Universiteit Antwerpen

Faculty of Business and Economics

Dissertation

Explaining prediction models to address ethical issues in business and society

Thesis submitted for the degree of Doctor of Applied Economics
at the University of Antwerp to be defended by
Sofie Goethals

Author:
Sofie Goethals

Supervisors:
Prof. dr. ir. David Martens
Prof. dr. Kenneth Sörensen

Antwerp, 2024

Supervisors
Prof. dr. ir. David Martens (University of Antwerp)
Prof. dr. Kenneth Sörensen (University of Antwerp)

Members of the Examination Committee
Prof. dr. Tim Verdonck (University of Antwerp) - *Chair of the Committee*
Prof. dr. Toon Calders (University of Antwerp)
Prof. dr. Theodoros Evgeniou (INSEAD Business School)
Prof. dr. Dolores Romero Morales (Copenhagen Business School)
Prof. dr. Foster Provost (NYU Stern)

This research was supported by the Research Foundation–Flanders (grant number: 11N7723N)

Research portfolio

Publications

The following publications are part of this thesis:

- Sofie Goethals, David Martens, and Theodoros Evgeniou. The Non-linear Nature of the Cost of Comprehensibility. *Journal of Big Data*, 9(1):1–23, 2022.
- Sofie Goethals, David Martens, and Toon Calders. PreCoF: Counterfactual Explanations for Fairness. *Machine Learning*, pages 1–32, 2023b.
- Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–24, 2023c.
- Sofie Goethals, David Martens, and Theodoros Evgeniou. Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (forthcoming)*. Springer, 2024c.
- Sofie Goethals, Sandra Matz, Foster Provost, Yanou Ramon, and David Martens. The Impact of Cloaking Digital Footprints on User Privacy and Personalization. *Under review*, 2024d.
- Sofie Goethals, Toon Calders, and David Martens. Beyond Accuracy-Fairness: Stop evaluating bias mitigation methods solely on between-group metrics. *Under review*, 2024a

The following publications were part of my research, but are not covered in this thesis:

- Raphael Mazzine, Sofie Goethals, Dieter Brughmans, and David Martens. Counterfactual explanations for employment services. In *International workshop on Fair, Effective And Sustainable Talent management using data science*, pages 1–7, 2021
- Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira, and David Martens. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, pages 1–21, 2022a.

- Travis Greene, Sofie Goethals, David Martens, and Galit Shmueli. Monetizing Explainable AI: A double-edged sword. *Under review*, 2023.

- Sofie Goethals, Eoin Delaney, Brent Mittelstadt, and Chris Russell. Resource-constrained fairness. *Under review*, 2024b

- Lauren Rhue, Sofie Goethals, and Arun Sundararajan. Evaluating LLMs for gender disparities in notable persons. *arXiv preprint arXiv:2403.09148*, 2024

Conference presentations

- Raphael Mazzine and Sofie Goethals. Counterfactual explanations for employment services. FEAST workshop at ECML: International workshop on Fair, Effective And Sustainable Talent management using data science, 2021 (online)

- Sofie Goethals. How counterfactual explanations can be used to detect bias in a machine learning model. EURO: 32th European Conference on Operational Research, Espoo, Finland, 2022.

- Sofie Goethals. The trade-offs of obscuring your digital footprints. ORBEL: 36th Annual Conference of the Belgian Operational Research Society, Liege, Belgium, 2023d

- Sofie Goethals. Explainability methods to measure discrimination in machine learning models. EWAF: European Workshop on Algorithmic Fairness, Winterthur, Switzerland, 2023c

- Sofie Goethals. The trade-offs of obscuring your digital footprints. ECDA: European Conference on Data Analysis, Antwerp, Belgium, 2023b

- Sofie Goethals. Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem. XKDD workshop at ECML: International Workshop on Explainble Knowledge Discovery in Data Mining, Turin, Italy, 2023e

- Sofie Goethals. Explainability methods to measure discrimination in machine learning models. BIAS workshop at ECML: Third Workshop on Bias and Fairness in AI, Turin, Italy, 2023a

- Sofie Goethals, Travis Greene, David Martens, and Galit Shmueli. Algorithmic Explanations as Ad Opportunities. Poster in the Workshop on Decision Intelligence and Analytics for Online Marketplaces at KDD, Long Beach, California, 2023a (online)

- Sofie Goethals and Toon Calders. Reranking individuals: The effect of fair classification within-groups. Poster in the European Workshop on Algorithmic Fairness, Mainz, Germany, 2024

Other activities

- Participant in the EURO PhD School on Data-Driven Decision Making and Optimization, 10-22 June,2022 in Seville, Spain.

- Research stay at New York University under the supervision of Professor Foster Provost, January-May, 2023 in New York, USA.

- Research stay at the Oxford Internet Institute, under the supervision of Professor Brent Mittelstadt, January-March, 2024 in Oxford, UK.

- Women in Data Science (WiDS) Ambassador 2022 and co-organiser of the event *WiDS Antwerp @ University of Antwerp* on Thursday May, 12th 2022 in Antwerp, Belgium.

- Women in Data Science (WiDS) Ambassador 2024 and co-organiser of the event *WiDS Belgium @ VIB* on Friday May,17th 2024 in Ghent, Belgium.

Acknowledgements

First and foremost, I want to thank my advisor David Martens. Thank you for hiring me and helping me find my place in the research domain. I really appreciate the freedom you gave me to pursue my own interests while at the same time providing guidance and ideas. Thank you also for introducing me to so many interesting people in our field and supporting my research stays! I also want to thank my other advisor, Kenneth Sörensen. Thank you for your support in my applications in the last year, for the fruitful research collaboration on the privacy paper, and for always being available for a nice chat! Thank you to the other members of my Examination Committee, Tim Verdonck, Dolores Morales, Foster Provost, Toon Calders and Theodoros Evgeniou for their time and thorough evaluation of my work. Your valuable feedback has helped me to improve this thesis in its final stage. A special shout-out to Nele and Aline for all their administrative help throughout my PhD!

All the papers in this PhD were conducted in collaboration with excellent co-authors, who all deserve acknowledgment. Theos, I enjoyed working on my very first research project together, although all our meetings had to be virtual at the time. It was great finally meeting you in Antwerp and enjoying a real Belgian meal together in De Bomma! Toon, thank you for introducing me to the field of fairness. I greatly enjoyed our collaborations, which could not have been done without your expertise in this field. Foster, thank you for inviting me to NYU Stern for a research visit. I had a wonderful time there and had the opportunity to meet so many interesting people. I also especially appreciate your detailed feedback and contributions, both on the cloaking paper and this PhD thesis. Sandra, thank you for the productive and fun meet-ups at Columbia Business School or in coffee shops across New York. It was great working with you, and I look forward to our collaboration next year! Galit and Travis, thank you for your commitment and the collaboration during the monetization project. I also enjoyed spending some time with you both in Antwerp! Chris, Brent, and Eoin, thank you for the collaboration and welcoming me during my research stay at the Oxford Internet Institute! I learned a lot from working with you all, and hopefully, our paths will cross again in the future. I also want to thank the Department of Engineering Management for supporting my research stay at the University of Oxford.

Thank you to everyone in the Applied Data Mining Team, who have been the best colleagues! Dieter, you were both my office and teammate for the longest period during my PhD. You significantly boosted the fun level of everything in the PhD,

as well as the conferences we attended together. Ine, I am very glad you joined the team in my third year. You definitely more than met the 'Marmite' level a colleague should have :). I really enjoyed our trip to Paris and wish you the best of luck in your future endeavors! Yanou, thank you for convincing me to do the PhD and for being a great colleague when I started. I enjoyed our collaboration, afterworks, runs, organizing WiDS together, ... Stiene, although we only overlapped for one year, it was great to have you as a running/afterwork/Secret Santa/office mate! You were always up for any social activity, which I really appreciated. Tom, thank you for always being entertaining and supportive. James, thank you for introducing us to vegan restaurants throughout Antwerp and for our trip to Turin (where there was a surprising lack of vegan food). Raphael, I greatly valued your guidance during my PhD. Although we did not meet in person often, you were always there virtually. Also, thank you to Bjorge, Pieter, Camille, Lize, Stephan, and Timour, who were part of our team for a shorter time; it was nice getting to know all of you!

The time at the office would not have been the same without the presence of many other great colleagues. Thank you to everyone within the Department of Engineering Management, and the Faculty of Economics in general! Although we did not see much of each other initially due to COVID, the ENM Runs every week were a nice way to get together! Lissa, thank you for the fun times at conferences (hopefully not over yet!), the meet-ups in New York, and the many afterworks. Amy, thank you for always being kind and available for a good chat! Thank you, Noemi, Celine, Lissa, Tijmen, Luc, Dieter & Ine, for the amazing trip to New York! Also, a big thanks to all the colleagues from other departments we got to know better in the Brain Embassy. I will not put names to avoid missing anyone, but the last two years of my PhD would not have been the same without all the fun activities, such as karaoke nights, long breaks at the coffee machine, Christmas market visits, running clubs, soup lunches, Pilates sessions at the Release, drag bingo, Marmite evenings, going to the Sinksenfoor, and evenings at the Ossenmarkt! Lastly, thank you to all the amazing people I met at NYU Stern and the University of Oxford for being so welcoming such as Emma, Chloe, Philip, Eoin, Daria, Errin, Maria, Eunsol, Rubing, Rui, Connor, and many more.

