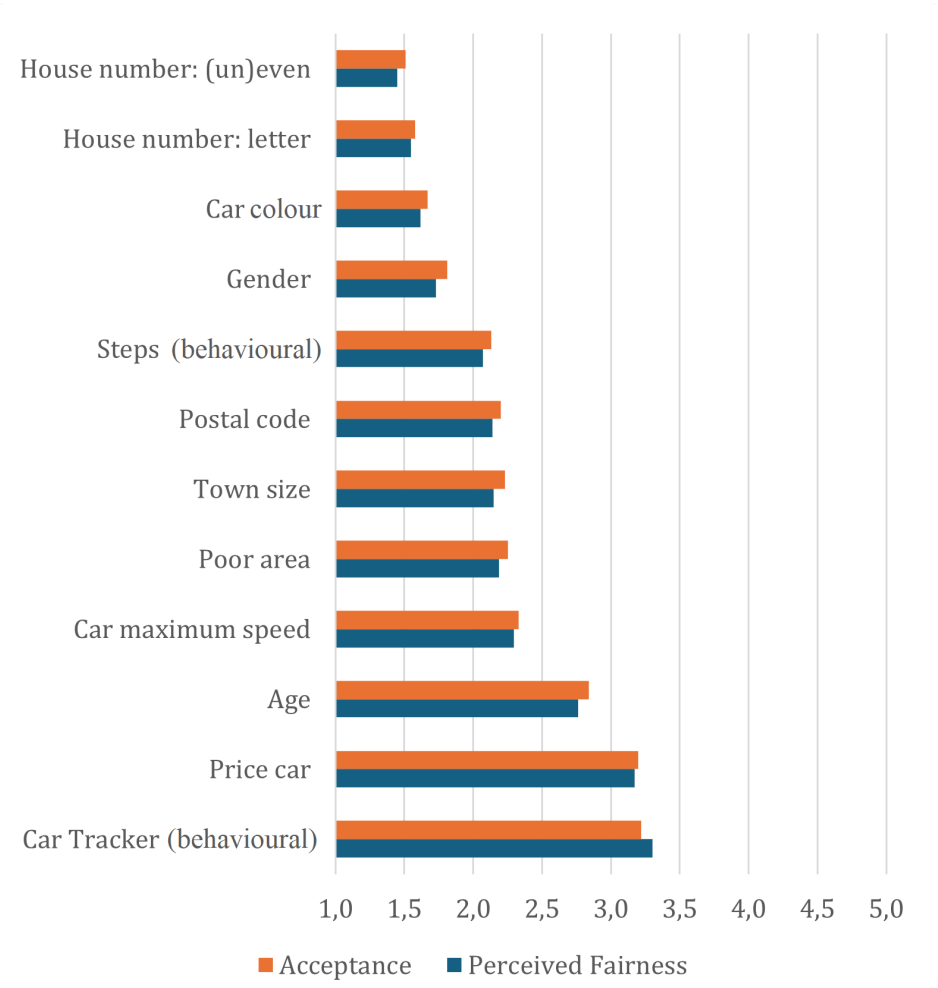


Figure 1: Public opinion of fairness and acceptance in insurance per data type.

As Figure 1 shows, respondents found it only (relatively) fair (i.e., scoring above the scale midpoint of 3) if an insurer adapts the premium based on the consumer's driving behaviour, which is an example of behaviour-

based insurance, and based on the maximum speed of the car, which is an example of data-intensive underwriting.[33] Respondents found it unfair and unacceptable if an insurer bases the premium on all the other types of data, including behaviour-based insurance based on a step counter (i.e., scoring below the scale midpoint of 3). Hence, the significant difference between behaviour-based insurance and data-intensive underwriting seems to be driven entirely by the higher fairness and acceptance values for the car tracker, and not by behaviour-based insurance with a step counter. In conclusion, respondents find almost all modern insurance practices that we described unfair. See Appendix A for an overview of significance levels of deviations from the scale's neutral mid-point for each data type.

4.2 If people feel they have more influence on the premium, do they find an insurance practice fairer and more acceptable?

As noted, insurers could calculate premiums based on characteristics that consumers cannot reasonably influence (see section 2). We asked respondents whether they felt that they could influence the premium if the insurer calculated the premium based on the following characteristics: the consumer's gender, age, or postal code, the car's price, colour, or maximum speed, the size of the city where the consumer lives, whether the consumer lived at an odd or even house number, and whether the consumer lived at a house number with a letter extension.

We also asked respondents whether they find it fair and acceptable if an insurer calculates the premium based on such characteristics.

We asked two questions about life insurance (we include the introductions in Appendix D). In the first example, the insurer charges higher prices to poorer people. We asked respondents whether they felt that the consumer can influence the premium in such a case. The question is based on a real-life example of an insurer charging higher premiums in poor neighbourhoods on purpose.[34]

33 See Section 3.3 (Questionnaire) for the description of the scenario. See also appendix D for life and burglary insurance.

34 College voor de Rechten van de Mens 2014.

We asked a second question about this life insurance example, but with an extra explanation: "In the Netherlands, people with a migration background are on average poorer than those without a migration background. This insurer's practice results in people with a migration background paying a higher premium on average than those without a migration background." (The example is based on a real-life insurance practice.[35]) Again, we asked respondents whether they thought that the consumer can influence the insurance premium.

We asked a similar question about burglary insurance.

We also asked respondents whether they felt that they could influence the insurance premium for two examples of behaviour-based insurance: behaviour-based car insurance (with a type of tracker in the car), and life insurance that gives a discount if the consumer walks more than 10.000 steps a day, as measured by a step counter.

35 College voor de Rechten van de Mens 2014.

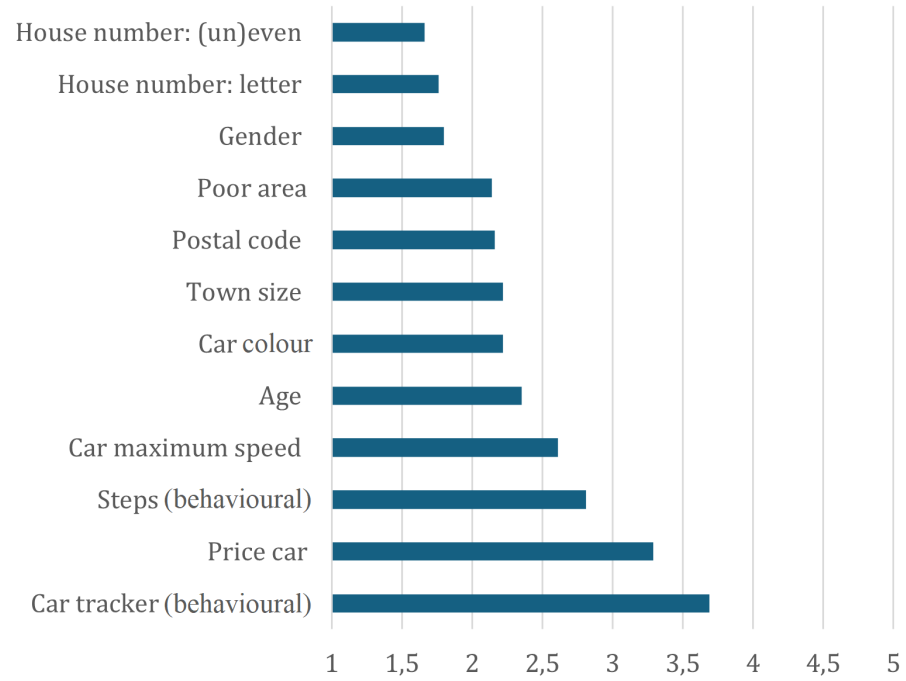


Figure 2: Influence perceptions by data type.

Figure 2 shows the influence perceptions across all data types. For most data types, respondents feel that they can hardly influence the insurance premium. A one-sample t-test showed all values significantly deviated from the scale's midpoint (3; See Appendix B for significance levels for each data type). Most practices deviate negatively from this neutral point (where respondents find the practices neither highly nor little influenceable; Figure 2; See also Appendix B). This means that for most practices, respondents feel that the consumer cannot really influence the premium. However, respondents feel that consumers have some influence on the premium in the case of behaviour-based insurance based on a car tracker and in the case of the price of the car they buy. Respondents do not feel that consumers have

much influence on the premium in the case of behaviour-based insurance with a step counter.

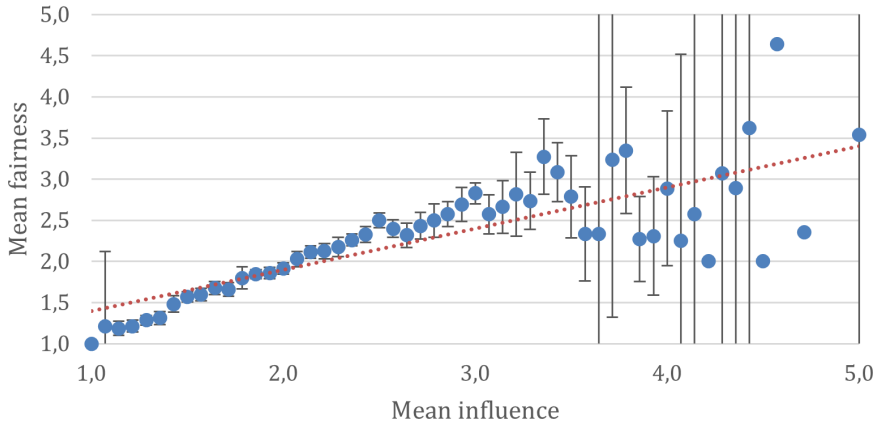


Figure 3: The relationship between perceived influence and fairness across all cases.

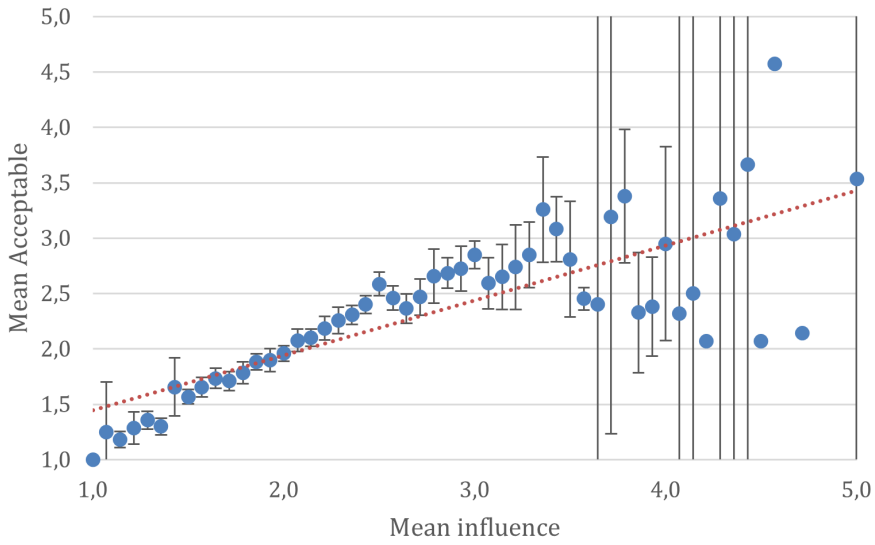


Figure 4: The relationship between perceived influence and acceptance across all cases below.

We found a positive relationship between perceived influence and fairness and acceptance. Figures 3 and 4 give a graphic representation of the relationship between, on the one hand, perceived influence, and on the other hand perceived fairness (Figure 3) and acceptance (Figure 4). Regression analysis confirms the pattern that is shown in Figure 3, indicating that there is a positive relationship between perceived influence and the perceived fairness of the insurance practice across all cases (R^2 = .54, F(1, 997) = 1147.52, β = .732, p < .001). A similar conclusion can be drawn regarding acceptance. Namely, as Figure 4 suggests, there is a positive relationship between perceived influence and perceived acceptance of the insurance practice (R^2 = .53, F(1, 997) = 1126.68, β = .728, p < .001).

In addition, Figures 3 and 4 suggest that if respondents feel that they have little influence on the insurance premium, then almost all respondents find the practice unfair (Figure 3) and unacceptable (Figure 4). However, if

respondents find that they have more influence on the premium, then some find the practice fairer than others; hence, respondents are more divided. We can see this in the error bar in the figure, which is more volatile, or varied, for data types for which respondents feel that they have more influence on the premium.[36]

4.3 If people find an insurance practice more logical, do they find the practice fairer and more acceptable?

We asked respondents whether they saw the logic of certain insurance practices (See Section 2 about data-intensive underwriting). For various insurance practices, we asked respondents whether they found it logical that the insurer uses certain consumer characteristics to set the insurance premium. We also asked respondents whether they find it fair and acceptable if an insurer calculates the premium based on such practices.

36 Note. The scatterplot and regression line represent the relationship between perceived influence and acceptance of all cases of insurance practices (i.e., data-intensive underwriting, behaviour-based practices and practices with indirect effects). The error bar is the vertical line attached to a dot in the figure. The error bar represents a 95% confidence interval. A 95% confidence interval means that there is only a 5% chance that the true value is NOT included within the span of the error bar.

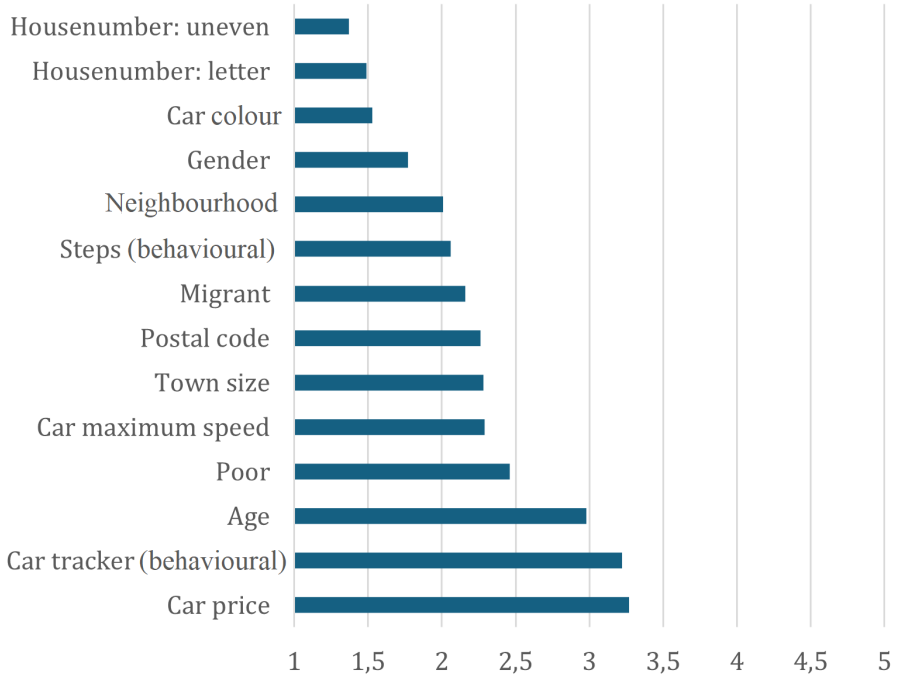


Figure 5: Perceived logic over all insurance practices.

In Figure 5, we show the perceived logic for all insurance practices. Respondents considered a situation where the insurer uses the price of the insured car as the most logical. Respondents considered a situation where an insurer uses the type of house number to calculate the premium for car insurance the least logical (i.e., an uneven house number and, to a slightly lesser extent, a house number with a letter).[37] Compared to the scales' neutral midpoint of 3, respondents only considered situations in which car tracker data, or the price of a car, determined insurance fees as relatively logical (p's <.001). Respondents considered all other situations as relatively illogical (p's

37 See Section 3.3 (Questionnaire) for the description of the scenario. See also appendix D for life and burglary insurance.

<.001), except for age, which the respondents considered neutral (p = .293; see Appendix B2 for all individual effects per insurance practice).

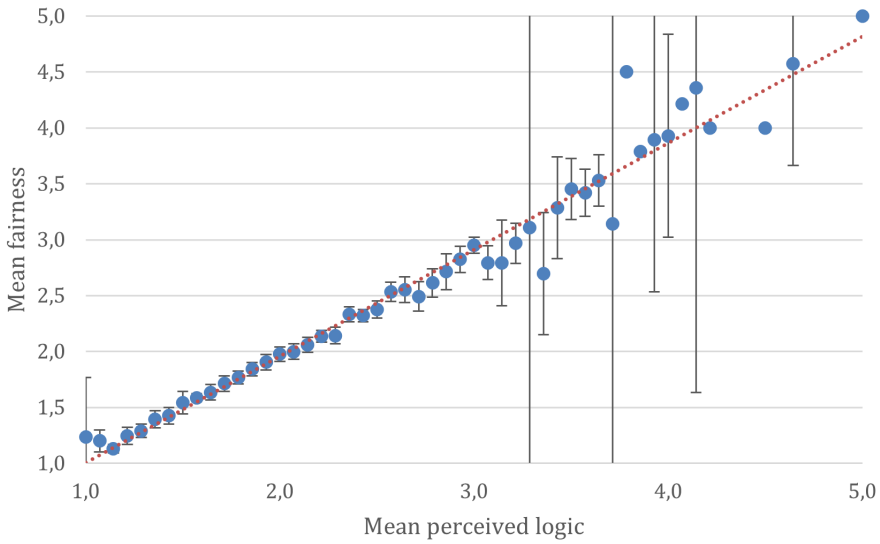


Figure 6: The relationship between perceived logic and fairness across all insurance practices.