Thank you to my family, especially my parents, who always supported me in whatever direction I wanted to go and never pushed me in another. Thank you for following my research journey with great interest. An additional thank you to my father for proofreading some of my papers, even when they were still in a very questionable state. To all my friends, thank you for being the best friends in the world and supporting me, even though most of you did not understand why I wanted to "keep studying". Finally, a huge thank you to Max for your endless support and interest, for discussing all my random ideas that might turn into a paper someday, for reading very rough drafts, for restyling many of the (initially very ugly) figures in this PhD, for upgrading my presentations with a "consultant touch", for making the best coffees, for your visits during my research stays, and for being great overall. Looking forward to our next adventure together!

Lastly, I acknowledge the support of the Flemish Research Foundation, who supported not only my doctoral grant but also provided funding for both my research stays.

Abstract

The field of artificial intelligence (AI) has experienced explosive growth in recent years, with applications ranging from medical diagnosis to financial forecasting. However, as these technologies become increasingly integrated into decision-making processes, it is crucial that we also consider the ethical implications of their use. In particular, the transparency, fairness and privacy of AI systems are major concerns, as these systems can have far-reaching impacts on individuals and society. In this PhD thesis, I focus on the ethics of explainable AI (XAI). XAI refers to the development of techniques that are able to provide human-understandable explanations for AI models. My research explores the importance of explainability in the context of ethical decision-making and investigates both opportunities and challenges that arise from the use of Explainable AI.

In this PhD thesis, I categorize my research contributions into three pillars: *Transparency*, *Fairness* and *Privacy*. Within the transparency pillar, I study the trade-off between transparency and performance of machine learning models, and investigate the manipulation issues that Explainable AI techniques can induce. Next, within the fairness pillar, I demonstrate how XAI techniques can be used to measure discrimination in machine learning models, and I discuss the opaqueness surrounding the impact of bias mitigation methods. Within the final pillar of privacy, I analyse the privacy issues of XAI techniques, and conduct an applied study to show the trade-off between privacy and personalization on a dataset of Facebook likes.

Dutch Abstract

Het domein van machine learning en data science heeft de afgelopen jaren een explosieve groei doorgemaakt, met toepassingen variërend van medische diagnose tot financiële voorspellingen. Echter, nu deze technologieën steeds meer geïntegreerd worden in besluitvormingsprocessen, is het cruciaal om de ethische implicaties van hun gebruik te bestuderen. Met name de transparantie, rechtvaardigheid en privacy van AI-systemen zijn essentiële kwesties, aangezien deze systemen vergaande gevolgen kunnen hebben voor individuen en onze samenleving. In dit proefschrift richt ik mij op de ethiek van begrijpbre AI (*Explainable Artificial Intelligence* in het Engels). XAI verwijst naar de ontwikkeling van technieken die menselijk begrijpbare verklaringen kunnen geven voor hun voorspellingen. Mijn onderzoek verkent het belang van transparentie in de context van ethische besluitvorming en onderzoekt zowel kansen als uitdagingen die voortvloeien uit het gebruik van XAI.

In dit proefschrift verdeel ik mijn onderzoeksbijdragen over drie pijlers: *Transparantie*, *Rechtvaardigheid* en *Privacy*. Binnen het domein van transparantie onderzoek ik de afweging tussen transparantie en performantie van machine learning modellen, en onderzoek ik de manipulatiekwesties die XAI-technieken kunnen veroorzaken. Vervolgens, binnen de pijler van rechtvaardigheid, toon ik aan hoe XAI-technieken kunnen worden gebruikt om discriminatie in machine learning-modellen te meten, en bespreek ik de onduidelijkheid rondom de impact van bias mitigatiemethodes. Binnen de laatste pijler van privacy analyseer ik de privacykwesties van XAI-technieken en voer ik een toegepaste studie uit om het conflict tussen privacy en personalisatie te demonstreren op een dataset van Facebook-likes.

Dutch Preface

"Het grootste risico bij AI is niet kwaadwilligheid, maar competentie. Een superintelligente AI is per definitie zeer bekwaam in het bereiken van zijn doelen, en als die doelen niet overeenkomen met de onze, hebben we een probleem." - Nick Bostrom, filosoof en schrijver.

Het Cambridge Analytica-schandaal, dat in 2018 uitbrak, legde de mogelijke gevaren bloot die gepaard gaan met het wijdverspreide verzamelen van gegevens en het gebruik ervan in Machine Learning. Het incident draaide om het misbruik van persoonlijke gegevens die zonder expliciete toestemming waren verkregen van miljoenen Facebook-gebruikers. Cambridge Analytica gebruikte deze gegevens om gedetailleerde psychologische profielen samen te stellen, wat de microtargeting van politieke advertenties met ongekende precisie mogelijk maakte. Dit incident diende als een scherpe herinnering aan de diepgaande invloed die technologie en gegevens kunnen uitoefenen op de vormgeving van de samenleving. Hoewel Kunstmatige Intelligentie (AI) de belofte inhoudt om innovatie te stimuleren en complexe problemen op te lossen, kan het ook het risico met zich meebrengen van privacyschending, discriminatie en de manipulatie van de publieke opinie.

Binnen deze context wordt de dringende noodzaak van ethische AI duidelijk. Dit veelzijdige veld omvat essentiële principes zoals transparantie, rechtvaardigheid, verantwoording en privacy, die allemaal van cruciaal belang zijn voor de bescherming van individuele rechten en de bevordering van het welzijn van de samenleving.

Met deze scriptie beoog ik bij te dragen aan dit vakgebied door de ethische kwesties rondom Explainable AI (XAI) te onderzoeken, met een specifieke focus op tegenfeitverklaringen (in het Engels *counterfactual explanations)*. Door de implicaties van counterfactual explanations binnen de ethische domeinen van transparantie, rechtvaardigheid en privacy te verkennen, streef ik ernaar de ontwikkeling van robuuste en ethisch verantwoorde AI-systemen te ondersteunen die in lijn zijn met het belang van de samenleving.

Deel I van deze scriptie legt de basis voor de rest van de scriptie. Het bevat een inleiding, een hoofdstuk over Machine Learning en een hoofdstuk over Ethical Machine Learning. Eerst introduceer ik het onderwerp van deze scriptie en geef ik een overzicht van mijn onderzoek en bijdragen in Hoofdstuk 1. In Hoofdstuk 2 bespreek ik de terminologie die gedurende de scriptie zal worden gebruikt. Vervolgens bespreek ik de classificatietechnieken en performantie-indicatoren die zullen worden toegepast om de prestaties van de Machine Learning modellen te meten. In

Hoofdstuk 3 wordt het veld van Ethical Machine Learning verkend door gebruik te maken van het FAT-framework. Ik organiseer mijn onderzoekscontributies langs drie pijlers van het FAT-framework: Transparantie, Rechtvaardigheid en Privacy. Het is echter belangrijk op te merken dat deze drie gebieden niet onderling uitsluitend zijn. Veel van mijn onderzoekscontributies hebben gevolgen voor meerdere gebieden, maar ik zal ze sorteren volgens de meest passende dimensie.

Deel II van deze scriptie richt zich op de bijdragen die het meest verband houden met het gebied van **transparantie**. In Hoofdstuk 4 analyseer ik de afweging tussen transparantie en performantie van machine learning modellen. Op basis van een analyse van 90 benchmark classificatiedatasets, maak ik de volgende bevindingen:

- Deze afweging bestaat voor de meeste (69%) van de datasets, maar in de meeste gevallen is deze vrij klein, terwijl deze voor slechts enkele datasets zeer groot is.

- De transparentie kan worden verbeterd door nog een andere algoritmische stap toe te voegen, namelijk die van het gebruik van zogenaamde 'surrogaat modellen'.

- Datasetkenmerken die verband houden met de complexiteit en het level van *noise* deze afweging significant kunnen verklaren.

Dit artikel is gepubliceerd in:

Sofie Goethals, David Martens, and Theodoros Evgeniou. The Non-linear Nature of the Cost of Comprehensibility. *Journal of Big Data*, 9(1):1–23, 2022.

In Hoofdstuk 5 analyseer ik de de risico's van manipulatie rond Explainable AI, wat een implicatie is van het *disagreement problem*. Dit probleem doet zich voor in het veld van XAI (Explainable AI) wanneer er meerdere verklaringen mogelijk zijn voor dezelfde AI-beslissing. Met dit onderzoek maak ik de volgende bijdragen:

- Een analyse van de verschillende strategieën die de aanbieders van een explanation kunnen inzetten om de gegeven explanation aan te passen ten voordele van henzelf.