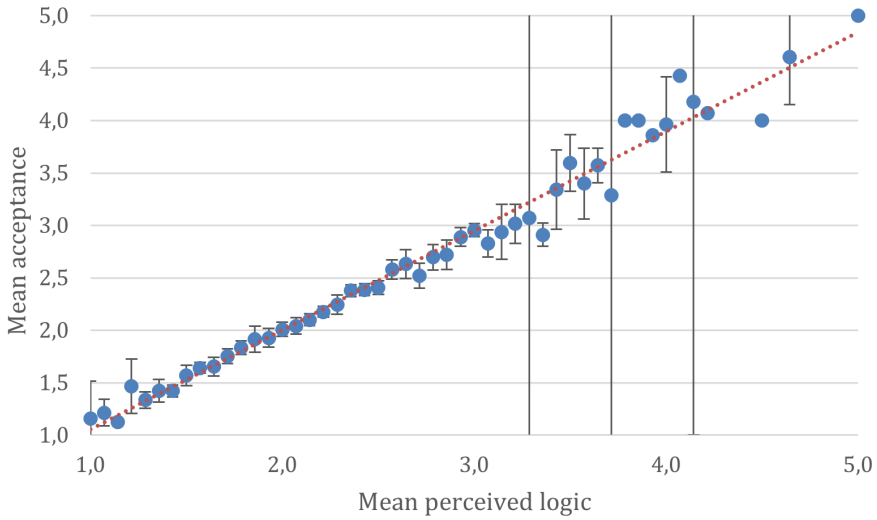


Figure 7: The relationship between perceived logic and acceptance across all insurance practices.

We found a strong positive relationship between perceived logic and fairness. The more logical respondents found a certain insurance practice, the fairer they thought the practice was (R^2 = .84, F(1, 997) = 5142.06, β = .915, p < .001). Similarly, we found a positive relationship between perceived logic and acceptance (R^2 = .83, F(1, 997) = 4948.89, β = .912, p < .001, see also Figure 6 and Figure 7). The error bars in the figures show that while respondents generally agree on the unfairness and unacceptability for examples that they find less logical, respondents disagree more often for more logical examples. In Section 4.3, we described a similar pattern for influence and fairness/acceptability. In sum, if respondents find an insurance practice more logical, they find the practice fairer and more acceptable.

4.4 If an insurer only accepts certain parts of the population, do people find such practices less fair and acceptable?

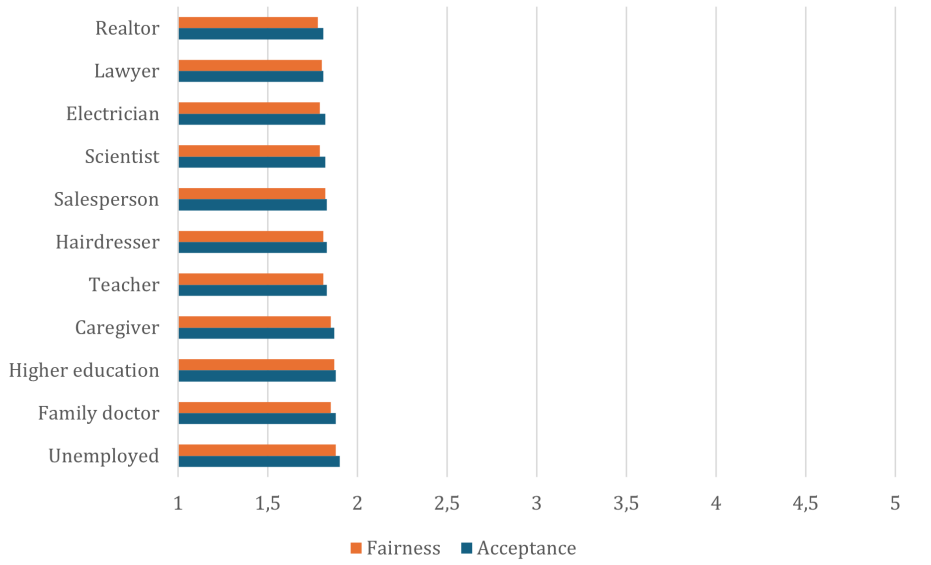


Figure 8: Respondents find target group insurance unfair and unacceptable.

We asked whether respondents find it fair and acceptable if insurers only accept specific groups (see section 2). On average, respondents find all target group insurance completely unfair and unacceptable (all *p*'s <.001; see Figure 8; see also Appendix C for the individual significance levels of deviations from the scale's midpoint). It hardly matters for respondents who the target group is. And, when they assess whether they find a specific target group insurance fair and acceptable, it seems that respondents find it irrelevant whether the target group has a low income (caregiver, unemployed) or a high income (lawyer, family doctor).

4.5 If an insurance practice leads to higher prices for poorer consumers, do people find the practice less fair and less acceptable?

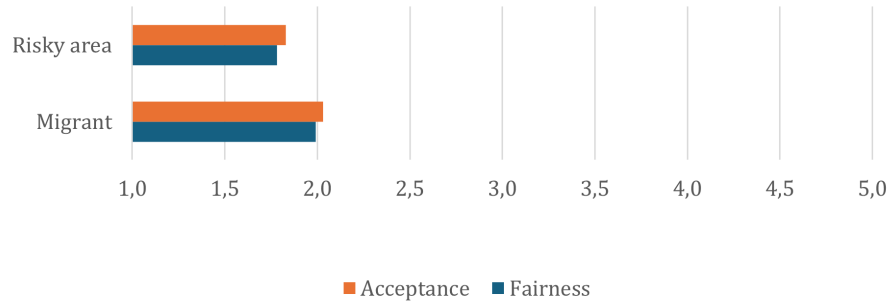


Figure 9: Public opinion of fairness and acceptance when poor population is indirect target group on the basis of....

Respondents find it unfair when insurance practices result in higher insurance premiums for poor people (Figure 9). We asked a question about an insurer who charges higher premiums in neighbourhoods where more burglaries occur. We added: "This leads to poorer people paying higher premiums on average." (The full question is quoted in section 4.2). Hence, in this question, the fact that poorer people pay higher prices is an accidental effect of an insurance practice. The insurer did not charge higher premiums to poorer people on purpose, according to the question.

Regardless, respondents find the burglary insurance practice unfair and unacceptable ($t(998) = -41.880$, Mean Difference = -1.22, and unacceptable, $t(998) = -38.27$, Mean Difference = -1.17). See figure 9. So, at least in this example, respondents find it unfair and unacceptable if an insurance practice leads to higher prices for poorer consumers.

We also asked a question in which a life insurer charged a higher premium to poorer consumers on purpose. The reasoning of the insurer is that poorer people live shorter, and thus pay a premium for fewer years (see section 4.2 for the full question).

We asked a second question about life insurance that was largely the same, but included some extra information: "In the Netherlands, people with a migration background are on average poorer than people without a migration background. This insurer's practice results in people with a migration background paying a higher premium on average than those without a migration background." In a similar vein, respondents find it unfair if people with an immigrant background pay higher prices ($t(998) = -29.19$, $p < .001$, Mean Difference = -1.01), and unacceptable ($t(998) = -27.78$, Mean Difference = -.97). Figure 9 shows the results. All in all, respondents find it unfair and unacceptable if poorer consumers pay more for insurance.

5 Discussion

In this section, we reflect on the main findings from the survey. First, *b*roadly speaking, if respondents find a practice unfair, they also find the practice unacceptable. For easy of reading, we only speak about "fair" and "unfair" below, rather than about "fair and acceptable" and "unfair and unacceptable". Second, the survey respondents find almost all modern insurance practices that we described very unfair. Insurers who want to apply data-intensive underwriting or behaviour-based insurance should be aware that many people dislike those practices. behaviour-based car insurance seems the least unfair of the practices: respondents found behaviour-based car insurance slightly fairer than neutral. This gives some preliminary evidence for a standard theory in insurance fairness: some say that behaviour-based insurance is more "actuarially fair" than traditional insurance.[38] A more expansive survey or experiment could confirm whether the same results hold.

Third, respondents find *insurance* practices fairer if they feel that they can influence the premium. This view of respondents is understandable. We can imagine that it feels unfair to be punished with a higher premium for your

38 Chapter II, Section 5.3. See also Meyers & Van Hoyweghen, *Science as Culture* 2018/27.

car insurance, just because you live in an apartment with a letter in the house number. Meanwhile, insurers sometimes set premiums based on such data (see section 2). In the future, insurers may use more data on which a consumer has little influence on to set premiums. By underwriting more data-intensively, insures could find more correlations that help them to predict claim costs. After all, machine learning is good at finding correlations in large datasets. It may be bad news for insurers that many consumers dislike it if insurers calculate premiums based on characteristics that the consumer cannot influence.

We asked about one other type of behaviour-based insurance: life insurance with discounts based on a step counter. Respondents find such step counter insurance unfair. What could explain this difference between respondents' fairness perceptions about behaviour-based car insurance and behaviour-based insurance based on a step counter? We provide some preliminary thoughts. *First*, perhaps people feel that they cannot help it if they make few steps per day. Perhaps their legs hurt, *or* perhaps physical activity does not fit in their busy schedules. People might think that, by comparison, they have more influence on their driving style. *Second*, perhaps some people do physical activity (e.g. strength training) that a step counter does not register well. Such people might see a step counter as a bad way to register physical activity. *Third,* we asked about a discount if you register more than 10.000 steps. Perhaps people's reaction of respondents would have been different if we asked them about a different number of steps. More research on such questions would be welcome.

Fourth, respondents dislike it if insurers use seemingly illogical consumer characteristics to set premiums. A controversial characteristic that respondents found logical is Age. Perhaps in the scenario we presented, age was deemed fairer. Other surveys could ask about other situations. Respondents consider behaviour-based the most logical, but this does not seem true for the step counter. This could be because of the scenario we presented (10.000 steps). Respondents find it unfair if an insurer calculates the premium based on a consumer characteristic when the respondent does not see the logic, or relevance, of using that characteristic. The results raise questions about whether insurers can underwrite data-intensively in a way that consumers find fair. Insurers may find it difficult to convince consumers of the logic

or relevance of the data that they use for underwriting, even if in the eyes of the insurer, the data are relevant. Suppose that an insurer only wants to use consumer characteristics to set premiums if it can explain the logic or relevance of using that characteristic. Such a fairness policy might mean that the insurer cannot use certain consumer characteristics, even though an AI system has found a correlation between those characteristics and filing claims.

Fifth, respondents dislike target group insurance. Respondents find all types of insurance that are only available for certain groups, such as a car insurance specifically for family doctors, unfair. Meanwhile, some insurers do offer such target group insurance. Because in the Netherlands, target group insurance for doctors is long-established, it is remarkable that respondents seem to consider all target group insurances equally unfair. Perhaps people feel discriminated if they are not allowed to buy certain insurance products. Perhaps people's opinions change if an insurer gives arguments for a certain type of target group insurance. Sixth, respondents find it unfair if poor people pay higher insurance prices. Maybe respondents feel that such practices discriminate against poor people.

An insurer could try to make practices fairer by using a technical "fairness metric" that helps to prevent some unfair practices. A fairness metric is, roughly speaking, a score that expresses how fair a system is in terms of, for example, the number of false positives and false negatives).[39] However, for the examples we used in our survey, many fairness metrics do not seem suitable. Many metrics, such as Equality of Odds, focus on the outcome rather than whether a characteristic seems logical or is influenceable for the consumer.[40] Such metrics are useful, for example, if an insurer wishes to prevent the poor paying more. Perhaps metrics based on causality or control could provide some insight into the logic behind certain characteristics and whether consumers can influence the characteristic. But even then, it may be difficult for an insurer to foresee what characteristics and practices consumers see as logical.[41] Similar to recent fairness research, we think that

39 Barocas, Hardt & Narayanan 2023, p. 54.
40 Barocas, Hardt & Narayanan 2023.
41 Chapter II, Section 4.2.2. See also Fröhlich & Williamson, *FAccT* 2024, p. 411.

bringing down fairness to a single metric is difficult.[42] Explainability may help insurers understand the correlations they come across while applying machine learning to new datasets. But even if insurers can explain a decision to the consumer, that explanation does not necessarily make the decision itself fairer.[43] For example, even if an insurer explains why it finds a house number relevant for car insurance, the consumer may disagree with the explanation. Most consumers are not statisticians.[44]

It seems that the law in Europe can hardly protect people against the unfairness types highlighted in this paper. For instance, most insurance practices described in this paper do not violate non-discrimination law. Most non-discrimination statutes only protect people against discrimination based on certain protected characteristics, such as *age*, *disability*, *gender*, *religion*, ethnicity, or *sexual orientation*. Through the concept of indirect discrimination (disparate impact) some insurance practices might be illegal, if they disproportionally harm people with a certain ethnicity or other protected characteristic.[45] Most rules in the AI Act only apply to health and life insurance, not to other types of insurance. And even for health and life insurance, the AI Act does not ban or clearly regulate the practices described in this paper.[46] EU law specific for insurance does not focus on the unfairness types in this paper either.[47] The best source for some, albeit limited, legal protection is probably the General Data Protection Regulation.[48] The GDPR

42 See e.g. De Jonge & Hiemstra, *ACM Conference on Fairness, Accountability, and Transparency* 2023, p. 837.

43 For an overview of the similarities and differences between fairness and explainability, see Deck and others, *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 1579-1595.

44 See also Chapter II.

45 Chapter II, Section 4.1 & 5.1.

46 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

47 Directive (EU) 2016/97 of the European Parliament and of the Council of 20 January 2016 on insurance distribution (recast) (2016/97).

48 Regulation (EU) 2016/179 of the European Parliament and the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

may limit the amount of data that insurers can gather and analyse, and the GDPR's transparency requirements could help to make some insurance practices less of a black box.[49] The GDPR's rules on certain fully automated decisions with legal or similar effects may also offer some protection.[50] But the GDPR does not outright ban data-intensive underwriting or behaviour-based insurance. All in all, it appears that current law can offer, at most, some limited protection against the perceived unfairness discussed in this paper. An in-depth analysis of possible legal protection falls outside the scope of this paper.

We do not think that policymakers should ban every practice that people dislike or find unfair. To illustrate: many people dislike paying taxes. But the state needs tax income to run the country. Policymakers should consider many aspects of a problem. Similarly, we do not think that a practice is always morally fine, if a majority agrees with the practice. In our opinion, policymakers should give some attention to what people find fair and acceptable according to surveys, as long as policymakers scrutinize the survey in question, e.g. for data quality.[51]

Finally, this paper did not show that, according to the respondents, certain insurance practices must be prohibited by a lawmaker or withdrawn by an insurer. What people consider unacceptable may not necessarily overlap with their buying decisions - a "fairness paradox". People's stated preferences (what people say in surveys) may differ from their observed behaviour (how people act). For example, somebody might say that it is unacceptable if a life insurer offers discounts to people who show an active lifestyle with a step counter, but that same person might buy such an insurance product because the option is more affordable. Related discussions exist about a "privacy paradox": stated privacy preferences versus privacy behaviour.[52] Even if a fairness paradox exists, our paper can still inform the debate about the fairness and acceptance of data-intensive underwriting and behaviour-based insurance.

49 Article 12-14, General Data Protection Regulation.

50 Article 22, General Data Protection Regulation.

51 Shamon & Berning, *Survey Research Methods* 2020, p. 55-77 Pages. Clifford & Jerit, *Public Opinion Quarterly* 2016/80.

52 Kokolakis, *Computers & Security* 2017/64. Solove, *SSRN Electronic Journal* 2020. Dencik & Cable 2017.

6 Limitations and suggestions for further research

In this section we highlight limitations of this research and give suggestions for further research. This survey research was a first attempt to learn more about people's opinions about two modern trends in insurance. The sample of 999 respondents gives a reasonable cross-section of Dutch society, but the sample is not fully representative.

There are exciting opportunities for further research, in many disciplines. Social scientists could do more empirical research. Survey research in more countries could highlight the differences between countries. And surveys could contain more, and more in-depth, questions about issues such as behaviour-based insurance, health insurance, and insurance practices that reinforce financial inequality. Focus groups or interviews could provide more insight into why people find certain practices unfair. Further research could investigate whether there is a "fairness paradox" in relation to data-intensive underwriting and behaviour-based insurance, i.e. the relationship between people's stated fairness preferences and their buying decisions.

Computer scientists could analyse whether fairness metrics, as proposed by many papers of the ACM FAccT Conference for example, could help to protect people against the types of (perceived) unfairness as highlighted in this paper.[53]

There are many open normative and ethical questions. For instance, from a normative perspective, when is it acceptable for insurance practices to increase financial inequality? Is it fair for people to bear the cost of characteristics or behaviours that they have no or little control over or find illogical?[54] Could a comprehensive normative framework be created for insurance?[55] Is behaviour-based insurance fairer than traditional insurance or data-intensive underwriting? Should insurers communicate the shift to

53 One paper exploring such a topic is Fröhlich & Williamson, *FAccT* 2024.

54 While some papers discuss this question in general, in the past, insurers rejected 'the imposition of a requirement that so-called 'risk-factors' must have a demonstrable causal relationship to variations in the cost of claims [...] largely because it is based on data mining' Gandy 2016, p. 117.

55 Avraham made a first contribution toward such a framework Avraham, Logue & Schwarcz, *87 S. CAL. L. REV. 195* 2014.

more behaviour-based insurance with language implying solidarity or with language implying rationality?[56]

Many legal questions deserve more research too. Perhaps an in-depth analysis of current law could show that the GDPR or other law offer some protection against the perceived unfairness discussed in this paper. And maybe additional regulation should be considered. Should policymakers do more to mitigate the risk that insurance practices reinforce social inequality? There is some precedent: in many situations, the law aims to protect the weaker party, such as in the field of labour law and consumer protection law. Are sector-specific rules needed for the use of AI in the insurance sector? Should the law do more to ensure that consumers are not excluded from certain insurance products, or should contract freedom stay the same? Normative and ethical research could discuss whether, and if so when, it is appropriate for the law to intervene.

7 Concluding remarks

With a survey in the Netherlands, we explored people's opinions about two modern trends in insurance: (i) data-intensive underwriting and (ii) behaviour-based insurance. The main results include the following. First, if survey respondents find an insurance practice unfair, they also find the practice unacceptable. Second, respondents find almost all modern insurance practices that we described unfair. Third, respondents find practices fairer if they can influence the premium. For example, respondents find behaviour-based car insurance based on a car tracker relatively fair. Fourth, respondents find it unfair if an insurer calculates the premium based on a consumer characteristic, while the respondent does not see the logic of using that characteristic. Fifth, respondents find all types of insurance that are only available for certain groups, such as a car insurance specifically for family doctors, unfair. Sixth, respondents find it unfair if an insurance practice leads to higher prices for poorer people. Overall, respondents worry about these practices, finding many of them unfair. Many open questions remain,

56 D. Minty, 'Insurance and the Importance of Words and Language', https://www.ethicsandinsurance.info/insurance-and-the-importance-of-words-and-language/, 10 April 2025.

for many disciplines: we suggest further research for social scientists, legal scholars, ethicists, and computer scientists.

References of chapter V

Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB) 2019
Autoriteit Financiële markten (AFM) & De Nederlandse Bank (DNB), *Artificial intelligence in the insurance sector. An exploratory study*, AFM & DNB 2019, https://www.afm.nl/~/profmedia/files/rapporten/2019/afm-dnb-verzekeringssector-ai-eng.pdf.

Avraham, Logue & Schwarcz, *87 S. CAL. L. REV. 195* 2014
R. Avraham, K.D. Logue & D. Schwarcz, 'Understanding Insurance Antidiscrimination Laws', *87 S. CAL. L. REV. 195* 2014, https://scholarship.law.umn.edu/faculty_articles/576.

Barocas, Hardt & Narayanan 2023
S. Barocas, M. Hardt & A. Narayanan, *Fairness and Machine Learning*, MIT Press 2023, https://fairmlbook.org/.

Barry & Charpentier, *Big Data & Society* 2020/7
L. Barry & A. Charpentier, 'Personalization as a promise: Can Big Data change the practice of insurance?', *Big Data & Society* 2020/7, issue 1, p. 205395172093514, http://journals.sagepub.com/doi/10.1177/2053951720935143, doi:10.1177/2053951720935143.

Barry & Charpentier, *Ethics and Information Technology* 2023/25
L. Barry & A. Charpentier, 'Melting contestation: insurance fairness and machine learning', *Ethics and Information Technology* 2023/25, issue 4, p. 49, https://link.springer.com/10.1007/s10676-023-09720-y, doi:10.1007/s10676-023-09720-y.

Binns and others 2018
R. Binns and others, ''It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions', 2018.

Charpentier & Vamparys, *Big Data & Society* 2025/12
A. Charpentier & X. Vamparys, 'Artificial intelligence and personalization of insurance: Failure or delayed ignition?', *Big Data & Society* 2025/12, issue 1, p. 20539517241291817, https://journals.sagepub.com/doi/10.1177/20539517241291817, doi:10.1177/20539517241291817.

Clifford & Jerit, *Public Opinion Quarterly* 2016/80
S. Clifford & J. Jerit, 'Cheating on Political Knowledge Questions in Online Surveys: An Assessment of the Problem and Solutions', *Public Opinion Quarterly* 2016/80, issue 4, p. 858-887, https://academic.oup.com/poq/article-lookup/doi/10.1093/poq/nfw030, doi:10.1093/poq/nfw030.

College voor de Rechten van de Mens 2014
College voor de Rechten van de Mens, *Advies aan Dazure B.V. over premiedifferentiatie op basis van postcode bij de Finvita overlijdensrisicoverzekering*, 2014, https://publicaties.mensenrechten.nl/publicatie/29c613ac-efab-4d2a-a26c-fe1cb2556407.

De Jonge & Hiemstra, *ACM Conference on Fairness, Accountability, and Transparency* 2023
T. De Jonge & D. Hiemstra, 'UNFair: Search Engine Manipulation, Undetectable by Amortized Inequity', *ACM Conference on Fairness, Accountability, and Transparency* 2023, https://dl.acm.org/doi/10.1145/3593013.3594046, doi:10.1145/3593013.3594046.

Deck and others, *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 1579-1595
L. Deck and others, 'A Critical Survey on Fairness Benefits of Explainable AI', *ACM Conference on Fairness, Accountability, and Transparency* 2024, p. 1579-1595, https://dl.acm.org/doi/10.1145/3630106.3658990, doi:10.1145/3630106.3658990.

Dencik & Cable 2017
L. Dencik & J. Cable, 'The Advent of Surveillance Realism: Public Opinion and Activist Responses to the Snowden Leaks', 2017.

Dutch Financial Authority (AFM) 2021
Dutch Financial Authority (AFM), *The personalisation of prices and conditions in the Insurance Sector. An exploratory study*, The Netherlands: AFM 2021, https://www.afm.nl/en/sector/actueel/2021/juni/aandachtspunten-gepersonaliseerde-beprijzing.

European Commission. Directorate General for Justice and Consumers. and others 2022

European Commission. Directorate General for Justice and Consumers. and others, *Indirect discrimination under Directives 2000/43 and 2000/78.*, LU: Publications Office 2022, https://data.europa.eu/doi/10.2838/93469.

Fröhlich & Williamson, *FAccT* 2024
C. Fröhlich & R.C. Williamson, 'Insights From Insurance for Fair Machine Learning', *FAccT* 2024, https://dl.acm.org/doi/10.1145/3630106.3658914, doi:10.1145/3630106.3658914.

Gandy 2016
O.H. Gandy, *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*, Routledge 2016.

H Rubin, Aas & Williams, *Journal of Information Technology Teaching Cases* 2023/13
T. H Rubin, T.H. Aas & J. Williams, 'Big data and data ownership rights: The case of car insurance', *Journal of Information Technology Teaching Cases* 2023/13, issue 1, p. 82-87, https://doi.org/10.1177/20438869221096859, doi:10.1177/20438869221096859.

Henckaerts & Antonio, *Insurance: Mathematics and Economics* 2022/105
R. Henckaerts & K. Antonio, 'The added value of dynamically updating motor insurance prices with telematics collected driving behavior data', *Insurance: Mathematics and Economics* 2022/105, p. 79-95, https://linkinghub.elsevier.com/retrieve/pii/S0167668722000385, doi:10.1016/j.insmatheco.2022.03.011.

Kiviat, *Sociological Science* 2021/8
B. Kiviat, 'Which Data Fairly Differentiate? American Views on the Use of Personal Data in Two Market Settings', *Sociological Science* 2021/8, p. 26-47, https://sociologicalscience.com/articles-v8-2-26/, doi:10.15195/v8.a2.

Kokolakis, *Computers & Security* 2017/64
S. Kokolakis, 'Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon', *Computers & Security* 2017/64, p. 122-134, https://linkinghub.elsevier.com/retrieve/pii/S0167404815001017, doi:10.1016/j.cose.2015.07.002.

Krippner & Hirschman, *Theory and Society* 2022/51
G.R. Krippner & D. Hirschman, 'The person of the category: the pricing

of risk and the politics of classification in insurance and credit', *Theory and Society* 2022/51, issue 5, p. 685-727, https://doi.org/10.1007/s11186-022-09500-5, doi:10.1007/s11186-022-09500-5.

McLoughney and others, *IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* 2023
A.J. McLoughney and others, '"Emerging proxies" in information-rich machine learning: a threat to fairness?', *IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)* 2023, https://ieeexplore.ieee.org/document/10155045/, doi:10.1109/ETHICS57328.2023.10155045.

Meyers 2018
G. Meyers, *Behaviour-based Personalisation in Health Insurance: a Sociology of a not-yet Market*, KU Leuven 2018, https://lirias.kuleuven.be/2087689&lang=en.

Meyers & Hoyweghen, *Big Data & Society* 2020/7
G. Meyers & I.V. Hoyweghen, '"Happy failures": Experimentation with behaviour-based personalisation in car insurance', *Big Data & Society* 2020/7, issue 1, p. 205395172091465, http://journals.sagepub.com/doi/10.1177/2053951720914650, doi:10.1177/2053951720914650.

Meyers & Van Hoyweghen, *Science as Culture* 2018/27
G. Meyers & I. Van Hoyweghen, 'Enacting Actuarial Fairness in Insurance: From Fair Discrimination to Behaviour-based Fairness', *Science as Culture* 2018/27, issue 4, p. 413-438, https://www.tandfonline.com/doi/full/10.1080/09505431.2017.1398223, doi:10.1080/09505431.2017.1398223.

Moor & Lury, *Journal of Cultural Economy* 2018/11
L. Moor & C. Lury, 'Price and the person: markets, discrimination, and personhood', *Journal of Cultural Economy* 2018/11, issue 6, p. 501-513, https://doi.org/10.1080/17530350.2018.1481878, doi:10.1080/17530350.2018.1481878.

Moosavi & Ramnath, *Accident Analysis & Prevention* 2023/192
S. Moosavi & R. Ramnath, 'Context-aware driver risk prediction with telematics data', *Accident Analysis & Prevention* 2023/192, p. 107269, https://linkinghub.elsevier.com/retrieve/pii/S0001457523003160, doi:10.1016/j.aap.2023.107269.

Muszyński, *Ask: Research and Methods* 2023/32
M. Muszyński, 'Attention checks and how to use them: Review and practical recommendations', *Ask: Research and Methods* 2023/32, issue 1, p. 3-38, https://kb.osu.edu/handle/1811/103631, doi:10.18061/ask.v32i1.0001.

O'Neil 2016
C. O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, New York: Penguin Books 2016.

OSTP 2022
OSTP, *Blueprint for an AI Bill of Rights. Making automated systems work for the American people*, The White house | OSTP 2022, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

Shamon & Berning, *Survey Research Methods* 2020, p. 55-77 Pages
H. Shamon & C.C. Berning, 'Attention Check Items and Instructions in Online Surveys: Boon or Bane for Data Quality?', *Survey Research Methods* 2020, p. 55-77 Pages, https://ojs.ub.uni-konstanz.de/srm/article/view/7374, doi:10.18148/SRM/2020.V14I1.7374.

Solove, *SSRN Electronic Journal* 2020
D.J. Solove, 'The Myth of the Privacy Paradox', *SSRN Electronic Journal* 2020, https://www.ssrn.com/abstract=3536265, doi:10.2139/ssrn.3536265.

Starke and others, *Big Data & Society* 2022/9
C. Starke and others, 'Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature', *Big Data & Society* 2022/9, issue 2, p. 205395172211151, http://journals.sagepub.com/doi/10.1177/20539517221115189, doi:10.1177/20539517221115189.

Tanninen, *Big Data & Society* 2020/7
M. Tanninen, 'Contested technology: Social scientific perspectives of behaviour-based insurance', *Big Data & Society* 2020/7, issue 2, p. 205395172094253, http://journals.sagepub.com/doi/10.1177/2053951720942536, doi:10.1177/2053951720942536.

The Geneva Association (Noordhoek) 2023
The Geneva Association (Noordhoek), *Regulation of Artificial Intelligence in Insurance: Balancing consumer protection and innovation*, The Geneva Assoca-

tion 2023, https://www.genevaassociation.org/publication/public-policy-and-regulation/regulation-artificial-intelligence-insurance-balancing.

Tselentis, Yannis & Vlahogianni, *Accident Analysis & Prevention* 2017/98
D.I. Tselentis, G. Yannis & E.I. Vlahogianni, 'Innovative motor insurance schemes: A review of current practices and emerging challenges', *Accident Analysis & Prevention* 2017/98, p. 139-148, https://linkinghub.elsevier.com/retrieve/pii/S0001457516303670, doi:10.1016/j.aap.2016.10.006.

Appendix A

Table A1: Fairness by data type.

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Postal code	−25.042	998	<.001	<.001	−0.862	−0.93	−0.79
Uneven	−59.407	998	<.001	<.001	−1.551	−1.60	−1.50
Letter	−51.713	998	<.001	<.001	−1.453	−1.51	−1.40
Price	4.747	998	<.001	<.001	0.173	0.10	0.24
Speed	−19.872	998	<.001	<.001	−0.703	−0.77	−0.63
Colour	−47.656	998	<.001	<.001	−1.381	−1.44	−1.32
Location	−25.630	998	<.001	<.001	−0.851	−0.92	−0.79
Gender	−42.661	998	<.001	<.001	−1.269	−1.33	−1.21
Age	−6.999	998	<.001	<.001	−0.238	−0.31	−0.17
Tracker	8.408	998	<.001	<.001	0.302	0.23	0.37
Steps	−27.432	998	<.001	<.001	−0.928	−0.99	−0.86
Neighbourhood	−41.880	998	<.001	<.001	−1.219	−1.28	−1.16
Migrant	−29.187	998	<.001	<.001	−1.010	−1.08	−0.94
Poor	−23.398	998	<.001	<.001	−0.811	−0.88	−0.74

Note. Negative t-values represent a negative deviation from the scale midpoint, whereas positive t-values represent positive deviations from the scale midpoint. Any mean differences are significant at the p<.001 level.

Table A2: Acceptance by data-type.

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Postal code	−22.345	998	<.001	<.001	−0.797	−0.87	−0.73
Uneven	−51.931	998	<.001	<.001	−1.494	−1.55	−1.44
Letter	−48.404	998	<.001	<.001	−1.417	−1.47	−1.36
Price	5.493	998	<.001	<.001	0.202	0.13	0.27
Speed	−18.503	998	<.001	<.001	−0.674	−0.75	−0.60
Colour	−43.070	998	<.001	<.001	−1.328	−1.39	−1.27
Location	−21.368	998	<.001	<.001	−0.750	−0.82	−0.68
Gender	−35.978	998	<.001	<.001	−1.189	−1.25	−1.12
Age	−4.639	998	<.001	<.001	−0.158	−0.23	−0.09
Tracker	6.141	998	<.001	<.001	0.223	0.15	0.29
Steps	−25.161	998	<.001	<.001	−0.875	−0.94	−0.81
Neighbourhooc	−38.270	998	<.001	<.001	−1.171	−1.23	−1.11
Migrant	−27.784	998	<.001	<.001	−0.970	−1.04	−0.90
Poor	−21.787	998	<.001	<.001	−0.768	−0.84	−0.70

Note. Negative t-values represent a negative deviation from the scale midpoint, whereas positive t-values represent positive deviations from the scale midpoint. Any mean differences are significant at the p<.001 level.

Appendix B

Table B1: Influence by data-type.