- Een overzicht van verschillende doelstellingen en concrete scenario's die de aanbieders kunnen hebben om dit gedrag te vertonen.

Deze positie paper is gepresenteerd op de Workshop of Explainable Knowledge Discovery and Data Mining bij ECML en opgenomen in de post-workshopverslagen:

Sofie Goethals, David Martens, and Theodoros Evgeniou. Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem. In *Joint European*

Conference on Machine Learning and Knowledge Discovery in Databases (forthcoming). Springer, 2024c.

Deel III omvat de studies die verband houden met het gebied van **rechtvaardigheid**. In Hoofdstuk 6 analyseer ik hoe counterfactual explanations gebruikt kunnen worden om bias in machine learning modellen te beoordelen. Ik maak de volgende contributies:

- De introductie van *PreCoF*, ofwel *Predictive Counterfactual Fairness*, een nieuwe techniek om de counterfactual explanations samen te voegen.

- Aantonen dat *PreCoF* kan worden gebruikt om expliciete bias (wanneer het model het sensitive attribuut rechtstreeks gebruikt) te detecteren door te zoeken naar verklaringen die dit attribuut bevatten.

- Illustreren dat *PreCoF* ook kan worden gebruikt om impliciete bias te detecteren, wanneer het model het sensitieve attribuut niet direct gebruikt, maar wel andere gecorreleerde attributen gebruikt die kunnen leiden tot aanzienlijk nadeel voor de beschermde groep.

Dit artikel is gepubliceerd in een speciaal nummer over 'Fair and Safe Machine Learning' in:

Sofie Goethals, David Martens, and Toon Calders. PreCoF: Counterfactual Explanations for Fairness. *Machine Learning*, pages 1–32, 2023b.

In Hoofdstuk 7, breng ik transparantie naar het domein van bias mitigation methoden (methoden die proberen machine learning modellen minder bevooroordeeld te maken). De bijdragen van deze paper zijn als volgend:

- Inzichten in de operationele dynamiek van bias mitigation methoden en illustreren hoe sommige methoden significant invloed zullen hebben op de rangschikking binnen elke groep, terwijl anderen dat niet zullen hebben.

- Kritiek op de huidige benadering van het vergelijken van bias mitigation methoden, aangezien het situaties vergelijkt die significant verschillend zijn en niet toepasbaar in applicaties in de echte wereld.

In **Deel IV** bundel ik mijn bijdragen aan het gebied van **privacy** in Machine Learning. In Hoofdstuk 8 onderzoek ik de privacy problemen die counterfactual explanations kunnen veroorzaken. Dit leidt tot de volgende bijdragen:

- De introductie van een *explanation linkage attack*, die kan optreden wanneer instantiegebaseerde strategieën worden ingezet om counterfactual explanations te vinden.

- Een mogelijke oplossing: een algoritme CF-K om *k*-anonieme counterfactual explanations te creëren en de introductie van *puurheid* als een metriek om de *geldigheid* van deze explanations te beoordelen.

- De evaluatie van *k*-anonieme counterfactual explanations met CF-K door deze te vergelijken met een algoritme dat de hele dataset *k*-anonymiseert. Ik laat zien dat alleen de explanations *k*-anoniem maken, voordelig is voor de kwaliteit van de explanations.

Dit artikel is gepubliceerd in ACM Transactions on Intelligent Systems and Technology:

Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–24, 2023c.

In Hoofdstuk 9 bestudeer ik het effect op zowel privacy als personalisatie wanneer gebruikers een deel van hun digitale voetafdruk verbergen, ofwel *cloaken*, zoals ik dit in het artikel noem. Ik lever drie bijdragen aan de bestaande literatuur:

- Beoordeling van de langetermijn-effectiviteit van het cloaken van digitale voetafdrukken, waarbij het percentage van individuen die beschermd blijven gedurende een bepaalde tijdspanne, gemeten wordt. De resultaten tonen aan dat de effectiviteit van het cloaken van fijnmazige kenmerken gestaag en aanzienlijk afneemt voor de meeste inferentietaken.

- De introductie van een nieuw type cloaking-strategie gebaseerd op metafeatures, waarbij aangetoond wordt dat dit de langetermijnbescherming van het cloaken verbetert.

- Onderzoek naar de privacy-personalisatieafweging die inherent is aan het gebruik van cloaking om ongewenste inferenties te voorkomen. Specifiek laat ik zien dat cloaking voor één taak de voorspellende prestaties van andere personalisatietaken kan beïnvloeden. Bovendien wordt dit effect groter wanneer de metafeature-based cloaking gebruikt wordt, en dit toont het conflict aan tussen privacy en personalisatie waar users mee geconfronteerd worden.

Deze studie is momenteel onder review bij Big Data.

Deel V vat de conclusies van de thesis samen en lijst een aantal interessante mogelijkheden op voor toekomstig onderzoek.

Contents

III Fairness

List of Figures

List of Tables

Acronyms

AI	Artificial Intelligence
AUC	Area Under the Curve
CCPA	California Consumer Privacy Act
CDD	Conditional Demographic Disparity
CE	Counterfactual Explanation
CM	Classification Metric
DL	Deep Learning
DT	Decision Tree
DS	Data Science
FN	False Negative
FP	False Positive
GDPR	General Data Protection Regulation
GRASP	Greedy Randomized Adaptive Search Procedure
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
NCP	Normalized Certainty Penalty
NN	Neural Networks
PMLB	Penn Machine Learning Benchmark

PreCoF	Predictive Counterfactual Fairness
RF	Random Forest
RIPPER	Repeated Incremental Pruning to produce Error Reduction
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
XAI	Explainable Artificial Intelligence

Part I

BACKGROUND

1

Introduction

"The greatest risk with AI is not malice but competence. A super intelligent AI is by definition very good at attaining its goals, and if those goals aren't aligned with ours, we're in trouble." - Nick Bostrom, philosopher and writer.

The Cambridge Analytica scandal, which erupted in 2018, exposed the potential dangers associated with the widespread harvesting of data and its deployment in machine learning. The incident centered around the misuse of personal data obtained from millions of Facebook users without their explicit consent. Cambridge Analytica used this data to construct detailed psychological profiles, facilitating the micro-targeting of political advertisements with unprecedented precision. This unsettling incident served as a stark reminder of the profound influence technology and data can wield in shaping society's landscape. While Artificial Intelligence (AI) holds the promise of driving innovation and solving complex problems, it also harbors the potential to amplify privacy breaches, perpetuate biases and manipulate the public opinion.

Within this context, the necessity of ethical AI becomes evident. This multifaceted field encompasses essential principles such as transparency, fairness, accountability, and privacy, all of which are paramount for protecting individual rights and promoting societal well-being. With this thesis, I aim to contribute to this research field by investigating the ethical issues surrounding Explainable AI (XAI), with a specific focus on counterfactual explanations. By exploring the implications of counterfactual explanations within the ethical domains of transparency, fairness, and privacy, I aim to support the development or robust and ethically sound AI systems that align with the best interests of individuals and society at large.

1.1 OVERVIEW OF RESEARCH AND CONTRIBUTIONS

Part I lays the foundation for the remainder of this thesis. It contains an introduction, a chapter on Machine Learning and a chapter on Ethical Machine Learning. First, I introduce the topic of this thesis and give an overview of my research and contributions in Chapter 1. In Chapter 2 , I discuss the terminology that will be used throughout the thesis. Next, I discuss the classification techniques and performance metrics that will be deployed to measure the performance of the machine learning models. In Chapter 3, the field of Ethical Machine Learning is explored by utilizing the FAT framework.[1] I organize my research contributions along three pillars of the FAT framework: Transparency, Fairness and Privacy. However, it is important to acknowledge that these three areas are not mutually exclusive. Many of my research contributions will have implications across multiple areas, but I will sort them across the most appropriate dimension.

Part II of this thesis focuses on the contributions I make that are the most closely related to the field of **transparency**. In Chapter 4, I analyse the trade-off between accuracy and comprehensibility of machine learning models. Based on an analysis of 90 benchmark classification datasets, I make the following contributions:

- I find that this trade-off exists for most (69%) of the datasets, but that for the majority of cases it is rather small, while only for a few it is very large.
- I analyse how comprehensibility can be enhanced by adding yet another algorithmic step, that of surrogate modelling.
- My results show that some datasets characteristics can significantly explain this trade-off and thus the cost of comprehensibility.

This paper was published in the Journal of Big Data:

Sofie Goethals, David Martens, and Theodoros Evgeniou. The Non-linear Nature of the Cost of Comprehensibility. *Journal of Big Data*, 9(1):1–23, 2022

In Chapter 5, I analyse the manipulation risks surrounding Explainable AI, which is an implication from the disagreement problem. This problem arises when multiple explanations are possible for the same AI decision or problem. In this study, I make the following contributions:

- I provide an overview of the different strategies the explanation providers could deploy to adapt the returned explanation to their benefit.

[1] The FAT framework in AI stands for Fairness, Accountability, and Transparency, focusing on ensuring algorithms are equitable, their decisions and operations are understandable and explainable, and their processes are open and clear to users and stakeholders.

- I analyse several objectives the providers could have to engage in this behavior.

This position paper was presented at the Workshop of Explainable Knowledge Discovery and Data Mining at ECML, and included in the post-workshop proceedings:

Sofie Goethals, David Martens, and Theodoros Evgeniou. Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (forthcoming)*. Springer, 2024c

Part III encompasses the studies related to the field of **fairness**. In Chapter 6, I analyse how I can use counterfactual explanations to assess bias in machine learning models. The following contributions were made:

- I introduce *PreCoF*, or *Predictive Counterfactual Fairness*, a new technique to aggregate the counterfactual explanations.

- I show how *PreCoF* can be used to detect explicit bias (when the model is directly using the sensitive attribute) by searching for explanations that contain the sensitive attribute.[2]

- I illustrate that *PreCoF* can also be used to detect implicit bias, when the model does not use the sensitive attribute directly, but does use other correlated attributes leading to a substantial disadvantage for the protected group.[3]

This paper was published in a special issue on 'Fair and Safe Machine Learning' in the Machine Learning Journal:

Sofie Goethals, David Martens, and Toon Calders. PreCoF: Counterfactual Explanations for Fairness. *Machine Learning*, pages 1–32, 2023b

In Chapter 7, I bring transparency to the domain of bias mitigation methods, which are methods that attempt to make machine learning models less biased. The contributions of this paper are:

- I provide insights into the operational dynamics of bias mitigation methods and illustrate how some methods will significantly impact the ranking within each group, while others will not.

[2] Note that this scenario is often not legally allowed, which I also discuss in the paper.

[3] The protected group is the group that typically has been historically disadvantaged, and for which we have to ensure that the machine learning algorithm does not replicate or even worsen these biased patterns.

- I criticize the current approach to benchmark bias mitigation methods, as it compares situations that are significantly distinct and not applicable in real-world applications.

This paper is currently under review:

Sofie Goethals, Toon Calders, and David Martens. Beyond Accuracy-Fairness: Stop evaluating bias mitigation methods solely on between-group metrics. *Under review*, 2024a

In **Part IV**, I bundle the contributions made to the field of **privacy** in Machine Learning. In Chapter 8, I investigate the privacy issues that counterfactual explanations create. I make the following contributions:

- I introduce an *explanation linkage attack*, which can occur when deploying instance-based strategies to find counterfactual explanations.

- As solution, I propose an algorithm CF-K to create *k*-anonymous counterfactual explanations, and introduce *pureness* as a metric to evaluate the *validity* of these counterfactual explanations.

- I evaluate the performance of creating *k*-anonymous counterfactual explanations with CF-K by comparing it with the performance of an algorithm that makes the whole dataset *k*-anonymous. I show that only making the explanations *k*-anonymous is beneficial for the quality of the explanations.

This paper was published in ACM Transactions on Intelligent Systems and Technology:

Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–24, 2023c

In Chapter 9, I study the effect on both privacy and personalization when users hide, or *cloak* as I name it in the paper, part of their digital footprints. I offer three contributions to the existing literature:

- I assess the longer-term effectiveness of cloaking digital footprints, measuring the percentage of targeted individuals whose privacy remains protected over time. The results show that the effectiveness of cloaking fine-grained features decreases steadily and markedly over time for most inference tasks.

- I introduce a new type of cloaking strategy based on metafeatures, and show that it enhances longer-term cloaking protection (as intended).

- I examine the privacy-personalization trade-off inherent in using cloaking to protect against unwanted inferences. Specifically, I show that cloaking for one task can affect the predictive performance of other personalization tasks. Moreover, the metafeature-based strategies affect other tasks more, highlighting the trade-offs faced by users: better longer-term privacy protection indeed can reduce desired personalization performance

This study is currently under review:

Sofie Goethals, Sandra Matz, Foster Provost, Yanou Ramon, and David Martens. The Impact of Cloaking Digital Footprints on User Privacy and Personalization. *Under review*, 2024d

Finally, in **Part V**, I conclude the thesis. I summarize my main findings, list some general limitations of my research and point to some interesting avenues for future research.

2

Machine Learning

2.1 TERMINOLOGY

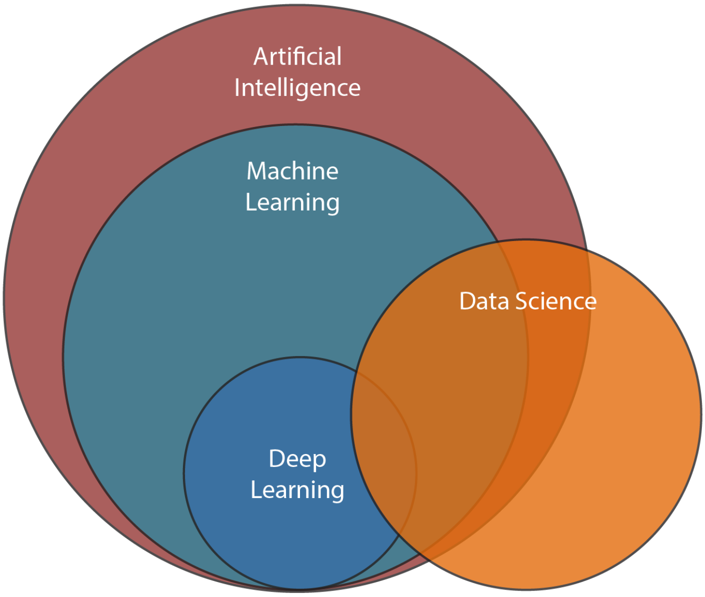


Figure 2.1: Terminology. Source: Dubovikov [2019]

In the latest years, it is easy to get confused about the difference between Artificial Intelligence (AI), Machine learning (ML), Data Science (DS) and many other terms [Dubovikov, 2019]. AI refers to the development of computer systems that would normally require human intelligence, and encompasses many subfields such as natural language processing, visual perception, logic, robotics and many others [Provost and Fawcett, 2013]. ML is one of the subfields of AI and involves the development of algorithms and models that allow computers to learn from data. Deep learning is a subset of machine learning that involves the use of neural networks with multiple layers (deep neural networks) to model and solve complex problems [LeCun et al., 2015]. Also related is Data Science, an interdisciplinary field that combines expertise from statistics, mathematics and computer science to draw knowledge from the data and provide insights [Dubovikov, 2019]. A graphical representation of this overview can be seen in Figure 2.1.

Within ML we focus on predictive modeling, which focuses on estimating an unknown value of interest, and on *supervised learning* in specific [Berry et al., 2019]. Supervised learning is a type of predictive modeling where the learning algorithm is trained using labeled data [Berry et al., 2019]. When the target variable is not known, the term *unsupervised learning* is used. Techniques such as *clustering* can be used, but this is out of the scope of this PhD thesis. Throughout the thesis, we will use the following terminology: We will model the relationship between a set of selected variables, which we call the *attributes* or *features*, and the variable we want to predict, which we call the *target variable* [Provost and Fawcett, 2013].

We are interested in predicting the values of instances we have not yet observed; we *generalize* the model to new data that was not used to build the model. When a model does not generalize well beyond the data it has been trained on, it is too tailored to the training data which we call *overfitting* [Provost and Fawcett, 2013]. To examine overfitting, we should 'hold out' some labeled data that will not be used during model building, which we name the *test set*. This set is used to measure the *generalization performance* of the model, as accuracy on this data should not diverge too much from the accuracy on the training data [Provost and Fawcett, 2013]. To tune the parameters of our models, we can split the remaining data that can be used for model building again in a *training set* and a *validation set*. The validation set is used to test the performance of different parameter settings for each of the classification models. This splitting process is depicted in Figure 2.2.

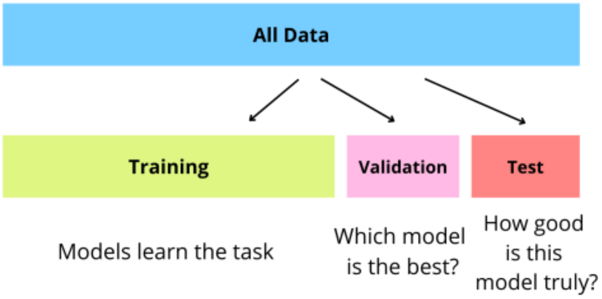


Figure 2.2: Process of splitting the dataset. Source: Chavan [2023]

2.2 CLASSIFICATION TECHNIQUES

Two of the main tasks within supervised learning are *classification* and *regression* [Provost and Fawcett, 2013]. Classification tries to estimate to which class an individual will belong. A popular example of a classification task is predicting whether a bank customer will default on its loan or not (*credit scoring*). In this case, there are only two possible classes and this kind of classification is also called *binary* classification. When dealing with more than two potential classes, the task is referred to as *multiclass* classification. Instead of a class prediction, classification algorithms could also return the probability that the individual belongs to each class. On the other hand, regression attempts to estimate a numerical target variable, such as predicting the price of a house or the future price of a stock.