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Postal code	−23.985	998	<.001	<.001	−0.840	−0.91	−0.77
House number: uneven	−41.272	998	<.001	<.001	−1.339	−1.40	−1.28
House number: letter	−37.216	998	<.001	<.001	−1.244	−1.31	−1.18
Price car	8.320	998	<.001	<.001	0.293	0.22	0.36
Car maximum speed	−9.930	998	<.001	<.001	−0.394	−0.47	−0.32
Car colour	−18.669	998	<.001	<.001	−0.782	−0.86	−0.70
Town size	−22.520	998	<.001	<.001	−0.785	−0.85	−0.72
Gender	−35.177	998	<.001	<.001	−1.204	−1.27	−1.14
Age	−17.354	998	<.001	<.001	−0.647	−0.72	−0.57
Car tracker (behavioural)	19.994	998	<.001	<.001	0.693	0.62	0.76
Steps (behavioural)	−4.478	998	<.001	<.001	−0.189	−0.27	−0.11
Poor area	−24.879	998	<.001	<.001	−0.861	−0.93	−0.79

Note. Negative t-values represent a negative deviation from the scale midpoint, whereas positive t-values represent positive deviations from the scale midpoint. Any mean differences are significant at the p<.001 level.

Table B2: Perceived logic by insurance practice.

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Postal code	−20.011	998	<.001	<.001	−0.739	−0.81	−0.67
House number: even	−69.493	998	<.001	<.001	−1.635	−1.6	−1.59
House number: letter	−57.445	998	<.001	<.001	−1.507	−1.56	−1.46
Price car	7.240	998	<.001	<.001	0.272	0.20	0.35
Car maximum speed	−19.239	998	<.001	<.001	−0.714	−0.79	−0.64
Car colour	−54.640	998	<.001	<.001	−1.473	−1.53	−1.42
Location	−20.424	998	<.001	<.001	−0.725	−0.79	−0.66
Gender	−38.598	998	<.001	<.001	−1.232	−1.29	−1.17
Age	−0.544	998	.293	.587	−0.019	−0.09	0.05
Car tracker (behavioural)	6.056	998	<.001	<.001	0.222	0.15	0.29
Steps (behavioural)	−27.671	998	<.001	<.001	−0.940	−1.01	−0.87
Neigbourhood	−28.901	998	<.001	<.001	−0.993	−1.06	−0.93
Migrant	−22.874	998	<.001	<.001	−0.837	−0.91	−0.77
Poor area	−14.108	998	<.001	<.001	−0.544	−0.62	−0.47

Note. Negative t-values represent a negative deviation from the scale midpoint, whereas positive t-values represent positive deviations from the scale midpoint. Any mean differences are significant at the p<.001 level.

Appendix C

Table C1: Fairness by target group.

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Family doctor	−35.159	998	<.001	<.001	−1.147	−1.21	−1.08
Lawyer	−37.321	998	<.001	<.001	−1.199	−1.26	−1.14
Scientist	−38.081	998	<.001	<.001	−1.209	−1.27	−1.15
Teacher	−36.478	998	<.001	<.001	−1.187	−1.25	−1.12
Realtor	−38.571	998	<.001	<.001	−1.216	−1.28	−1.15
Electrician	−38.210	998	<.001	<.001	−1.206	−1.27	−1.14
Caregiver	−34.704	998	<.001	<.001	−1.147	−1.21	−1.08
Salesperson	−36.369	998	<.001	<.001	−1.178	−1.24	−1.11
Hairdresser	−37.366	998	<.001	<.001	−1.194	−1.26	−1.13
Unemployed	−33.479	998	<.001	<.001	−1.118	−1.18	−1.05
Higher education	−33.849	998	<.001	<.001	−1.133	−1.20	−1.07

Note. Negative t-values represent a negative deviation from the scale midpoint, whereas positive t-values represent positive deviations from the scale midpoint. Any mean differences are significant at the p<.001 level.

Table C2: Acceptance by target group.

	t	df	Significance		Mean Difference	95% Confidence Interval of the Difference	
			One-Sided p	Two-Sided p		Lower	Upper
Family doctor	−33.177	998	<.001	<.001	−1.124	−1.19	−1.06
Lawyer	−36.683	998	<.001	<.001	−1.192	−1.26	−1.13
Scientist	−36.179	998	<.001	<.001	−1.184	−1.25	−1.12
Teacher	−35.322	998	<.001	<.001	−1.174	−1.24	−1.11
Realtor	−35.971	998	<.001	<.001	−1.187	−1.25	−1.12
Electrician	−35.563	998	<.001	<.001	−1.178	−1.24	−1.11
Caregiver	−33.135	998	<.001	<.001	−1.131	−1.20	−1.06
Salesperson	−35.425	998	<.001	<.001	−1.171	−1.24	−1.11
Hairdresser	−34.936	998	<.001	<.001	−1.172	−1.24	−1.11
Unemployed	−32.270	998	<.001	<.001	−1.101	−1.17	−1.03
Higher education	−33.090	998	<.001	<.001	−1.124	−1.19	−1.06

Note. Negative t-values represent a negative deviation from the scale midpoint, whereas positive t-values represent positive deviations from the scale midpoint. Any mean differences are significant at the p<.001 level.

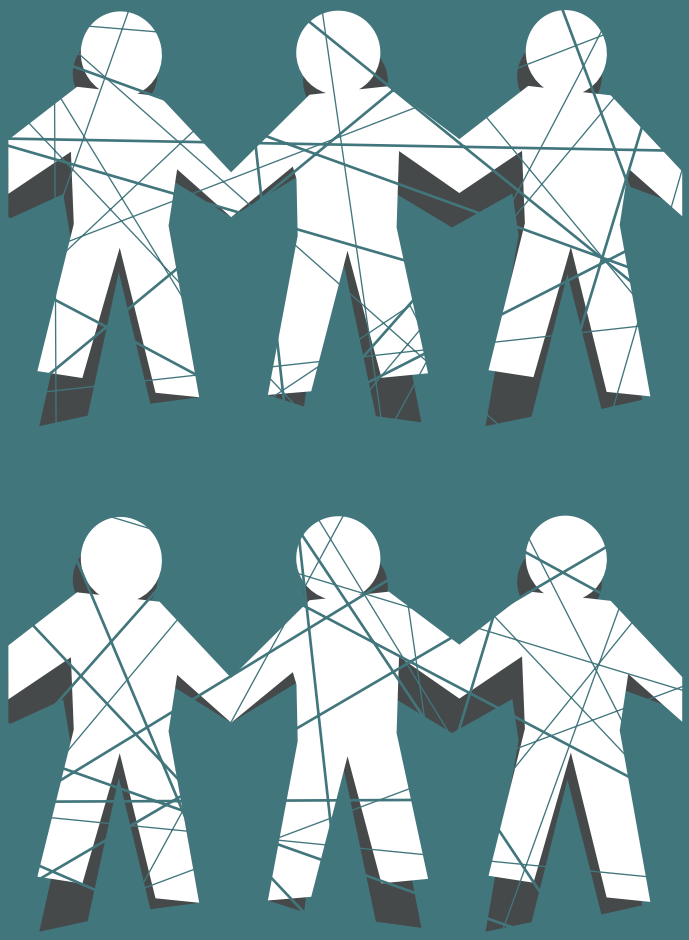
Appendix D

Survey questions on life and burglary insurance

"With life insurance, you pay a higher premium if the insurer expects you to die earlier. Because if you die earlier, you will pay a premium for a shorter period. The insurer wants to charge higher premiums to people who are likely to die earlier. In the Netherlands, poorer people live on average seven years shorter than richer people. The insurer therefore charges higher premiums to poorer people. People with a certain income often live in certain neighbourhoods.

The insurer charges higher prices in poorer neighbourhoods."

"With burglary insurance, you pay a higher premium if the risk of burglary in your home is higher. Suppose poorer neighbourhoods have more burglaries. An insurer charges higher premiums in neighbourhoods with more burglaries. This leads to poorer people paying higher premiums on average."



VI.
Conclusion

In this PhD thesis, I investigated the central research question: *to what extent do insurers introduce discrimination-related risks by using AI systems for underwriting*? I investigated four gaps in the literature. First, I discussed how insurers currently use AI systems for underwriting and identify possible discrimination-related threats. Second, I investigated whether the ban in Art. 9 GDPR could hinder insurers from testing whether their AI systems pose discrimination-related threats. Next, I investigated the exception to the ban in the GDPR enshrined in Article 10(5) AI Act. Finally, I investigated the consumer's perspective about the discrimination-related threats I identified.

1 Answers by Chapters II-V

The individual chapters answer the research question as follows. In Section 2, I give a general answer.

Chapter II highlights two AI-related trends in modern insurance: data-intensive underwriting and behaviour-based insurance. Under the two trends, insurers apply data differently. For data-intensive underwriting, the insurer applies more characteristics to calculate the insurance premium than previously. In calculating the insurance price, the insurer compares the claims history across groups, which happens at the collective level. For example, in assessing the price a senior consumer will pay, the insurer looks at the claims history of seniors. For behaviour-based insurance, the insurer processes the data of consumers (often) in real-time, at the level of the individual.[1]

1 Some work presents another view about the individualisation of behaviour-based insurance but is (in my opinion) having a distinctly different discussion. Barocas, Hardt & Narayanan 2023, p. 36; Moor & Lury, Journal of Cultural Economy 2018/11; McFall

Without thorough testing and monitoring for possible risks by the insurer, data-intensive underwriting and behaviour-based insurance make several threats more likely. With data-intensive underwriting, insurers decide based on a proxy attribute what premium a consumer pays, and insurers may find many such proxy attribute attributes.

As a result, especially *indirect indiscrimination* could happen more often. Insurers may end up using relationships for their underwriting that disadvantage certain ethnicities. Insurers could also use characteristics that are *seemingly irrelevant*. Insurers are likely to use characteristics that are at least statistically relevant. Nevertheless, insurers may use a characteristic that is statistically relevant, but seems irrelevant to the consumer: consumers are not statisticians. The house number-example at the beginning of the thesis illustrates such a seemingly irrelevant characteristic. Insurers may also use characteristics for underwriting that consumers have *no influence* on. Insurers could – possibly unintentionally – discriminate against the poor (*the poor pay more*). Finally, insurers may – possibly unintentionally – *exclude certain groups*.[2] Overall, insurers introduce new threats by using AI systems for underwriting risks.

Chapter III explored whether insurers are allowed to collect the sensitive data necessary for testing whether an AI system discriminates against, for example, certain ethnicities, and whether the GDPR should have an exception for that purpose. In principle, the GDPR bans the use of certain 'special categories of data' (sometimes called 'sensitive data'), which include data on ethnicity, religion, and sexual preference. The ban does not include information on gender and age. EU non-discrimination law contains two protected characteristics for insurance, which have different statuses under the GDPR: ethnicity, which is sensitive data under the GDPR, and gender, which is not sensitive data.[3]

Whether the GDPR should allow insurers to collect sensitive data for preventing discrimination is not clear. I presented arguments for and against. The first argument for an exception is that insurers could collect sensitive

& Moor, Distinktion: Journal of Social Theory 2018/19. I discuss this further in Section 4 (Research agenda).

2 For a summary, see the conclusion of Chapter II, Table 2.

3 Compare Chapter II, section 3 with Chapter III, Figure 1.

data to prevent AI discrimination. Second, AI discrimination testing could increase the trust consumers have in an AI system.[4] There are also arguments against allowing insurers to collect sensitive data: First, the mere collection of sensitive data is already a breach of the consumer's fundamental right to data protection. Second, the data could be misused: the exception creates a relatively large risk of some data leaks. Finally, the field of *fairness of AI* is still relatively immature.[5] Overall, I cannot decide whether an exception is justifiable. On the one hand, the threats I identified in chapter II of the thesis seem serious. On the other hand, scholars should not dictate policy, which should be democratically decided.[6]

Chapter IV analyses the exception in the AI Act allowing the collection of sensitive data. The exception enables insurers to evaluate the datasets they use to develop an AI system, for possible discriminatory effects. The exception does not seem promising for insurers: the new exception only applies to 'high-risk' systems, which only includes life and health insurance. Insurers renting an AI system, for example *AI as a service*, cannot make use of the exception: the exception only applies for the developers of an AI system.[7] This affects up to a third of the European insurers.[8]

The AI Act's exception sets a high bar of proportionality, meaning that life and health insurers would have to be very careful with the data they

4 While I did not talk about trust in Chapter II, especially in the context of insurance, consumer trust seems important: insurers themselves seem skeptical of machine learning, and consumers seem to find insurance in general unfair (see Chapter V). In the Netherlands, following some scandals (as part of the 'Woekerpolis' affair), insurers seem cautious of losing the trust of their consumers. For a discussion of the *Woekerpolis*-case, see (in Dutch) Van Meerten & Schmidt, *Pensioen magazine* 2015. The Woekerpolis-affair still follows Dutch insurers, in the form of collective action by consumers. See (in Dutch): Hoge Raad 11 February 2022, ECLI:NL:HR:2022:166 (*Vereniging Woekerpolis.nl*); Gerechtshof Den Haag 26 September 2023, ECLI:NL:GHDHA:2023:1852 (*Vereniging Woekerpolis.nl*).

5 Many insurers seem to have much experience with statistics: the fields of AI and insurance can learn from each other about fairness. See e.g. Fröhlich & Williamson, *FAccT* 2024. See also the research agenda at the end of this Chapter.

6 See also the discussion in Section 3 of this chapter.

7 The exception in Article 10(5) AI Act only applies to providers, not to deployers. See Chapter IV.

8 European insurers outsource about 34% of their AI from third parties. Some of these were developed together with third parties. See EIOPA 2024, p. 26.

can collect. The AI Act leaves open which biases insurers must (and can) correct with the collected sensitive data. I expect that there will be much uncertainty about which biases insurers must correct under the AI Act.[9] Finally, the obligation in Article 10 AI Act targets biases that may cause *legal* discrimination: under Art. 10 AI Act, insurers are not obligated to fight unfair differentiation.

In Chapter V, I presented a survey that gives a first glance at the opinion of the Dutch public about many controversial examples. Chapter V concluded: (i) if survey respondents find an insurance practice unfair, they also find the practice unacceptable. (ii) Respondents find almost all modern insurance practices unfair. (iii) Respondents find practices fairer if they can influence the premium. For example, respondents find behaviour-based car insurance based on a car tracker relatively fair. (iv) Respondents find it unfair if an insurer calculates the premium based on a consumer characteristic, while the respondent does not see the logic of using that characteristic. (v) Respondents find all types of insurance that are only available for certain groups, such as a car insurance specifically for family doctors, unfair. (vi) Respondents find it unfair if an insurance practice leads to higher prices for poorer people. Overall, respondents seem to find the examples of unfair differentiation in the survey *more* unfair than existing insurance practices.

2 To what extent do insurers introduce discrimination-related risks by using AI systems for underwriting?

My answer to the central research question is: insurers are likely to introduce new threats to fairness and non-discrimination by introducing AI systems into underwriting. I focused on machine learning models, not on generative AI models.[10] Perhaps insurers could prevent the identified threats if they could properly test whether an AI system discriminates against certain groups.

Because of legal hurdles, insurers cannot always properly test the characteristics they use to *mitigate* the possible discrimination-related risks.

9 As I noted in Chapter IV, the AI Act's standardization bodies and regulatory bodies could yet clarify which biases must be corrected under Article 10(5) AI Act.

10 See Introduction. Chapter I, Section 2 of the thesis.

My research shows that it is not trivial for insurers to give guarantees to consumers: current laws do not allow insurers to test for discrimination based on ethnicity, and insurers would need to hold surveys to know whether consumers can 'influence' certain characteristics or find them 'irrelevant'. Such surveys are a costly, time-consuming process. On a more positive note, once insurers decide on a specific characteristic (or a limited number of characteristics) that they want to use, an independent survey seems a worthwhile method to involve society.[11]

3 Discussion

The answer to the research question leaves some questions that are worth discussing further. In the following subsections, I highlight to what extent current laws protect the consumer against discrimination (3.1), whether similar discriminatory effects can also happen outside of the insurance sector (3.2), and finally whether (and possibly, how) the lawmaker should intervene (3.3).

3.1 Discrimination and unfair differentiation: gaps in current laws

The survey in Chapter V showed that people care not only about the output of an AI system or whether the system excludes certain groups, but also about biases in the *data* insurers use to develop AI systems. For example, respondents care whether an insurer bases decisions on data with seemingly irrelevant characteristics.

The results of the survey in Chapter V raise the question to what extent current laws protect consumers against discrimination-related threats. Non-discrimination law applies as far as insurers might discriminate illegally: insurers have an obligation to not discriminate against people based on certain protected characteristics.