A distinction that will be important throughout the paper, is that of the prediction model and the decision-making context. The decision threshold is not part of the prediction model but part of the decision logic. As argued by Scantamburlo et al. [2024], the ultimate decision of an automated system is informed by the prediction model, but in nearly all cases is also influenced by additional parameters such as quota, business rules and the costs and benefits of each decision. In the studies conducted in this thesis, we will often use the default threshold of 0.5 for the different prediction models, but this is not the procedure that will be used in standard decision-making contexts.

In this thesis, our experiments are conducted on binary classification tasks, but all our findings are applicable to multiclass classification and in most cases regression as well. Many learning algorithms exist for these kinds of tasks, and we will discuss the ones that will be used in later chapters for the experiments, which are *Decision Trees*, *Rule Induction*, *Logistic Regression*, *Random Forests*, *Support Vector Machines*, and *Neural Networks*. We will not discuss other popular supervised learning techniques such as *Naive Bayes* and *k-Nearest Neighbors*.

2.2.1 *Decision tree (DT)*

A decision tree algorithm will segment the data in the form of a 'tree'. The algorithm will recursively split the training data until all instances belong to the same class or a stopping criteria is reached [Quinlan, 1986, Song and Ying, 2015]. The tree is made up of *nodes*, internal nodes and leaf nodes, and branches that connect them. Each internal node represents a test on one of the attributes of the training data, and each branch represent the outcome on that test. Each branch ultimately ends in a terminal or leaf node, which represent a class label or probability estimation. The most common splitting criterion is *information gain*, which measures the change in *entropy* due to new information being gathered. Entropy is a purity measure that measures the disorder in a set, and in this case of predictive modeling it measures how *pure* each of the resulting decision segments is [Provost and Fawcett, 2013].

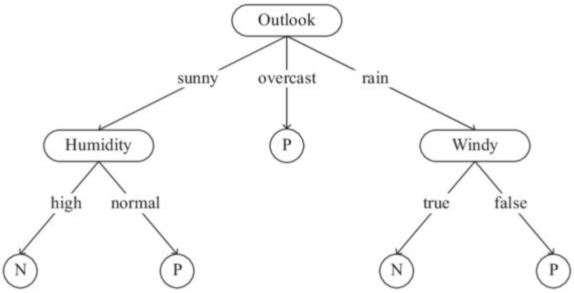


Figure 2.3: Simple decision tree to predict the weather, where P signifies that the weather is suitable and N that the weather is unsuitable. Source: Quinlan [1986].

An example of a simple decision tree to predict the weather can be seen in Figure 2.3. To avoid overfitting, a stopping criterion a can be used to stop the growing of the tree before it gets too complex, or the tree can be pruned to reduce its complexity.

2.2.2 *Rule induction*

Decision trees can also be interpreted as logical statements. If we would descend the tree from to the root node to one of the leaf nodes, a rule would be generated. Each rule contains all the attribute tests in the internal nodes along the path, connected with AND. If we would follow the left path in the decision tree depicted in Figure 2.3, we would get the rule: *If the outlook is sunny AND the humidity is high, the weather is predicted as not suitable.*

2.2.3 *Logistic regression (LR)*

The learning algorithm will estimate a linear model and tune the parameters so that the model fits the data as well as possible. The linear model can be written as follows:

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b \tag{2.1}$$

where $\mathbf{x}$ represents the feature vector, b the intercept and $\mathbf{w}$ the vector of weights. This returns a numerical value, but we are interested in the probability of class membership, which is why we will use a logistic function (as shown in Figure 2.4) to estimate this probability:

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\mathrm{T}\mathbf{x}+b)}} \tag{2.2}$$

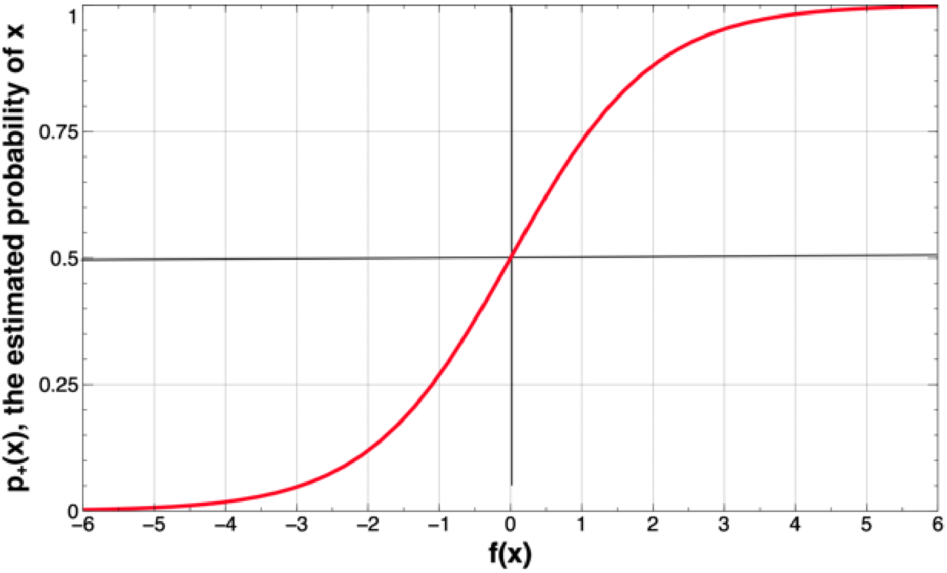


Figure 2.4: The logistic function maps the outcome of a linear function $f(x)$ to the range $[0, 1]$. Source: Provost and Fawcett [2013]

The unknown parameters **w** and b are estimated by maximizing the conditional likelihood that the predicted class label equals the true label, which is done by minimizing the loss function (the distance between the predicted and the true labels). The loss function can be defined as (when using the notation $y_i \in \{-1, 1\}$):

$$\mathcal{L} = \sum_{i=1}^{n} log(1 + e^{-y_i(\mathbf{w}^\mathrm{T} x_i + b)}) \tag{2.3}$$

To avoid overfitting the data, regularization constraints should be added to control the complexity of the model. We do this by adding a penalty for complexity to the objective function with regularization weight λ, which determines how much importance this penalty should get. This weight determines the trade-off between limiting the complexity of the model, and minimizing the prediction error. Different penalty functions can be used, and the most frequently used are ℓ_1-regularization (Lasso regression) and ℓ_2-regularization (Ridge regression). Both functions penalize the weights of the coefficients in the logistic regression model, but ℓ_1-regularization adds the sums of the absolute values of the coefficients as the penalty term, while ℓ_2-regularization adds the sum of the squared values of the coefficients as penalty term. ℓ_1-regularization will zero out many coefficients, and perform an automatic form of feature selection, while ℓ_2-regularization will lead to small, non-zero weights. The resulting objective functions for both types of regularization can be seen in 2.4 and 2.5:

ℓ_1-**regularization**:

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} log(1 + e^{-y_i(\mathbf{w}^T x_i + b)}) + \lambda \|\mathbf{w}\| \tag{2.4}$$

ℓ_2-**regularization**:

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} log(1 + e^{-y_i(\mathbf{w}^T x_i + b)}) + \lambda \|\mathbf{w}\|^2 \tag{2.5}$$

The optimal penalty function and the optimal value of λ are determined on the validation set.

2.2.4 *Random Forest (RF)*

A key issue of decision trees is that they might exhibit high variance: small variations in the training data, might result in very different decision trees. Combining decision trees into one 'super' model reduces the variance and thus improves the generalization performance [Provost and Fawcett, 2013]. Combining multiple models into one overarching model is also called *ensemble modelling*, and a *Random Forest* is a specific kind of ensemble model.

Random Forests insert randomness in two ways: each decision tree only uses a random sub-sample of the data, and each split in the decision tree only uses a random selection of features. The final RF model will average over the predictions of all individual decision trees, and end up with a more robust final prediction of the target class [Breiman, 2001a].

In Random Forest models, typical parameters that are tuned during the training process include *n estimators* (the number of trees in the forest), *max depth* (the maximum depth of the trees), *min samples split* (the minimum number of samples required to split an internal node), *min samples leaf* (the minimum number of samples required to be at a leaf node), and *max features* (the number of features to consider when looking for the best split).

2.2.5 *Support vector machine (SVM)*

Support Vector Machines attempt to find the decision boundary that maximally separates the margin (or distance) between the different classes. This process is illustrated in Figure 2.5. One can make a distinction between linear SVM's that are used when the data is linearly separable by a straight line (2D), a plane (in 3d) or a hyperplane (in higher dimensions). Non-linear SVM's are used when the data is not linearly separable, by mapping the data into a higher-dimensional feature space using the 'kernel trick'. This transformation makes it possible to find a hyperplane that separates the data points in the higher-dimensional space, even though they were not linearly separable in the original feature space.