Insurers can justify indirect discrimination. If an insurer can objectively justify a certain differentiation, based on all the circumstances of the case, then the differentiation does not qualify as indirect discrimination. At face

11 As I argued in the Introduction of the thesis, fairness requires participation of many stakeholders in society.

value, the test for justifying indirect discrimination is in the advantage of the insurer: it justifies differentiation. I argue that the test can also be a *burden* for the insurer that protects the consumer. Apart from a legitimate aim, an objective justification requires a proportionality test: the differentiation must be 'appropriate and necessary'.[12]

In general, for differentiation to be 'appropriate and necessary', the insurer must choose the least intrusive form of differentiation: insurers must weigh different alternative proxy attributes against each other.[13] Take, for example, an insurer who wonders whether they can choose a proxy attribute that might discriminate indirectly against a certain ethnicity. Let us assume the insurer has a legitimate aim (courts are generally lenient in recognizing an aim as legitimate). Insurers must choose the proxy attribute that leads to the 'least intrusive' differentiation towards that ethnicity, which means insurers may (in principle) not choose a proxy attribute that discriminates more than the proxy attributes the insurer currently uses, and the insurer must ensure that the differentiation does not cause disproportionate disadvantages for protected groups such as the ethnic group in question.

The new AI Act seems to be in tune with non-discrimination law: Article 10 AI Act targets biases that affect the 'health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law'.[14] In other words, under the AI Act, if insurers encounter biases that could lead to illegal discrimination in the datasets they use for AI training, then they must mitigate the risk.[15]

For unfair differentiation, the research results are less positive. In Chapter II, I identified four types of differentiation that fall *outside* of current non-discrimination law. For example, insurers do not illegally discriminate if they

12 For justifying indirect discrimination, the differentiation must be 'appropriate and necessary'. See Chapter II, Section 4.1.

13 See also Hildebrandt, *European Data Protection Law Review* 2021/7, p. 365. Deciding which proxy is the least intrusive can be difficult, requiring a complex weighing of interests. A case where insurers had to weigh many different types of attributes (e.g. income, postal code, 'other health factors') is the *Finvita* case before the Dutch human rights authority. See College voor de Rechten van de Mens 2014, p. 20-22.

14 See Chapter IV.

15 This is assuming the caveats with the exception I mentioned in the discussion of Chapter IV do not hinder insurers too much.

use a characteristic that seems irrelevant to the consumer, but that does not differentiate 'based on' a protected characteristic, such as a person's house number. Neither non-discrimination law, nor Article 10 AI Act, cover such cases.[16] It is possible that other laws, not covered by the thesis, could help somewhat.[17]

On the positive side: the ban in Article 9 GDPR does not hinder insurers in collecting the necessary data to test whether a certain characteristic *unfairly differentiates*. In sum, while current laws apply to possible direct and indirect discrimination, I find that they will not explicitly help the consumer against cases of unfair differentiation.

3.2 Similarities with other sectors, outside of insurance

Arguably, insurance can 'act as an analogon for the social situatedness of machine learning systems, hence allowing machine learning scholars to take insights from the rich and interdisciplinary insurance literature'.[18] Can other sectors also learn from the insights in this thesis?

The four factors for unfair differentiation from Chapter II could also apply to other sectors, but the normativity is different.[19] For example, energy companies could target consumers based on their postal code, a characteristic that consumers cannot easily influence. However, in the energy sector, such differentiation may be easier to justify than in insurance. For example, the costs of bringing energy to a consumer are not the same across a country, which could require an energy company to differentiate by region: there is a causal reason for the differentiation. In other words, in the energy sector, a postal code may generally seem a relevant characteristic for the consumer. By contrast, as demonstrated by Chapter V, a postal code appears irrelevant for car insurance prices in the eyes of the consumer.

In sectors other than insurance, differentiation may also result in the poor paying more. For example, in The Netherlands, the government used the system *Systeem Risico Indicatie* to combat fraud, which resulted in poor people

16 Much is still unknown about the AI Act. It is possible that the AI Act deals with some of the factors in other ways, such as the ban in social scoring. See Section 4.

17 See the limitations and research agenda in Section 4.

18 Fröhlich & Williamson, *FAccT* 2024, p. 407.

19 For an overview of the factors, see Chapter II, Table 2.

being targeted more often.[20] Where an organisation differentiates between people – for example, by deciding which neighbourhood to check for possible fraud – there comes a possibility that the differentiation disproportionately affects the poor. In other sectors, money is perhaps not the object at stake: for example, in the SyRI case, people felt constantly under suspicion of fraud. Perhaps in many sectors outside insurance, the poor pay more *in kind*.

Finally, in sectors outside insurance, organisations also risk unintentionally excluding certain groups: for example, in the labour market.[21] In sum, the four factors from chapter II can also apply to other sectors than insurance, but the normativity is still sector-specific: underwriting is a very specific context. While there is much other sectors could learn from insurance, organisations from other sectors should be careful in following context-specific examples of fairness from insurance.

3.3 Should the legislator intervene?

Overall, the thesis gave some hints about why a legislator should intervene in the insurance sector, but also arguments about why such an intervention would be hasty. Two arguments for an intervention are as follows. First, the discrimination-related threats I identified in the thesis are impactful: they may cause discrepancies between people or groups of people. While the odds of an insurer suddenly switching to machine learning completely seem low at the moment,[22] the impact would be great. Insurance is a market with a great deal of competition. A fintech company or a large company like Microsoft could enter the insurance sector with a war chest, which could force insurers to innovate faster.[23]

20 For a paper about the SyRI case, see Van Bekkum & Zuiderveen Borgesius, *European Journal of Social Security* 2021/23;Appelman, Fathaigh & Van Hoboken, *JIPITEC* 2021/12.

21 For a book that warns for exclusion across many sectors, see O'Neil 2016. O'Neil calls such AI systems 'weapons of math destruction'.

22 Insurers currently experience hurdles in introducing AI systems in their processes, slowing down the two trends. See Charpentier & Vamparys, *Big Data & Society* 2025/12.

23 While not an example of a big tech invasion, some big tech companies already offer Large Language Models specifically to insurers. Quess GTS, 'Microsoft 365 Copilot to Revolutionize Insurance Underwriting', https://www.quessgts.com/empowering-insurance-microsoft-365-copilots-journey-in-revolutionizing-underwriting/, https://quessgts.com/, 3 April 2024.

A second argument is that certain norms create legal uncertainty. For example, the AI Act has created some confusion among Dutch insurers: while the insurers successfully lobbied to limit the 'high-risk AI' to life and health insurance, it seems that the AI Act still introduced many norms that might be a burden on the entire sector.[24] The result is that insurers do not know the full extent of the new obligations the AI Act brings compared to their existing (self-regulating) framework. By intervening, legislators and regulators could help prevent legal uncertainty.

There are also arguments against (further) regulation. First, one could say that the problem of unfair differentiation seems like a slippery slope. As the survey in Chapter V highlighted, people find *all* cases of unfair differentiation unfair, so it seems that insurers are always on the defence: they must choose characteristics that people find *the least* unfair. When treading on such thin ice, scholars should not recommend policy too hastily.

Another argument against overregulation is that insurers, in principle, have contractual freedom: fairness is not the main foundation of insurance. Regulators can intervene for severe cases, but insurers traditionally have the freedom to set prices that consumers find unfair, to some extent. More strongly expressed, the survey in Chapter V is not binding for insurers. The chapter is not a referendum for what should or should not be regulated.

Finally, data-intensive underwriting and behaviour-based insurance are gradual trends: they have not changed the entire sector completely so far, therefore regulators still have time to intervene against these possible threats later. This is supported by recent work arguing that insurers experience many hurdles in introducing AI in insurance, some of which are not due to the law.[25]

If legislators intervene, how should they? From the point of view of a regulator, I see common ground in the *sensitive data* problem that follows from Chapter III of the thesis and *discrimination and unfair differentiation*: it should be clear for insurers which biases are so harmful for society, that

24 See, for example, Dutch Association Of Insurers, 'Comments on draft text of AI Act [Bedenkingen bij concepttekst AI Act]', https://www.verzekeraars.nl/publicaties/actueel/bedenkingen-bij-concepttekst-ai-act, 6 February 2024.

25 Charpentier & Vamparys, *Big Data & Society* 2025/12.

insurers may collect the sensitive data necessary for correcting those biases.[26] Or at the very least, the legislator could help by guiding insurers, which is in the spirit of the law: data protection law was not created for *hindering* non-discrimination law, but rather for helping it. Legislators should find the right balance between the two fields of law.

Sector-specific rules seem the most necessary, as follows from the conflict between the AI Act and non-discrimination law.[27] In subsection 3.2, I argued what sets insurance part from another sector such as the energy sector. Another argument is that general laws seem available, but specific laws seem to be missing or inadequate. By creating the AI Act, the legislator created norms that apply to all sectors: a *lex generalis*. Yet for insurance, *lex specialis* rules seem to be missing or how they fit with the AI Act is unknown.

4 Limitations and interdisciplinary research agenda

The thesis has some limitations. On the factual side, the thesis did not interview insurers on how they use AI in practice. It is possible that insurers use far more AI in practice than what is publicly known.[28] I (partly) compensated for that limitation by using surveys by the European Insurance and Occupational Pensions Authority and reports by insurance associations as a source, which already show what type of AI insurers used 'behind the scenes' in 2023.[29] A second limitation follows from the scope of the thesis: it is possible that AI applications in insurance have changed the sector in other parts of insurance than underwriting.[30]

26 For a similar view on data protection and fairness/justice, see e.g. Van Hoboken/Van Der Sloot, Broeders & Schrijvers 2016, p. 252.

27 Existing work also argues for sector-specific regulation. According to Zuiderveen Borgesius, the risk pooling in insurers warrants a sector-specific approach. See Zuiderveen Borgesius, *The International Journal of Human Rights* 2020/24, p. 1585. See also the next section.

28 For a review of scientific papers discussing AI applications in automotive, health, and property insurance, see e.g. Bhattacharya and others, *Frontiers in Artificial Intelligence* 2025/8.

29 A survey conducted by the European Insurance and Occupational Pensions Authority in June 2023 shows that 'to date insurance undertakings have predominantly opted for more simple and explainable AI algorithms such as advanced regressions or tree-based algorithms'. EIOPA 2024, p. 26.

30 See the research agenda at the end of this chapter.

The conclusions of the legal research are limited by the following circumstances. First, the thesis does not focus on national laws. In certain EU countries, for example, income is a protected characteristic, in which case non-discrimination law applies for that category.[31] Another example is that, in specific contexts, there may be an exception in national laws allowing for AI de-biasing, which is not accounted for by the thesis.

Second, I did not investigate every possibly relevant law. While the AI Act seems relevant, the AI Act came into force relatively late in my PhD project and is long and detailed.[32] Healthcare insurance is heavily regulated in the EU and heavily fragmented across member states: perhaps, some of the threats in the thesis do not apply to healthcare in specific countries. In my opinion, such topics would benefit from a dedicated PhD thesis.

The insights about unfair differentiation are limited by the following. It is possible that more aspects of fairness exist that were not discussed in the thesis or the work it is based on. For a decision to be fair, it is important to decide – in context – based on all the circumstances of the case, whether the decision is fair. For example, in health insurance, characteristics that consumers cannot influence are more problematic than in car insurance. The sector the differentiation occurs in, is therefore also a possible factor, and can be a mitigating circumstance in deciding whether a certain differentiation is unfair. Many more such factors could exist, and a perfect 'fairness assessment' seems impossible.[33]

Another limitation of unfair differentiation is in its definition: whether a certain differentiation *feels* unfair even if it is not covered by current laws.[34] For Chapter V, I chose to interpret this in terms of consumer perceptions, even if the survey does not aim to measure the consumer's feelings.

31 Socio-economic status is recognized in at least ten EU countries. Ganty & Benito Sanches 2021, p. 95.

32 See the research agenda later in this section. In my opinion, one could write an entire thesis about the AI Act and insurance. Chapter IV offers some anecdotal evidence for that claim: an entire chapter about a mere subsection of a single legal provision in the AI Act.

33 Work on algorithmic fairness supports this. See e.g. Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023. De Jonge & Hiemstra, *ACM Conference on Fairness, Accountability, and Transparency* 2023, p. 837.

34 For definitions, see the Key Terms in Chapter 1, Table 1.

The survey I presented in Chapter V has some limitations. It is a first indication of what people think, but does not account for *why* people may have certain perceptions, such as why people find a certain type of data irrelevant. How consumers perceive fairness may also depend on whether the insurer includes an explanation.[35]

Chapter V is intended as a first survey, not (for example) an analysis based on focus groups, which could help uncover the reasons for people's perceptions.[36] Finally, Chapter V may suffer from what I call in that chapter the *fairness paradox*: people's perceptions and behaviour may differ.[37] It is unclear how much we can trust on people's stated preferences when it comes to fairness. On the other hand, as I argued in that chapter, people's stated preferences can still inform policy.

In this thesis, I argue that before legislators intervene to regulate AI in insurance, far more insights from different disciplines are necessary. In my opinion, that is not a limitation of the current research. In the rest of this section, I present an interdisciplinary research agenda for the non-discrimination and unfair differentiation of AI in insurance sector.

4.1 Research agenda: topics for different disciplines

The research has revealed new angles for disciplinary and interdisciplinary research. *Philosophers and ethicists* could think about a fairness framework for the insurance sector: perhaps more factors exist. Philosophers could look at older, existing factors: perhaps many factors of fairness become a bigger threat when adding AI systems to the analysis just as, for example, it becomes more likely that insurers might choose a 'seemingly illogical characteristic'.[38]

There is still much work for *lawyers*. In the thesis, I did not research the entire AI Act in relation to insurance. At a glance, the AI Act does affect

35 While fairness and explainability of AI are distinct fields in computing science literature, the two can arguably interact in practice, influencing the *perceptions* of consumers. See footnote 50 of this chapter.

36 See also the explanation in Chapter 1, Section 5 (Method).

37 Future research could clarify this, see research agenda. The fairness paradox seems akin to the privacy paradox, see the conclusion of Chapter V.

38 Another example: philosophers could refer to the Rawlsian veil of ignorance for discussing 'the poor paying more' or actuarial fairness. Hildebrandt 2020, p. 305.

the entire insurance sector, despite many insurances not being high-risk. Yet some provisions in the AI Act could have consequences for insurance.[39] Perhaps the social scoring ban in Article 5 AI Act could (politically or legally) affect how insurers underwrite.[40] Legal scholars could outline *how* the AI Act applies to insurance, and the overlap between the AI Act and obligations from the European Union's financial legislation.[41] Non-discrimination lawyers could continue from the work in Chapter II of the thesis, investigating whether behaviour-based insurance is discriminatory (see also the research agenda of Chapter II).[42] Another topic is whether legislators should amend the list of protected grounds in national laws to tackle data-intensive underwriting.[43]

Future research could look into what the duties in insurance-specific laws mean compared to the right to non-discrimination and the EU AI Act.[44] While the AI Act is partly a product safety regulation, the insurance sector already had a type of product safety regime before the AI Act: the Product Approval and Review Process (PARP) from the Insurance Distribution Directive.[45] I did

39 For example, rules about AI using biometric identification could affect 'pay-as-you-live' insurance, where a company monitors whether the consumer smokes. See Glintić & Bezbradica 2025, p. 109.

40 D. Minty, 'The Social Scoring Ban – Confusion for Insurers', https://www.ethicsandinsurance.info/the-social-scoring-ban-confusion-for-insurers/, 19 February 2025. D. Minty, 'Why the Social Scoring Ban is Seismic for Counter Fraud', https://www.ethicsandinsurance.info/why-the-social-scoring-ban-is-seismic-for-counter-fraud/, 4 September 2023.

41 For a global discussion focusing on credit underwriting, see Hacker & Eber, *Harvard Data Science Review* 2025/7.

42 For example, one view is that behaviour-based insurance results in discrimination, e.g. because men might drive differently than women overall. Hildebrandt 2015, p. 96. Another view: Regardless of gender, individuals can change their behaviour. Rebert & Van Hoyweghen, *Tijdschrift voor Genderstudies* 2015/18; Thiery 2010, p. 565. See also the research agenda of Chapter II.

43 See for a relevant discussion e.g. (in Dutch) Terlouw, *Nederlands Tijdschrift voor de Mensenrechten* 2022/4.

44 For example, the Insurance Distribution Directive, national laws that implement the directive such as the Wet op het Financieel Toezicht.