For Support Vector Machines (SVM), common parameters to tune are *C* (the regularization parameter controlling the trade-off between achieving a low training error and a low testing error), *kernel* (the kernel type, such as 'linear', 'poly', 'rbf', 'sigmoid') and *gamma* (the kernel coefficient for 'rbf', 'poly', and 'sigmoid' kernels, defining the influence of a single training example).

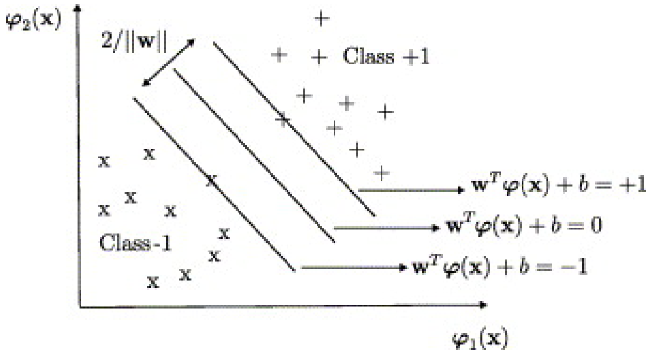

$\varphi_2(\mathbf{x})$
$2/||\mathbf{w}||$
Class +1
Class-1
$\mathbf{w}^T\varphi(\mathbf{x}) + b = +1$
$\mathbf{w}^T\varphi(\mathbf{x}) + b = 0$
$\mathbf{w}^T\varphi(\mathbf{x}) + b = -1$
$\varphi_1(\mathbf{x})$

Figure 2.5: Example of a support vector machine. Source: Martens et al. [2007]

.

2.2.6 *Neural networks (NN)*

Neural networks, often referred to as artificial neural networks or simply 'neural nets', are a class of machine learning models inspired by the structure and function of the human brain. The core building block of a neural network is the artificial neuron. The typical architecture of a neural network is depicted in Figure 2.6. It consists of an input layer, hidden layers and an output layer. Each neutron takes input from multiple sources, applies a weight to each input, and combines them through a non-linear activation function to produce an output. The outputs of the input layer are used as input to the next hidden layer. By adjusting the weights of these connections during the training process, the neural network learns to recognize patterns and make accurate predictions.

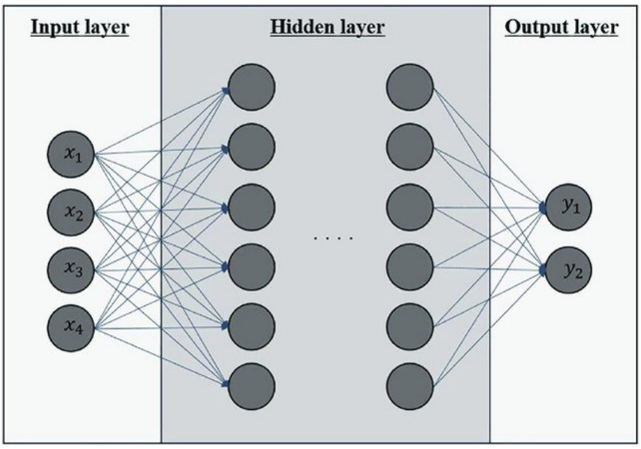


Figure 2.6: Example of an artificial neural network. Source: Pyo et al. [2017].

The training of a neural network typically involves two main phases: forward propagation and backpropagation. During forward propagation, the input data flows through the network, and predictions are made. The output is then compared to the ground truth to calculate the prediction error, also known as the loss or cost function. Backpropagation is the process through which the neural network updates its weights to minimize the prediction error. This is achieved by calculating the gradient of the loss function with respect to the network's parameters and using optimization algorithms like stochastic gradient descent (SGD) to iteratively adjust the weights.

In Neural Networks, typical parameters tuned include *learning rate* (the step size for updating weights), the *number of layers* (total hidden layers in the network), *number of neurons per layer* (neurons in each hidden layer), *activation function* (function introducing non-linearity, such as 'relu', 'sigmoid', 'tanh'), *batch size* (samples processed before updating the model), and *epochs* (number of times the entire training dataset is passed through the network).

Deep learning (DL) is a subset of machine learning that focuses on deep neural networks with multiple hidden layers. These models have achieved exceptional performance in areas such as computer vision, natural language processing and speech recognition.

2.3 PERFORMANCE METRICS

To evaluate the performance of machine learning models, it is important to first establish the difference between scores (probabilities) and classifications (decisions). A classification model assigns a prediction *score* to each instance, and this score is transformed into a classification by using the threshold of the model [Provost and Fawcett, 2013]. As mentioned in Section 2.1, we use the hold-out or test data (not used during model building) to evaluate the performance of a classifier. The target labels of the test data are known, which allows us to compare the true class membership of the instances in the test data with the predicted output of the classification model.[1]

ACCURACY Accuracy is a popular metric as it is very intuitive to understand [Provost and Fawcett, 2013].

$$Accuracy = \frac{Number\ of\ correct\ classifications\ made}{Total\ number\ of\ classifications\ made} \tag{2.6}$$

Unfortunately, evaluating on accuracy has some well-known problems [Provost et al., 1998]: when the class distribution becomes very skewed, always choosing the most prevalent class can return a very high accuracy (equal to the imbalance in the dataset), while the model is not actually predicting anything. Furthermore, it makes no distinction between the different kinds of errors, but they can have different costs.

CONFUSION MATRIX A confusion matrix can help in showing how each class is being confused for the other. A confusion matrix for a problem of n classes consists out of n rows, where each row denotes the predicted class, and n columns, where each column denotes the actual class. An example of a confusion matrix for a binary classification problem can be seen in Table 2.1

Table 2.1: Confusion matrix

	Actual positive	**Actual negative**
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

[1] The predicted output can be the scores or the decisions, depending on the evaluation metric that was used.

Based on the values in this matrix, various evaluation metrics can be described:

- **Precision** = $TP/(TP + FP)$
- **Recall** (or Sensitivity or TPR) = $TP/(TP + FN)$
- **Specificity** = $TN/(FP + TN)$
- **F1-score** (harmonic mean between precision and recall) = $2 \times \frac{precision \times recall}{precision + recall}$

We can also include cost-sensitive metrics, and explicitly take into account the cost or weight of each type of prediction:

- **Weighted accuracy** = $\frac{w_{TP} \cdot TP + w_{TN} \cdot TN}{w_{TP} \cdot TP + w_{TN} \cdot TN + w_{FP} \cdot FP + w_{FN} \cdot FN}$
- **Total Cost** = $C(0, 1) \cdot FN + C(1, 0) \cdot FP$

AUC A useful approach to show the entire space of performance possibilities is the Receiver Operating Characteristics curve. This is a two-dimensional plot with the false positive rate of the classifier on the x axis against true positive rate of the classifier on the y-axis. The output of a discrete classifier will result in a point in this space. However, we can also look at the output of the scoring function of our classifier, ranking the scores in descending order and range over all possible thresholds from high to low.

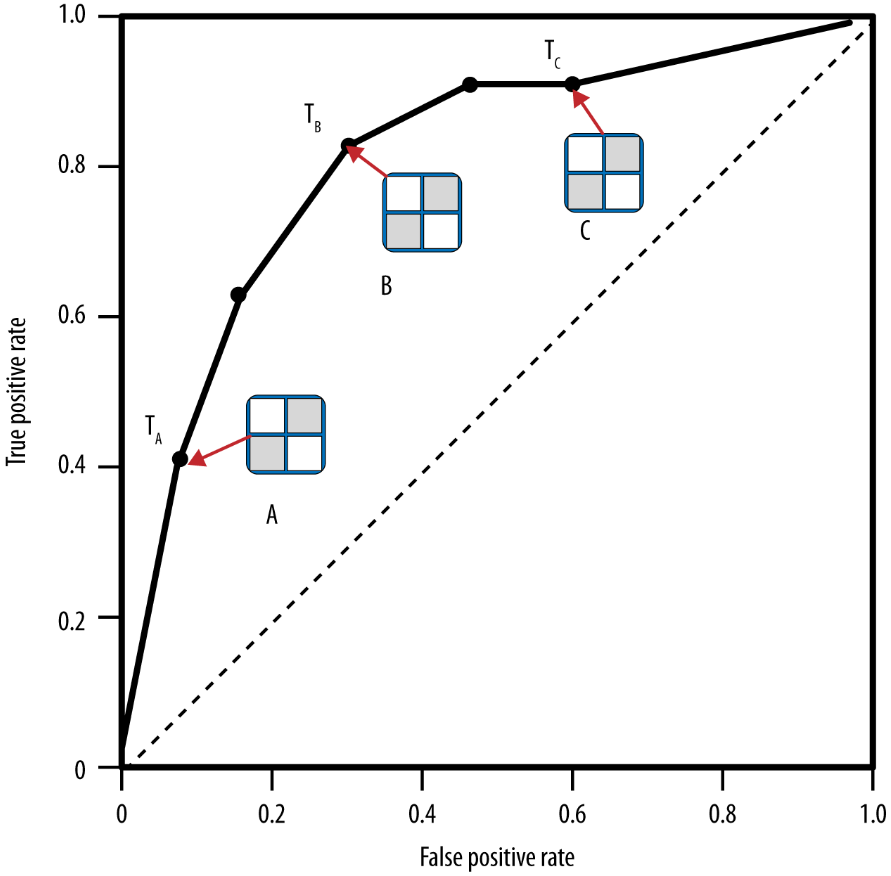


Figure 2.7: An illustration of how each point in this space corresponds to a specific threshold and confusion matrix. Source: Provost and Fawcett [2013]