45 Directive (EU) 2016/97 of the European Parliament and of the Council of 20 January 2016 on insurance distribution (IDD), https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016L0097. See in particular Chapter VI of the IDD. The European Commission worked out these norms in a delegated regulation: Commission Delegated Regulation (EU) 2017/2358 of 21 September 2017 supplementing Directive (EU) 2016/97

not research this directive and its implementation in national law, because it has a financial rationale rather than focusing on non-discrimination. The directive could require insurers to explain their choice for a certain characteristic better.[46] Another topic is that certain insurance law-specific principles could bring notification obligations for the insurer or consumer.[47]

Scholars could dive into whether the threats identified in this thesis apply in the healthcare sector, or look for tension between healthcare regulations, the AI Act, non-discrimination law and the Right to Health. Scholars could investigate how the exception in Article 10(5) AI Act interacts with health insurance laws. Finally, lawyers could investigate whether national laws and self-regulatory frameworks cover the threats mentioned in the thesis.

Computing Scientists could discuss which insights of the thesis can be translated into fairness metrics, and make concrete what role such metrics could play in assessing discrimination risks.[48] Since AI systems can create new types of vulnerable (algorithmic) groups, researchers could combine research on vulnerable groups as a basis for clustering algorithms, allowing such algorithms to identify new vulnerable groups in datasets.[49]

---

of the European Parliament and of the Council with regard to product oversight and governance requirements for insurance undertakings and insurance distributors, https://eur-lex.europa.eu/eli/reg_del/2017/2358/oj/eng. In the Netherlands, these norms are further encoded in the 'Besluit Gedragstoezicht financiële ondernemingen Wft'.

46 Note that fairness and explainability are distinctly different: it is possible to plausibly explain an unfair characteristic.

47 For example, It is unclear what the principle of uberrima fides (utmost good faith) means for AI in insurance. See (in Dutch) Nelemans & Van Ark, *Nederlands Tijdschrift voor Handelsrecht* 2024/6. On the principle (in Dutch) Hendrikse & Rinkes/Hendrikse, Van Huizen & Rinkes 2023, ch. 6.

48 There is a distinction between 'measuring' fairness using a metric, and normatively weighing whether a case of illegal discrimination has taken place. The interaction between measuring and weighing could be studied further. Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023. For work in the direction of making a translation, see e.g. Wachter, Mittelstadt & Russell, *SSRN Electronic Journal* 2021. Sauce, Chancel & Ly 2023. Sargeant & Magnusson, *Conference on Fairness, Accountability, and Transparency* 2025.

49 For example, clustering algorithms such as 'Hierarchical Bias-Aware Clustering' can identify vulnerable groups in datasets, but require a fairness metric to make the groups. Researchers could translate work on vulnerability into new metrics for such algorithms. Misztal-Radecka & Indurkhya, *Information Processing & Management* 2021/58.

In computer science literature, explainability and fairness are separate fields.[50] However, insurers must 'sell a product that the buyer can understand': explanations and fairness may interact, both influencing the fairness *perception* of a decision.[51] Future research could focus on the interaction between explanations of the fairness of a decision and how people perceive the fairness of a decision.

Social scientists can help in understanding many of the issues regarding AI in insurance from different perspectives. Publicly available work focusing on the *consumer* perspective on AI and insurance remains scarce.[52] Future work could continue where the survey in Chapter V left off, answering whether a fairness paradox indeed exists and uncovering the motivations behind the respondents' perceptions (see also the research agenda of Chapter V). Social science scholars can also reflect about whether policymakers can trust on people's stated preferences when assessing fairness.

Scholars in *economics and sociology* could clarify whether some information asymmetry is important in insurance: in insurance, contrary to other sectors, information asymmetry seems important for the sector to exist.[53] More generally, scholars in finance can investigate how and whether AI is changing the insurance sector.[54] Finally, empirical work can be important

S. Muhammad, 'Auditing Algorithmic Fairness with Unsupervised Bias Discovery', http://amsterdamintelligence.com/posts/bias-discovery, 14 September 2021.

50 A subfield of AI explainability is how to *explain* whether a decision is fair. See e.g. Zhou, Chen & Holzinger, *xxAI - Beyond Explainable AI* 2020, p. 375-386. Shulner-Tal, Kuflik & Kliger, *Ethics and Information Technology* 2022/24. Binns and others 2018.

51 D. Minty, 'How to Explain the Fairness of a Particular Rating Factor', https://www.ethicsandinsurance.info/explaining-the-fairness-of-a-particular-rating-factor/, 10 February 2022. See for previous work about the readability of insurance contracts and consumer expectations, for example, Van Boom, *Journal of Consumer Policy* 2011/34. Van Boom, Desmet & Van Dam, *Journal of Consumer Policy* 2016/39.

52 It seems that most work of the work on insurance by social scientists does not focus on the experiences of the consumer. Tanninen, *Big Data & Society* 2020/7, p. 9. For a general literature review of empirical work on the perceptions of individuals regarding fairness, see Starke and others, *Big Data & Society* 2022/9.

53 See Eling, Nuessle & Staubli, *The Geneva Papers on Risk and Insurance - Issues and Practice* 2022/47. SCOR 2018, p. 8.

54 For a recent paper in that direction, see Kirov, *Financial Navigator Journal* 2025/10.

in discovering whether cases of discrimination by AI may be occurring in insurance.[55]

There is probably room for scholars from many more disciplines within the topic of the thesis. To give an example: *ethnographers* could focus on defining vulnerability, identifying possible vulnerable groups affected by AI in insurance. The list goes on.

4.2 Research agenda: Interdisciplinary topics

Some topics go beyond the scope of this research, and are interdisciplinary. First, interdisciplinary scholars could focus on different types of AI technology. Future research could look into how generative AI systems affect the workflow of the insurer. For example, Microsoft has introduced an experimental version of their Large Language Model *Copilot* for specifically insurance.[56] Another approach is to compare AI developments in insurance across different countries: in this thesis, I did not compare explicitly between different countries. Another focus would be non-risk-based pricing in insurance, such as Willingness to Pay or other types of non-risk-based pricing.[57]

Future work could continue the research on data-intensive underwriting and behaviour-based insurance. For example, perhaps the *combination* of behaviour-based insurance and data-intensive underwriting creates new problems as well. A caveat, as I wrote in Chapter II, is that the two trends I presented are not completely separate: insurers can collect the data for pattern-matching or aggregation. In other words, while I treat both trends differently, we can 'mix the trends up' to analyse the effects further.

55 See, for example, Rondina and others 2025. Raji & Buolamwini, *Conference on AI, Ethics, and Society* 2019.

56 Quess GTS, 'Microsoft 365 Copilot to Revolutionize Insurance Underwriting', https://www.quessgts.com/empowering-insurance-microsoft-365-copilots-journey-in-revolutionizing-underwriting/, https://quessgts.com/, 3 April 2024.

57 EIOPA warns that such practices are becoming more and more sophisticated, currently especially in the United Kingdom. EIOPA 2023. Hildebrandt warns for a mixing of code-driven law with AI in insurance: 'a smart contract may trigger an increase in the premium of my car insurance after my car has detected a certain threshold of fatigue or risky driving.' Hildebrandt/Deakin & Markou 2020, p. 67. See also Hildebrandt 2015.

It is also not black-and-white whether behaviour-based insurance implies truly individualised insurance pricing. According to Barocas, Hardt & Narayanan, insurers do not only look at individuals when they decide a behaviour-based insurance price: they *compare* individuals, looking at patterns indicating (for example) unsafe driving. Based on such patterns, the insurers build a model. The argument is that behaviour-based insurance is not true individualization.[58] In Chapter II, I presented a different view of what 'individualization' means: individuals can change their behaviour in real-time, choosing to drive more safely or not, which decides their discount.[59]

Future work can analyse whether there are different *types* of individualization at stake, and perhaps show these on a scale from individual to the group level. Such research may be relevant for scholars from all disciplines in the field of non-discrimination, fairness, AI and insurance.

Another suggestion would be to focus on *re*insurers: insurers who insure insurers, in case of (for example) difficult insurances or insurance against large disasters. Many existing insurances depend on reinsurers.

I see the research agenda that this section presents as an initial step towards understanding non-discrimination and AI in the insurance sector, showing that there is ample inspiration for both disciplinary and interdisciplinary research. The future will reveal how AI develops in insurance.

5 Looking back: where Law meets Artificial Intelligence

Some final reflections about interdisciplinary research. Interdisciplinary research answered new questions about the topic of non-discrimination by AI, that non-discrimination law alone could not have answered. It also uncovered new issues. It offered a lens for a more holistic perspective

58 Barocas, Hardt & Narayanan 2023, p. 36. A related argument by sociologists exists in the context of 'persons' (risk personification in insurance): Moor & Lury, *Journal of Cultural Economy* 2018/11. McFall & Moor, *Distinktion: Journal of Social Theory* 2018/19.

59 See also footnote 58. Perhaps the difference in opinion can be explained because much of the sociological work on behaviour-based insurance comes from the fields of *critical data studies* and *sociology of insurance*, which have their own views and ways of thinking. See Tanninen, *Big Data & Society* 2020/7, p. 9.

by including other disciplines.[60] Overall, it seems that including other disciplines in legal research is not only worth it, but *necessary*, to investigate the discrimination that artificial intelligence could bring. For that reason, I presented an interdisciplinary research agenda.

In fact, two disciplines such as Law and Computing Science do not seem enough: society must be involved and more fact-finding would be useful, requiring social scientists. The economic nature of insurance requires involving economists, sociologists and actuaries. Perhaps in the future, for Law to meet Artificial Intelligence, increasingly more interdisciplinarity is necessary.

60 For example, while the *special category data* problem is a hurdle for insurers, there are also practical reasons for why insurers do not innovate that would not surface from purely legal research. For a paper discussing possible hurdles, see footnote 22 of this chapter.

References of chapter VI

Appelman, Fathaigh & Van Hoboken, *JIPITEC* 2021/12
N. Appelman, R.Ó. Fathaigh & J. van Hoboken, 'Social welfare, risk profiling and fundamental rights: The case of SyRI in The Netherlands', *JIPITEC* 2021/12, p. 257.

Barocas, Hardt & Narayanan 2023
S. Barocas, M. Hardt & A. Narayanan, *Fairness and Machine Learning*, MIT Press 2023, https://fairmlbook.org/.

Bhattacharya and others, *Frontiers in Artificial Intelligence* 2025/8
S. Bhattacharya and others, 'AI revolution in insurance: bridging research and reality', *Frontiers in Artificial Intelligence* 2025/8, p. 1568266, https://www.frontiersin.org/articles/10.3389/frai.2025.1568266/full, doi:10.3389/frai.2025.1568266.

Binns and others 2018
R. Binns and others, *'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions* (preprint), SocArXiv 2018, https://osf.io/9wqxr, doi:10.31235/osf.io/9wqxr.

Charpentier & Vamparys, *Big Data & Society* 2025/12
A. Charpentier & X. Vamparys, 'Artificial intelligence and personalization of insurance: Failure or delayed ignition?', *Big Data & Society* 2025/12, issue 1, p. 20539517241291817, https://journals.sagepub.com/doi/10.1177/20539517241291817, doi:10.1177/20539517241291817.

College voor de Rechten van de Mens 2014
College voor de Rechten van de Mens, *Advies aan Dazure B.V. over premiedifferentiatie op basis van postcode bij de Finvita overlijdensrisicoverzekering*, 2014, https://publicaties.mensenrechten.nl/publicatie/29c613ac-efab-4d2a-a26c-fe1cb2556407.

De Jonge & Hiemstra, *ACM Conference on Fairness, Accountability, and Transparency* 2023
T. De Jonge & D. Hiemstra, 'UNFair: Search Engine Manipulation, Undetectable by Amortized Inequity', *ACM Conference on Fairness, Accountability,*

and Transparency 2023, https://dl.acm.org/doi/10.1145/3593013.3594046, doi:10.1145/3593013.3594046.

EIOPA 2023
EIOPA, 'Supervisory statement on differential pricing practices in non-life insurance lines of business EIOPA-BoS-23/076', 2023.

EIOPA 2024
EIOPA, *Report on the digitalisation of the insurance sector EIOPA-BoS-24/139*, 2024, www.eiopa.europa.eu/publications/eiopas-report-digitalisation-european-insurance-sector_en.

Eling, Nuessle & Staubli, *The Geneva Papers on Risk and Insurance - Issues and Practice* 2022/47
M. Eling, D. Nuessle & J. Staubli, 'The impact of artificial intelligence along the insurance value chain and on the insurability of risks', *The Geneva Papers on Risk and Insurance - Issues and Practice* 2022/47, issue 2, p. 205-241, https://link.springer.com/10.1057/s41288-020-00201-7, doi:10.1057/s41288-020-00201-7.

Fröhlich & Williamson, *FAccT* 2024
C. Fröhlich & R.C. Williamson, 'Insights From Insurance for Fair Machine Learning', *FAccT* 2024, https://dl.acm.org/doi/10.1145/3630106.3658914, doi:10.1145/3630106.3658914.

Ganty & Benito Sanches 2021
S. Ganty & J.C. Benito Sanches, *Expanding the List of Protected Grounds within Anti-Discrimination Law in the EU (Equinet Report)*, Equinet 2021, https://equineteurope.org/publications/expanding-the-list-of-protected-grounds-within-anti-discrimination-law-in-the-eu-an-equinet-report/.

Glintić & Bezbradica 2025
M. Glintić & A. Bezbradica, *Artificial Intelligence as a cause of discrimination in insurance law*, Belgrade: Contemporary challenges in the protection of human rights 2025, https://zbornik.pravni.pr.ac.rs/index.php/test/article/view/135.

Hacker & Eber, *Harvard Data Science Review* 2025/7
P. Hacker & M. Eber, 'The Future of Credit Underwriting and Insurance Under the EU AI Act: Implications for Europe and Beyond', *Harvard*

Data Science Review 2025/7, issue 3, https://hdsr.mitpress.mit.edu/pub/19cwd6qx, doi:10.1162/99608f92.171157d2.

Hendrikse & Rinkes/Hendrikse, Van Huizen & Rinkes 2023
M.L. Hendrikse & J.G.J. Rinkes/M.L. Hendrikse, H.J.G. van Huizen & J.G.J. Rinkes, 'The notification obligation while entering an insurance contract ['De mededelingsplicht bij het aangaan van verzekeringen']', in: *Verzekeringsrecht (Recht en Praktijk Verzekeringsrecht)*, Deventer: Wolters Kluwer 2023, p. 245-316.

Hildebrandt 2015
M. Hildebrandt, *Smart Technologies and the End(s) of Law: novel entanglements of law and technology*, Cheltenham, UK Northampton, MA, USA: Edward Elgar Publishing 2015, https://china.elgaronline.com/monobook/9781849808767.xml.

Hildebrandt 2020
M. Hildebrandt, 'Closure: On Ethics, Code, and Law', in: M. Hildebrandt, *Law for Computer Scientists and Other Folk*, Oxford University Press: Oxford 2020, p. 283-318, https://academic.oup.com/book/33735/chapter/288379040, doi:10.1093/oso/9780198860877.003.0011.

Hildebrandt, *European Data Protection Law Review* 2021/7
M. Hildebrandt, 'Discrimination, Data-driven AI Systems and Practical Reason', *European Data Protection Law Review* 2021/7, issue 3, p. 358-366, http://edpl.lexxion.eu/article/EDPL/2021/3/4,doi:10.21552/edpl/2021/3/4.

Hildebrandt/Deakin & Markou 2020
M. Hildebrandt/S. Deakin & C. Markou, 'Code-driven Law: Freezing the Future and Scaling the Past', in: *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence*, Hart Publishing 2020, http://www.bloomsburycollections.com/book/is-law-computable-critical-perspectives-on-law-and-artificial-intelligence, doi:10.5040/9781509937097.

Kirov, *Financial Navigator Journal* 2025/10
S. Kirov, 'InsurAI – The Overtaking Lane in The Insurance Race', *Financial Navigator Journal* 2025/10, issue 1, p. 58-68, https://doi.org/10.56065/FNJ2025.1.58, doi:10.56065/FNJ2025.1.58.

McFall & Moor, *Distinktion: Journal of Social Theory* 2018/19
L. McFall & L. Moor, 'Who, or what, is insurtech personalizing?: persons, prices and the historical classifications of risk', *Distinktion: Journal of Social Theory* 2018/19, issue 2, p. 193-213, https://www.tandfonline.com/doi/full/10.1080/1600910X.2018.1503609, doi:10.1080/1600910X.2018.1503609.

Van Meerten & Schmidt, *Pensioen magazine* 2015
H. van Meerten & E.S. Schmidt, 'De woekerpoliszaak: een kwestie van interpretatie?', *Pensioen magazine* 2015, https://dspace.library.uu.nl/handle/1874/326985.

Misztal-Radecka & Indurkhya, *Information Processing & Management* 2021/58
J. Misztal-Radecka & B. Indurkhya, 'Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems', *Information Processing & Management* 2021/58, issue 3, p. 102519, https://www.sciencedirect.com/science/article/pii/S0306457321000285, doi:10.1016/j.ipm.2021.102519.