An illustration of how each this curve is constructed can be seen in Figure 2.7. Each point on the curve reflects a specific threshold, and therefore a different confusion matrix and resulting TPR and FPR. *TPR*, also known as sensitivity or recall, is the proportion of actual positive instances that are correctly identified by a classification model. *FPR* is the proportion of instances that are actually negative but are incorrectly classified as positive by a model. The dashed diagonal line represents a random classifier, as here the true positive rate will be equal to the false positive rate for each threshold. The Area Under the Curve (AUC) is a summary statistic for the Receiver Operating Characteristics curve that ranges from 0 to 1. A perfect model achieves an AUC of 100%, while a random model has an AUC of 50%. AUC allows for an objective comparison across classifiers, as it is unaffected by the choice of threshold or the frequency of classes [Hand, 2009].

3

Ethical Machine Learning

Artificial Intelligence (AI) and Machine Learning (ML) have witnessed an unprecedented surge in popularity in recent years, and consequently, they are revolutionizing numerous aspects of our lives, from optimizing daily routines to driving decisions in high-stakes scenarios. Machine learning models are now trusted to guide judgements in domains with profound social consequences, such as medical diagnosis, credit scoring, fraud detection, and criminal justice systems. The growing reliance on these systems in high-stakes decisions emphasizes the need to ensure that the effect on society is positive and aligned with our ethical objectives. In the past decades, the focus of predictive modeling was mostly on ensuring that accurate predictions were made. Yet, many cautionary tales exist, that show that a high test accuracy is no longer sufficient. In HR Analytics, a well-known case is that of an automated Amazon recruitment system. Based on historical data, a predictive model was built to predict whether a candidate was suitable for an engineering position by analyzing their resume. The model seemingly had a high accuracy. Yet, the model had learned a bias against women: if the name of some all-female universities were included in the resume, or the word 'women's' (as in 'president of the women's soccer team or chess club'), the candidate would automatically be down weighted. Upon revelation, Amazon pulled the system [Dastin, 2022]. Apple faced a media storm on Twitter about the potential sexist credit scoring of Apple Card, and its inability to explain the used machine learning [Agrawal, November 12, 2019]. Lastly, Google was criticized for using an image classifier that automatically tagged images of black peoples as gorillas [Barr, 2015]. These cases demonstrate that major technology players (such as Amazon, Google and Apple) with massive capabilities, are struggling with this challenge with large reputational, ethical and legal implications.

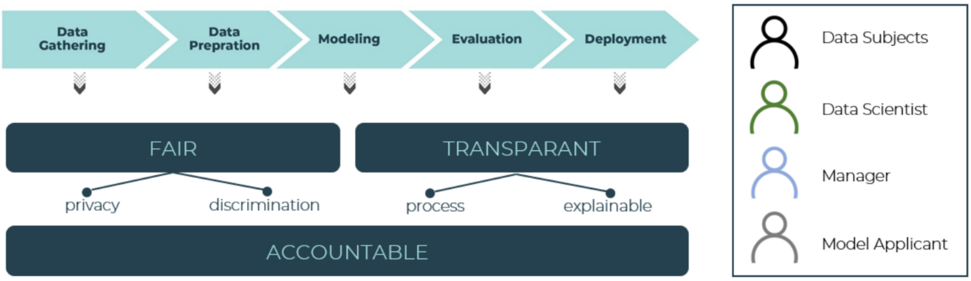


Figure 3.1: FAT Flow framework, using three dimensions: 1) role, (2) modeling stage, and (3) evaluation criterion. Source: Martens [2020]

3.1 FAT FRAMEWORK

We will use the FAT Flow framework to analyse the ethical aspects of machine learning. This framework looks at three dimensions: (1) the role of the humans involved in the project; (2) the stage of the data science project; and (3) the FAT evaluation criteria: *Fair*, *Accountable* and *Transparent*.

In this PhD thesis, we mostly focus on the Evaluation stage, and the role of Data Scientists to make the whole process more ethical. The criteria of *Fair* and *Transparent* encompass ethical concepts such as privacy, discrimination and explainability [Martens, 2022]. The criterion of *Accountable* is about the implementation of these concepts into effective, demonstrable measures and will be out the scope of my research. In my research I will focus on the criteria of *Fair* and *Transparent*, and how they relate to Explainable AI. The first criterion *Fair* encompasses two guidelines in the context of machine learning:

- **Fair(a): Not discriminating against sensitive groups**
- **Fair(b): Without cheating or trying to achieve unjust advantage with respect to privacy**

The goal of machine learning is to discriminate between groups, and the first guideline states that this should be done in a way that treats people equally without prioritizing or discriminating certain society groups. Beside the discrimination aspect, this criterion also takes into account the privacy aspect. The fair use of personal data entails that the privacy of the data subjects should be respected [Martens, 2020]. This criterion leads to the pillars of fairness (a) and privacy (b).

The *Transparent* criterion also contains two components:

- **Transparent(a): Clarity in the process**

- **Transparent(b): Ability to explain decisions made by data science models**

This means that all the stages of the model itself should be clear, but also that the information provided should be sufficiently comprehensive for the data subject to understand the reasons for the decision made by the model (as required by GDPR) [Martens, 2022].

3.2 TRANSPARENCY

It is important to first shed some light on the used terminology regarding transparency. We will use the terms *transparency*, *comprehensibility* and *interpretability* interchangeably (as is done by literature) and define this as *the degree to which a human can understand the cause of a decision* [Miller, 2019].

3.2.1 *Why do we need it?*

Within the field of Artificial Intelligence, providing insights into the decision-making process is crucial for various reasons. Following Ramon [2022], we can group the arguments for transparency in machine learning models into four categories: trust and acceptance, improved insights, model improvement and protection of data subjects.

TRUST & ACCEPTANCE First, it establishes trust and compliance with stakeholders, as they can understand and validate the reasoning behind the model's output. This is important as users who do not understand the inner workings of a machine learning model, will be skeptical and reluctant to use it [Kayande et al., 2009].

PROTECTION OF DATA SUBJECTS Machine learning models can pick up biases from the training data. Interpretability methods can be a useful tool for detecting bias, as we will show in Chapter 6. Lately, this has also been legally required. For example, the General Data Protection Regulation (GDPR) notes the 'right to explanation' for individuals who are affected by AI decisions [Wachter et al., 2017b].

MODEL DEBUGGING AND IMPROVEMENT Machine learning models can only be properly audited when they can be interpreted. Understanding the model can be an important safety measure during model testing, as it can not only pick up mistakes but also lead to to model improvements. For example, in Vermeire et al. [2022a], explanations were computed to explain why certain lighthouse images were wrongly predicted. The explanations revealed that the model was not focusing on the shape of the lighthouse, but on the presence of the clouds [Vermeire et al., 2022a].

DOMAIN INSIGHTS It can lead to improved insights about the domain, allowing practitioners and users to gain a deeper understanding of the problem space and uncover valuable knowledge. This also allows experts to better interact with the output of the model, and inform investigators what to (first) focus their attention on [Martens, 2022]. For example, when we are predicting mental states such as depression [Müller et al., 2020], insights in the machine learning model can help us to design targeted interventions.

3.2.2 *Taxonomy*

To reach these goals, various methods to achieve comprehensibility in AI models have been proposed. In general, there are two main approaches commonly used: inherently transparent models and post-hoc explanations. Inherently transparent models, such as small decision trees or sparse linear models, are comprehensible by nature due to their simple structure, without the need for additional explanations [Molnar, 2020]. However, in many real-world scenarios, data is becoming increasingly complex and black-box models are used due to their superior predictive performance.[1] These models lack inherent interpretability, and post-hoc explanations are used to provide insights into their decision-making process. Post-hoc interpretability refers to the application of interpretation methods after the building of the model [Molnar, 2020]. This field of research is commonly known as Explainable Artificial Intelligence (XAI).

Another distinction can be made between model-specific interpretation methods and model-agnostic interpretation methods. The former can only be used for one specific model, such as for example neural networks. On the other hand, model-agnostic tools can be used on any machine learning model. By definition, these methods do not need access to model internals such as weights or structural information [Molnar, 2020].