Moor & Lury, *Journal of Cultural Economy* 2018/11
L. Moor & C. Lury, 'Price and the person: markets, discrimination, and personhood', *Journal of Cultural Economy* 2018/11, issue 6, p. 501-513, https://doi.org/10.1080/17530350.2018.1481878, doi:10.1080/17530350.2018.1481878.

Nelemans & Van Ark, *Nederlands Tijdschrift voor Handelsrecht* 2024/6
M.D.H. Nelemans & D.B. van Ark, 'AI, big data and the pre-contractual duty of disclosure in insurance law ['AI, big data en de precontractuele mededelingsplicht in het verzekeringsrecht']', *Nederlands Tijdschrift voor Handelsrecht* 2024/6, https://repository.ubn.ru.nl/handle/2066/313873.

O'Neil 2016
C. O'Neil, *Weapons of math destruction: how big data increases inequality and threatens democracy*, New York: Penguin Books 2016.

Raji & Buolamwini, *Conference on AI, Ethics, and Society* 2019
I.D. Raji & J. Buolamwini, 'Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products',

Conference on AI, Ethics, and Society 2019, https://dl.acm.org/doi/10.1145/3306618.3314244, doi:10.1145/3306618.3314244.

Rebert & Van Hoyweghen, *Tijdschrift voor Genderstudies* 2015/18
L. Rebert & I. Van Hoyweghen, 'The right to underwrite gender: The Goods & Services Directive and the politics of insurance pricing', *Tijdschrift voor Genderstudies* 2015/18, issue 4, p. 413-431, https://www.aup-online.com/content/journals/10.5117/TVGN2015.4.REBE, doi:10.5117/TVGN2015.4.REBE.

Rondina and others 2025
M. Rondina and others, 'Testing software for non-discrimination: an updated and extended audit in the Italian car insurance domain [Preprint]', 2025, http://arxiv.org/abs/2502.06439, doi:10.48550/arXiv.2502.06439.

Sargeant & Magnusson, *Conference on Fairness, Accountability, and Transparency* 2025
H. Sargeant & M. Magnusson, 'Formalising Anti-Discrimination Law in Automated Decision Systems', *Conference on Fairness, Accountability, and Transparency* 2025, https://dl.acm.org/doi/full/10.1145/3715275.3732015, doi:10.1145/3715275.3732015.

Sauce, Chancel & Ly 2023
M. Sauce, A. Chancel & A. Ly, 'AI and ethics in insurance: a new solution to mitigate proxy discrimination in risk modeling [PREPRINT]', 2023, http://arxiv.org/abs/2307.13616, doi:10.48550/arXiv.2307.13616.

SCOR 2018
SCOR, *The impact of Artificial Intelligence on the (re)insurance sector*, SCOR 2018, https://www.scor.com/sites/default/files/focus_scor-artificial_intelligence.pdf.

Shulner-Tal, Kuflik & Kliger, *Ethics and Information Technology* 2022/24
A. Shulner-Tal, T. Kuflik & D. Kliger, 'Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system', *Ethics and Information Technology* 2022/24, issue 1, p. 2, https://link.springer.com/10.1007/s10676-022-09623-4, doi:10.1007/s10676-022-09623-4.

Starke and others, *Big Data & Society* 2022/9
C. Starke and others, 'Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature', *Big Data & Society* 2022/9, issue 2, p. 205395172211151, http://journals.sagepub.com/doi/10.1177/20539517221115189, doi:10.1177/20539517221115189.

Tanninen, *Big Data & Society* 2020/7
M. Tanninen, 'Contested technology: Social scientific perspectives of behaviour-based insurance', *Big Data & Society* 2020/7, issue 2, p. 205395172094253, http://journals.sagepub.com/doi/10.1177/2053951720942536, doi:10.1177/2053951720942536.

Terlouw, *Nederlands Tijdschrift voor de Mensenrechten* 2022/4
A. Terlouw, 'Classism: discrimination based on social status ['Klassisme: discriminatie op grond van sociale status']', *Nederlands Tijdschrift voor de Mensenrechten* 2022/4, https://njcm.nl/wp-content/uploads/2023/01/1.-NTM-47-4_Terlouw.docx.pdf.

Thiery 2010
Y. Thiery, *Discriminatie en verzekering. Economische efficiëntie en actuariële fairness getoetst aan het juridisch gelijkheidsbeginsel*, KU Leuven 2010, https://lirias.kuleuven.be/1844266&lang=en.

Van Bekkum & Zuiderveen Borgesius, *European Journal of Social Security* 2021/23
M. Van Bekkum & F. Zuiderveen Borgesius, 'Digital welfare fraud detection and the Dutch *SyRI* judgment', *European Journal of Social Security* 2021/23, issue 4, p. 323-340, https://journals.sagepub.com/doi/10.1177/13882627211031257, doi:10.1177/13882627211031257.

Van Boom, *Journal of Consumer Policy* 2011/34
W.H. Van Boom, 'Price Intransparency, Consumer Decision Making and European Consumer Law', *Journal of Consumer Policy* 2011/34, issue 3, p. 359-376, http://link.springer.com/10.1007/s10603-011-9163-8, doi:10.1007/s10603-011-9163-8.

Van Boom, Desmet & Van Dam, *Journal of Consumer Policy* 2016/39
W.H. Van Boom, P. Desmet & M. Van Dam, '"If It's Easy to Read, It's Easy to Claim"—The Effect of the Readability of Insurance Contracts on

Consumer Expectations and Conflict Behaviour', *Journal of Consumer Policy* 2016/39, issue 2, p. 187-197, http://link.springer.com/10.1007/s10603-016-9317-9, doi:10.1007/s10603-016-9317-9.

Van Hoboken/Van Der Sloot, Broeders & Schrijvers 2016
J. Van Hoboken/B. Van der Sloot, D. Broeders & E. Schrijvers, 'From collection to use in privacy regulation? A forward looking comparison of European and US frameworks for personal data processing', in: *Exploring the Boundaries of Big Data*, Wetenschappelijke Raad voor het Regeringsbeleid 2016, p. 231-252, https://www.ivir.nl/publicaties/download/1764.pdf.

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

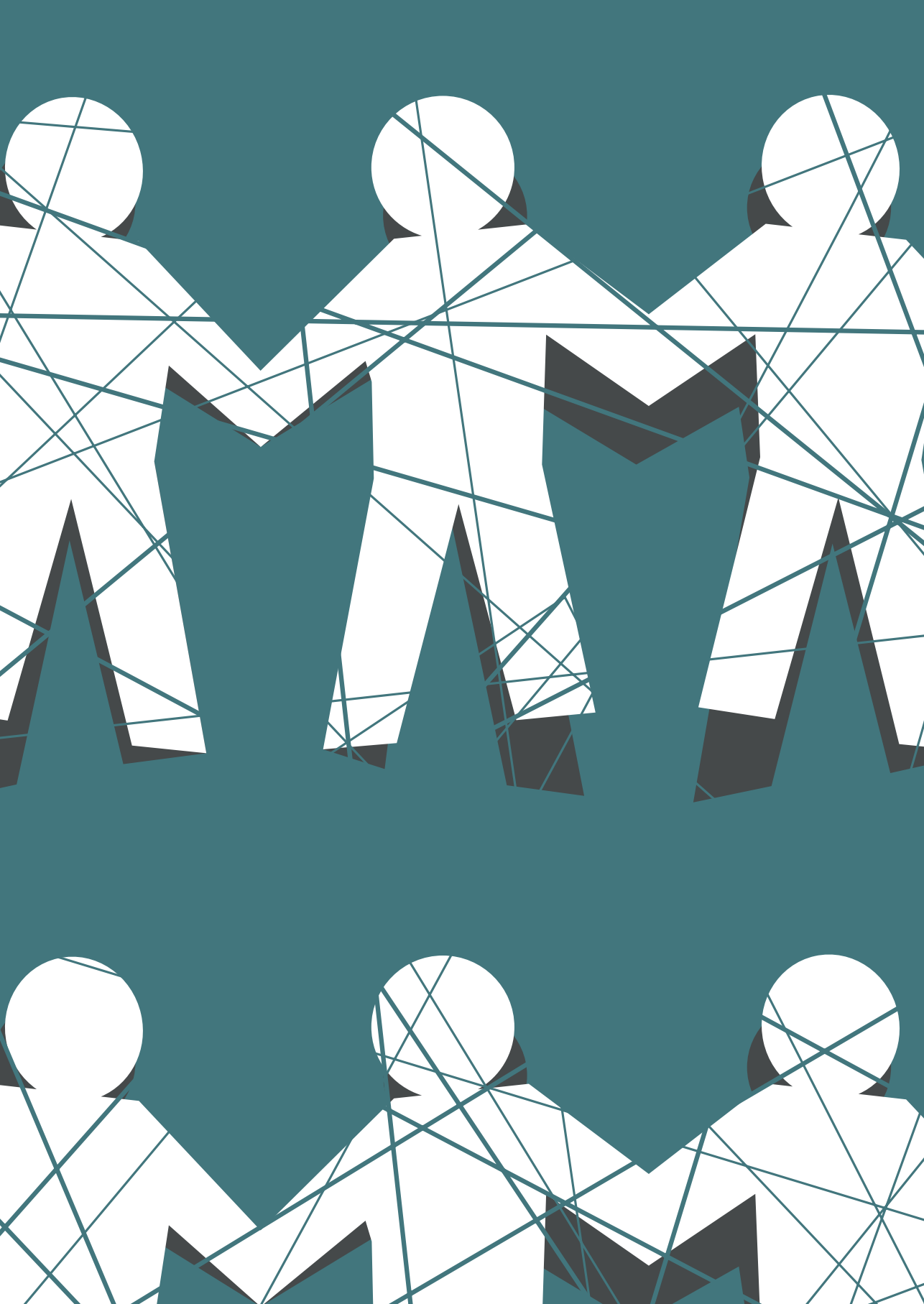
Weerts and others, *ACM Conference on Fairness, Accountability, and Transparency* 2023
H. Weerts and others, 'Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree', *ACM Conference on Fairness, Accountability, and Transparency* 2023, https://dl.acm.org/doi/10.1145/3593013.3594044, doi:10.1145/3593013.3594044.

Zhou, Chen & Holzinger, *xxAI - Beyond Explainable AI* 2020, p. 375-386
J. Zhou, F. Chen & A. Holzinger, 'Towards Explainability for AI Fairness', *xxAI - Beyond Explainable AI* 2020, p. 375-386.

Zuiderveen Borgesius, *The International Journal of Human Rights* 2020/24
F.J. Zuiderveen Borgesius, 'Strengthening legal protection against discrimination by algorithms and artificial intelligence', *The International Journal of Human Rights* 2020/24, issue 10, p. 1572-1593, https://www.tandfonline.com/doi/full/10.1080/13642987.2020.1743976, doi:10.1080/13642987.2020.1743976.

Gerechtshof Den Haag 26 September 2023, ECLI:NL:GHDHA:2023:1852 (*Vereniging Woekerpolis.nl*).

Hoge Raad 11 February 2022, ECLI:NL:HR:2022:166 (*Vereniging Woekerpolis.nl*).



Curriculum vitae

Marvin S.L. van Bekkum studied Dutch Law (LL.B., LL.M., Spec. Civil Law) and Computing Science (B.Sc., M.Sc., Spec. TRU/e master Cyber Security) at Radboud University Nijmegen. He is interested in the intersections between law and computing science, expressing them in his work. Marvin received a grant of €1000,- from the Netherlands Network of Human Rights Research, with which he funded a self-organised workshop during the term of his PhD. He was interviewed by VNAB Visie, a professional magazine by the Dutch business insurers (See also Other Contributions, page 252).

During his PhD, Marvin was involved in the organisation of several conferences and workshops:

- ACM Conference on Fairness, Accountability, and Transparency: Program Committee Member (reviewer)
- eLaw Symposium Leiden: Panel moderator
- iHub Brown Bag Sessions: Organiser
- See also Other Contributions, page 252.

Marvin peer reviewed for peer-review academic journals such as:

- Computer Law & Security Review
- Technology and Regulation (TechReg)
- International Data Privacy Law (IDPL)
- European Law Open
- Utrecht Law Review
- European Company and Financial law Review (ECFR)

Nederlandse samenvatting

Artificial Intelligence (AI) lijkt van invloed op de manier waarop verzekeraars risico's inschatten, prijzen bepalen en risico's beheren. Waar verzekeraars vroeger vooral statistische modellen gebruikten om de kans op claims te schatten, kunnen verzekeraars tegenwoordig grote hoeveelheden nieuwe soorten data analyseren — van Internet of Things (IoT)-apparaten tot sociale media. Daardoor kunnen zij risico's gedetailleerder inschatten. In dit proefschrift behandel ik de centrale onderzoeksvraag: In welke mate introduceren verzekeraars discriminatie-gerelateerde risico's door het gebruik van AI-systemen voor risico-onderschrijving? Ik behandel in dit proefschrift vier onderwerpen: trends in het gebruik van AI en de mogelijke risico's daarvan, de juridische obstakels voor het testen op bias (oftwel bevoordeeldheid, vertekeningen) onder de AVG, de effectiviteit van een uitzondering in de AI-verordening (AI Act), en hoe het publiek deze praktijken beoordeelt.

Het proefschrift laat twee belangrijke AI-trends zien in de moderne verzekeringswereld. De eerste, "data-intensive underwriting", betekent dat verzekeraars veel meer datapunten analyseren om risico's te schatten en premies te berekenen. Daarbij vertrouwen zij vaker op statistische verbanden in plaats van duidelijke oorzaken. In de tweede trend, "behaviour-based insurance", passen verzekeraars de premies aan op basis van het gedrag van de consument, bijvoorbeeld met apparaten zoals stappentellers of voertuigtrackers. Deze ontwikkelingen kunnen consumenten voordelen bieden, bijvoorbeeld lagere premies voor het hebben van een veilige levensstijl. Tegelijk brengen ze ook risico's met zich mee. Ze maken het bijvoorbeeld makkelijker voor verzekeraars om vooral mensen met een laag risico te accepteren ("cherry-picking") en mensen met een hoger risico uit te sluiten of weg te prijzen. Dit kan leiden tot nieuwe vormen van uitsluiting.

In het proefschrift maak ik een onderscheid tussen wettelijk verboden "discriminatie" en "oneerlijke differentiatie". Juridische discriminatie ontstaat wanneer verzekeraars verschillende premies of voorwaarden hanteren op

basis van wettelijk beschermde kenmerken zoals geslacht of etniciteit. "oneerlijke differentiatie" verwijst naar situaties waarin verzekeraars differentiëren op basis van kenmerken die niet wettelijk beschermd zijn, maar die mensen toch als oneerlijk ervaren. Een voorbeeld is het gebruik van zeer gedetailleerde data, zoals huisnummertoevoegingen (bijv. 11A versus 11B), om verschillende premies te rekenen aan buren met vrijwel hetzelfde risicoprofiel. Ik betoog in het proefschrift dat het gebruik van AI-systemen door verzekeraars dit soort verschillen waarschijnlijk zullen vergroten.

Om discriminatie te voorkomen, moeten verzekeraars controleren of hun AI-systemen bepaalde groepen benadelen. Zulke controles meten bijvoorbeeld hoe verschillende groepen vertegenwoordigd zijn in beslissingen van een AI-systeem. Daarbij lopen verzekeraars echter tegen een juridisch probleem aan. Om te testen of een systeem discrimineert tegen bijvoorbeeld een bepaalde etniciteit, moeten zij gegevens over die etniciteit verwerken. Artikel 9, lid 1, van de Algemene Verordening Gegevensbescherming (AVG) verbiedt in principe de verwerking van zulke "bijzondere categorieën van persoonsgegevens". Hierdoor ontstaat een spanningsveld: de AVG maakt het moeilijker om te controleren of AI-systemen discriminerend werken. Ik concludeer dat de AVG niet zelf in een uitzondering voorziet voor dit soort controles. Vervolgens bespreek ik de argumenten voor en tegen voor de stelling: heeft de AVG een nieuwe uitzondering nodig om dit probleem op te lossen. Ik zet onder meer de voordelen van betere biascontroles af tegen de risico's van het verzamelen van gevoelige gegevens.

Ik analyseer ook artikel 10, lid 5, van de Europese AI Act, dat een mogelijke uitzondering op artikel 9 AVG biedt. Deze bepaling staat aanbieders van "hoog-risico" AI-systemen toe om gevoelige gegevens te verwerken wanneer dat nodig is om bias op te sporen en te corrigeren, zolang zij strikte waarborgen toepassen. Dit is een belangrijke stap vooruit, maar er blijven beperkingen. De AI Act richt zich alleen op bias die leidt tot illegale discriminatie. Verzekeraars hoeven dus niet per se iets te doen aan "oneerlijke differentiatie". Daarnaast beschouwt de AI Act slechts bepaalde verzekeringen, zoals levens- en zorgverzekeringen, als "hoog-risico".

Om beter te begrijpen hoe mensen deze ontwikkelingen beoordelen, heb ik een enquête gehouden onder de Nederlandse bevolking. De resultaten laten zien dat respondenten veel moderne AI-gedreven acceptatiepraktijken

als oneerlijk ervaren, vooral wanneer ze de logica achter een beslissing niet begrijpen of wanneer ze leiden tot hogere premies voor armere consumenten. Tegelijkertijd zien respondenten gedragsverzekeringen als relatief eerlijk wanneer consumenten het gevoel hebben dat zij hun premie zelf kunnen beïnvloeden. Ideeën over rechtvaardigheid in verzekeringen worden niet alleen door de markt worden bepaald, maar ook door maatschappelijke opvattingen.

Ik concludeer dat het gebruik van AI in verzekeringen nieuwe risico's kan creëren voor rechtvaardigheid en non-discriminatie die de huidige wetgeving nog niet volledig opvangt. Hoewel de AI Act belangrijke instrumenten biedt om bias te detecteren en te corrigeren, richt deze wet zich vooral op juridische discriminatie. Daardoor blijven bredere vormen van oneerlijke behandeling grotendeels buiten beeld. Ik betoog dat wetgevers en toezichthouders sectorspecifieke regels moeten ontwikkelen.

Abstract

Artificial intelligence (AI) seems to be changing the way insurers assess risks, set prices, and manage risk. In the past, insurers mainly relied on statistical models to estimate the likelihood and impact of claims. Today, however, insurers can analyse large amounts of new types of data — ranging from Internet of Things (IoT) devices to social media. This allows insurers to estimate risks in greater detail. In this dissertation, I address the central research question: To what extent do insurers introduce discrimination-related risks when they use AI systems for underwriting? I organise the PhD thesis around four topics: trends in how insurers apply AI and the possible risks of these practices, the legal obstacle to testing for bias under the General Data Protection Regulation (GDPR), the effectiveness of an exception in the EU AI Act, and how the public evaluates these practices.

The PhD thesis identifies two important AI trends in the modern insurance sector. The first, 'data-intensive underwriting', means that insurers analyse far more data points to estimate risks and calculate premiums. In doing so, they more often rely on statistical correlations rather than clear causal relationships. The second trend, 'behaviour-based insurance', involves insurers adjusting premiums based on consumer behaviour, for example through devices such as step counters or vehicle trackers. These developments can benefit consumers, for instance through lower premiums for people with a safe lifestyle. At the same time, they also introduce risks. For example, AI makes it easier for insurers to focus on accepting people with a low risk ('cherry-picking') and to exclude or price out people with higher risks. These practices can create new forms of exclusion.

The thesis distinguishes between two types of discrimination: illegal 'discrimination' and 'unfair differentiation.' illegal discrimination occurs when insurers apply different premiums or conditions based on legally protected characteristics such as gender or ethnicity. 'Unfair differentiation' describes situations in which insurers differentiate on the basis of

characteristics that the law does not protect but that people nevertheless experience as unfair. For example, an insurer might use house number additions (such as 11A versus 11B) to charge different premiums to neighbours with almost identical risk profiles. In the PhD thesis, I argue that when insurers apply AI in underwriting, these kinds of differences will likely increase.

To prevent discrimination, insurers must verify whether the AI systems they operate may disadvantage certain groups. For example, insurers can examine how different groups appear in the decisions produced by an AI system. However, insurers encounter a legal problem at this stage: to test whether an AI system discriminates against, for example, a certain ethnicity, insurers must process data about that ethnicity. Article 9(1) of the General Data Protection Regulation (GDPR) in principle prohibits the processing of such 'special categories of personal data.' This creates a tension: data protection law makes it more difficult for insurers to verify whether the AI systems they operate behave in a discriminatory way. I conclude that the GDPR itself does not provide an exception for these types of checks. I then discuss the arguments for and against the claim that the GDPR needs a new exception to resolve this problem. Among other arguments, I weigh the benefits of better bias testing against the risks that come with collecting sensitive data.

I also analyse Article 10(5) of the European AI Act, which offers a possible exception to Article 9 GDPR. This provision allows providers of 'high-risk' AI systems to process sensitive data when they need those data to detect and correct bias, as long as they apply strict safeguards. This rule marks an important step forward, but important limitations remain. The AI Act focuses only on bias that leads to illegal discrimination. As a result, the law does not necessarily require insurers to address 'unfair differentiation'. In addition, the AI Act classifies only certain types of insurance, such as life and health insurance, as 'high-risk'. This choice creates a regulatory gap for other types of insurance, such as motor insurance, where insurers increasingly rely on AI.

To better understand how people evaluate these developments, I conducted a survey among the Dutch population. The results show that respondents consider many modern AI-driven underwriting practices unfair, especially when they do not understand the logic behind the practices or when the

practices lead to higher premiums for poorer consumers. At the same time, respondents view behaviour-based insurance as relatively fair when consumers feel that they can influence their premium themselves. Societal views therefore shape ideas about fairness in insurance alongside market forces. For this reason, policymakers and insurers should involve the public in the design of AI systems.

I conclude that in using AI, insurers can create new risks for fairness and non-discrimination that current legislation does not yet fully address. Although the AI Act provides important tools to detect and correct bias, the law focuses mainly on illegal discrimination. As a result, broader forms of unfair treatment largely remain outside its scope. I discuss whether lawmakers and regulators should develop sector-specific rules.

Acknowledgements (Thank-yous)

In defending this PhD thesis, I become a Doctor. Without the strength and support of many, the PhD thesis would not have made it into the world. In particular, I thank my father, mother for standing behind my PhD during the whole project. Especially because you are non-academics, I consider your support to my becoming an academic as no small feat. Similarly, I thank my sister for her love and support. I dedicate some of the strength and positivity with which I approach life and the thesis to the love and lessons by my grandparents, who are sadly no longer among the living to witness this achievement.

Without my PhD supervisors, Frederik and Tom, I could not have finished the thesis. With positivity rather than pressure, Frederik supervised my work and fueled my ambition. From the moment I started my internship with Frederik (before the PhD) I knew that he was the best supervisor I could wish for. I thank him for his enthusiasm, constructive criticism and professionalism. While they fueled my ambitions, Frederik and Tom also protected me from working too hard and supported me in managing my time. I learnt more lessons from both of them than I could write down in this section. A big THANK YOU to Frederik and Tom is an understatement.

Next, I thank my friends Michel and Robert for having my back for many years. We always had fun in hanging out together and occassionally in discussing my work. Michels first reaction was that he could see how getting a PhD fit me, and it seems he was right!

I also thank my neighbour and great friend Floor, for the time we spent together in Nijmegen. During the PhD, you supported me in the best, but also at the most difficult of times. You were always there to give me your great advice. Finishing a thesis such as this requires a dedication and mental fortitude. Or as you described it, the title PhD could also stand for 'Permanent head Damage' !

My colleagues at iHub were the great, positive, inspiring environment I needed. I especially thank Lotje. You were always there for me during the first three years of my PhD, even during the difficult pandemic years. Thank you to Marthe, Evi, Stein, Tim, Nina and Yiran for their humor and many inspiring conversations. Also thank you to my newer colleagues Liz, Jip and Yana, who supported me during the final year. I thank Tamar and Jaap-Henk for being role models and critical thinkers during the PhD. There are many more people at iHub to thank, and no acknowledgements section could do justice to everyone. I also say to you: Thank you!

I thank iCIS for funding the PhD, attending my presentations and the positive enthusiasm for this kind of research (different from the typical work in computer science). Finally, I thank all the people I met and the friends I made at conferences and venues. Without them, the research underlying the PhD thesis would have taken place in an ivory tower. Their contributions have been noted in a different section of the PhD thesis.

Author contributions

Attribution loosely based on Contributor Roles Taxonomy (CRediT):[1]
Brand and others, 'Beyond authorship: attribution, contribution, collaboration, and credit', *Learned Publishing* 2015/28, issue 2, https://onlinelibrary.wiley.com/doi/10.1087/20150211, doi:10.1087/20150211.

1 I made three main changes to the taxonomy: (i) added the term *valorisation* for contributions after a paper was published, but which (based on the paper) created value for the thesis output and had further societal impact. (ii) Added columns for non-authorship contributions, in line with the methodology of the thesis (*With thanks to* and *Conferences & Workshops*). (iii) Split the row for 'analysis' into 'data analysis' and 'other analysis' for Chapter V.

Chapter II

Author(s): Marvin van Bekkum (MvB), Frederik Zuiderveen Borgesius (FZB), Tom Heskes (TH)

With thanks to: Bart van der Sloot (BvdS), Paola Lopez (PL), Balázs Bodó (BB), Jos Schaffers (JS) and Maaike Harbers (MH).
Main conferences & workshops where we presented the draft paper: iHub Automated Decision-making Club (ADM), Radboud Data Science Seminar (DSS), Surveillance Studies Network (SV), Human(e) AI Conference (HAI), Digital Legal Talks (DLT)
Workshops (co-)organised by author: Netherlands Network of Human Rights Research Workshop (Org. MvB) (NNHRR (MvB)), Association for Computing and Machinery UMAP Conference (Org. MvB, FZB, Hanna Schraffenberger (HS)) (ACM (MvB,FZB, HS))[2]

	Substantial contribution	With thanks to	Conf. & Worksh.
Conceptualization	MvB, FZB		ACM(MvB,FZB, HS)
Methodology	MvB, FZB		
Analysis	MvB, FZB		
Writing (original draft)	MvB	BvdS, PL, BB, JS, MH	NNHRR(MvB), DLT
Writing (review and editing)	MvB, FZB, TH		ADM, DSS, SV, HAI
Visualisation	MvB, FZB, TH		

2 For the full reference of the workshops, see 'Other contributions', page 252 of the thesis.

Chapter III

Author(s): Marvin van Bekkum (MvB)*, Frederik Zuiderveen Borgesius (FZB)*
*joint first author

With thanks to: Tom Heskes (TH), Michael Veale (MV), Raphaële Xenidis (RX) and Steven Bellovin (SB)
Main conferences & workshops where we presented the draft paper: Privacy Law Scholars Conference (PLSC), Data Protection Scholars Network (DPSN), Lorentz Workshop (LW), iHub Brown Bag Session (BBS), Human(e) AI Conference (HAI), CNIL Privacy Research Day (CNIL), Dutch Data Protection Authority (DDPA), Opening contribution Commissie Meyers conference (Meyers), UK Government Department for Science, Innovation and Technology (DSIT), Academy of European Law Seminar (ERA)

	Substantial contribution	With thanks to	Conf. & Worksh.
Conceptualization	MvB, FZB		LW, DPSN, BBS, HAI
Methodology	MvB, FZB		
Analysis	MvB, FZB		
Writing (original draft)	MvB, FZB		
Writing (review and editing)	MvB, FZB	TH, MV, RX, SB	PLSC
Visualisation	MvB		
Valorisation			CNIL, DDPA, Meyers, DSIT, ERA

Chapter IV

Author(s): Marvin van Bekkum (MvB)

With thanks to: Frederik Zuiderveen Borgesius (FZB), Jaap-Henk Hoepman (J-HH), Tom Heskes (TH), Ylja Remmits (YR) and Francien Dechesne (FD)
Main conferences & workshops where I presented the draft paper: eLaw Conference Leiden (eLaw), TILTing Perspectives (TILT), iHub Automated Decision-making Club (ADM), Dutch Ministery of Internal Affairs (DMIA), Academy of European Law Seminar (ERA)

	Substantial contribution	With thanks to	Conf. & Worksh.
Conceptualization	MvB	FZB	
Methodology	MvB		
Analysis	MvB		
Writing (original draft)	MvB		
Writing (review and editing)	MvB	FZB, J-HH, TH, YR,FD	eLaw, TILT, ADM
Visualisation	MvB		
Valorisation			DMIA, ERA

Chapter V

Author(s): Marvin van Bekkum (MvB)*, Frederik Zuiderveen Borgesius (FZB)*, Gabi Schaap (GS), Iris van Ooijen (IvO), Maaike Harbers (MH), Tjerk Timan (TT)
*joint first author

With thanks to: Jeremias Adams-Prassl (JA-P)
Main conferences & workshops where we presented the draft paper: iHub Automated Decision-making Club (ADM), Radboud Data Science Seminar (DSS), Second ENS Fellows and Friends Paper Workshop (ENS), Dutch Association of Insurers (DAI) (upcoming, September 2025)

	Substantial contribution	With thanks to	Conf. & Worksh.
Funding acquisition	FZB, MvB, GS, IvO, MH, TT		
Conceptualization	MvB, FZB		
Methodology	GS, IvO, MvB, FZB		
Data collection	GS, IvO		
Data analysis	GS, IvO		
Other analysis	GS, IvO, MvB, FZB		
Writing (original draft)	MvB, FZB		
Writing (review and editing)	MvB, FZB, MH, TT	JA-P	ADM, DSS, ENS
Visualisation	GS, IvO		
Valorisation			DAI

Other contributions

Other published papers by author

M. van Bekkum & F. Zuiderveen Borgesius, 'Digital welfare fraud detection and the Dutch *SyRI* judgment', *European Journal of Social Security* 2021/23, issue 4, p. 323-340, https://journals.sagepub.com/doi/10.1177/13882627211031257, doi:10.1177/13882627211031257.

M.S.L. van Bekkum & F.J. Zuiderveen Borgesius, 'De spanning tussen het non-discriminatierecht en het gegevensbeschermingsrecht: Heeft de AVG een nieuwe uitzondering nodig om discriminatie door kunstmatige intelligentie tegen te gaan?', *Nederlands Juristenblad* 2023/1957.

G. Schaap, I. van Ooijen, F. Zuiderveen Borgesius & M. van Bekkum, 'What Drives Perceptions of AI-Driven Decision Making? Examining the Role of Data Relevance, Prediction Accuracy, and Decision Outcome', 2025. (preprint)

Blog posts

R. Gellert, M. van Bekkum & F. Zuiderveen Borgesius, 'The Ola & Uber judgments: for the first time a court recognises a GDPR right to an explanation for algorithmic decision-making', *EU Law Analysis* 28 April 2021, https://eulawanalysis.blogspot.com/2021/04/the-ola-uber-judgments-for-first-time.html.

F. Zuiderveen Borgesius & M. van Bekkum, 'Digital welfare fraud detection and the Dutch SyRI judgment', *International Association of Privacy Professionals* 23 September 2021, https://iapp.org/news/a/digital-welfare-fraud-detection-and-the-dutch-syri-judgment/.

M. van Bekkum & F. Zuiderveen Borgesius, 'Using sensitive data to prevent AI discrimination: Does the EU GDPR need a new exception?', *International Association of Privacy Professionals* 31 January

2023, https://iapp.org/news/a/using-sensitive-data-to-prevent-ai-discrimination-does-the-eu-gdpr-need-a-new-exception.

F. Zuiderveen Borgesius & M. van Bekkum, 'The AI Act's debiasing exception to the GDPR', *International Association of Privacy Professionals* 21 februari 2024, https://iapp.org/news/a/the-ai-acts-debiasing-exception-to-the-gdpr.

M. van Bekkum, F. Zuiderveen Borgesius & T. Heskes, 'AI, insurance, discrimination and unfair differentiation: An overview and research agenda', *International Association of Privacy Professionals* 14 May 2025, https://iapp.org/news/a/ai-insurance-discrimination-and-unfair-differentiation-an-overview-and-research-agenda/.

Interviews

Interview for *VNAB Visie* 2023/28, issue 2. VNAB visie is a professional magazine by the Dutch business insurers.

Workshops (co-)organised by author

F. Zuiderveen Borgesius, H. Schraffenberger & M. van Bekkum, 'Insurance, Algorithmic Decision-Making, and Discrimination', *Association for Computing and Machinery UMAP* online workshop, 22 June 2021, p. 333, https://doi.org/10.1145/3450614.3461451.

M. van Bekkum, 'Insurance, Algorithmic Decision-Making, and Discrimination', *Netherlands Network of Human Rights Research (NNHRR)* in-person workshop, 25 January 2023. Funded by a research grant of the NNHRR.

D. Minty & M. van Bekkum, Webinar 'Social Scoring and Insurance. The ban in Article 5 AI Act', *self-organised webinar*, 4 September 2024. Blog post based on webinar: D. Minty, 'The Social Scoring Ban – Confusion for Insurers', 19 February 2025, https://www.ethicsandinsurance.info/the-social-scoring-ban-confusion-for-insurers/.

Radboud Universiteit