3.2.3 *Explainable Artificial Intelligence (XAI)*

Explainable AI (XAI) refers to the capability of an AI system to provide understandable explanations for its decisions, actions or predictions. An important remark about this field is that there currently is not a clear definition of what an explanation

[1] We discuss the difference in performance between interpretable models and black-box models in Chapter 4

actually is. A lot of different methods exist that are all considered as explanations but display very different things. I will study these methods together as they are all considered as explanations by the research field. Generally, a distinction can be made between global and local explanations.

global explanation methods Global explanations aim to provide an understanding of the model's logic as a whole, allowing users to follow the reasoning that leads to every possible outcome. Techniques such as surrogate modeling [Martens et al., 2007], feature importance rankings and Partial Dependence plots [Friedman, 2001] fall under this category. A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model. Any interpretable model (linear model, decision tree, rules, ...) can be used and the closeness between the surrogate model and the black box model is measured by the *fidelity* of the surrogate model.

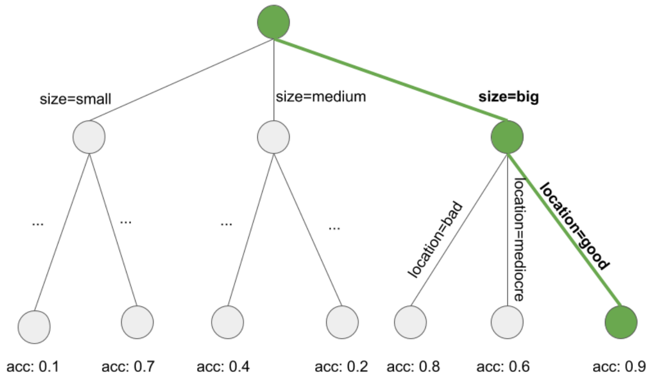


Figure 3.2: Surrogate decision tree to predict the value of a house. Source: Molnar [2020]

In Figure 3.2, we see how a simple decision tree can be used to mimic the behavior of a black box, and still result in explainable rules.

Feature importance rankings are also a popular global explanation method. Several implementations to calculate this exist, both model-specific and model-agnostic. For example, random forest have a model-specific function to determine the feature importance based on the Gini Impurity. Permutation feature importance rankings are an alternative way to calculate this, which involves changing the value of a feature and assessing the changes in the algorithm's performance. An example of a feature importance ranking can be seen in Figure 3.3

local explanation methods On the other hand, local post-hoc explanations focus on explaining the logic behind a specific prediction or decision made by the

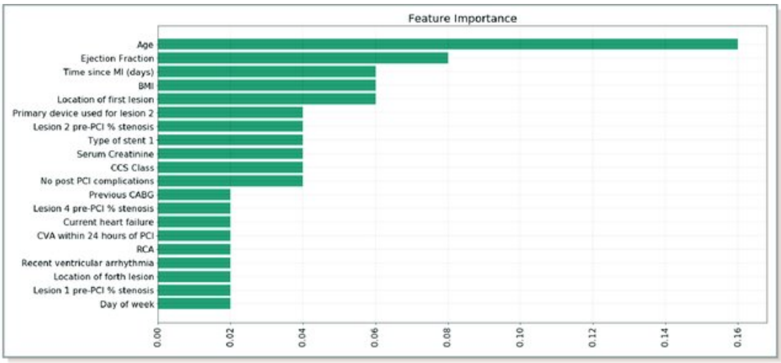


Figure 3.3: Feature importance ranking of a machine learning model that predicts the in-hospital mortality. Source: Al'Aref et al. [2019]

model. Methods like SHapley Additive exPlanations (SHAP) [Lundberg and Lee, 2017] and Local Interpretable Model-agnostic Explanations (LIME) [Ribeiro et al., 2016] are examples of post-hoc explanations that measure the impact of each feature for a given prediction score (feature importance methods). SHAP is based on Shapley values, a game-theoretic concept to estimate the contribution of each feature to a prediction [Shapley et al., 1953]. However, computing the theoretical Shapley values is very computationally expensive, and this is why SHAP uses an efficient approximation of this concept [Lundberg and Lee, 2017, Molnar, 2020]. LIME trains a local surrogate model in the neighborhood of the instance to be explained, and uses the weights of this surrogate model to explain the instance's decision [Ribeiro et al., 2016]. In Figure 3.4, we show an example of a SHAP waterfall plot for the Boston Housing dataset. [2]

However, both LIME and SHAP have some common drawbacks as they do not consider feature dependence [Aas et al., 2021], and do not return sparse explanations as they assign a value for ever input feature. Another local technique, known as counterfactual explanations, describes a combination of feature changes required to alter the predicted class [Martens and Provost, 2014, Wachter et al., 2017b, Guidotti, 2022]. An example of a counterfactual explanation in the context of credit scoring can be seen in Figure 3.5.

Factual instance						**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Lisa	***21***	*F*	***Brussels***	***$50K***	*Single*	*Reject*

Counterfactual explanation=
If you would be **three years older**, lived in **Antwerp** and your income would be **$10K** higher, you would have received a positive credit decision

[2] https://github.com/shap/shap

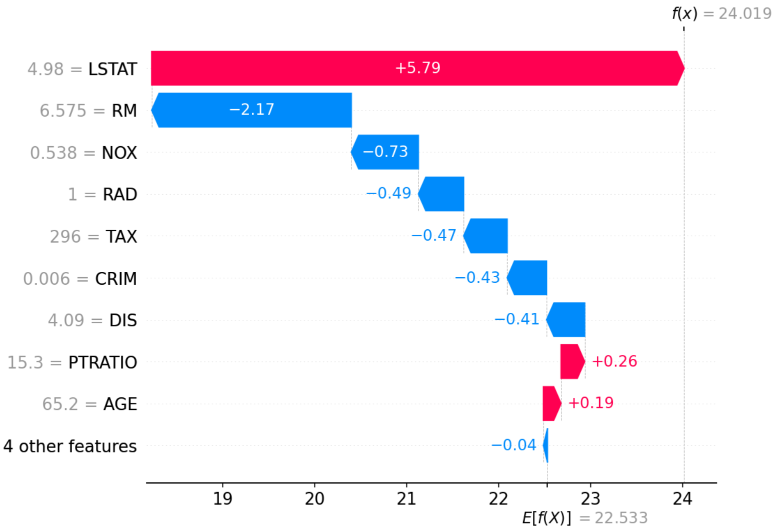


Figure 3.4: This SHAP plot shows how each feature contributes to the prediction score. Source: Lundberg et al. [2020]

Counterfactual instance						**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Fiona	***24***	*F*	***Antwerp***	***$60K***	*Single*	*Accept*

Figure 3.5: Example of a counterfactual explanation. Source: Goethals et al. [2023c]

The largest difference between SHAP and LIME on one side, and counterfactual explanations on the other is that the former will focus on features impacting the prediction score, while the latter focuses on the decision of the classification model [Fernandez et al., 2020].

It is important to note that no explanation method is perfect and that the preferred technique will depend on the task and end user at hand. However, this plethora of explanation methods and implementations will lead to a new problem, commonly known as *the disagreement problem*. We discuss some possible implications in Chapter 5.

3.3 FAIRNESS

Fairness has become one of the most popular topics in machine learning in recent years. The research community has been putting more and more emphasis on this field, which also led to several new conferences and workshops on fairness such as ACM FAccT and the European Workshop on Algorithmic Fairness. The sudden increase in papers around this topic can also be seen in Figure 3.6.

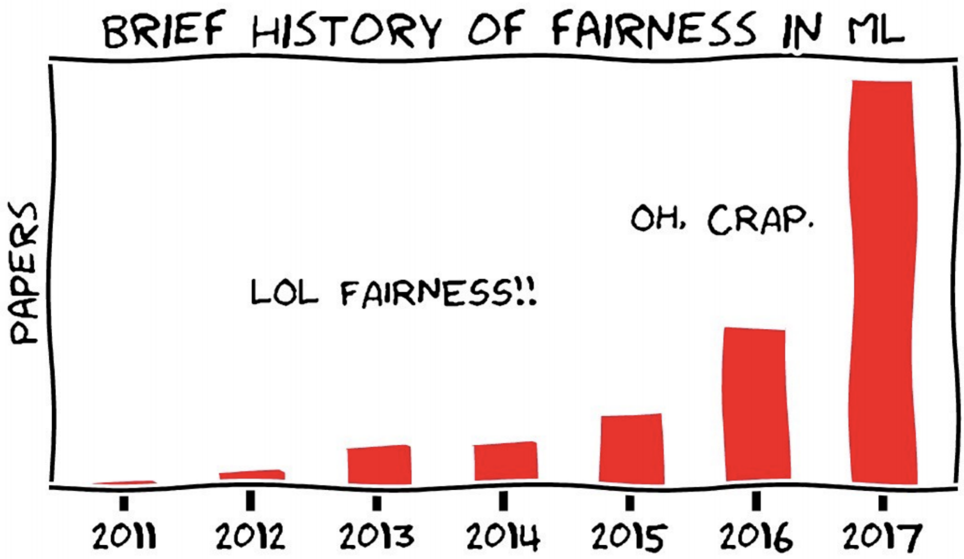


Figure 3.6: History of Fairness in machine learning. Source: Hardt and Barocas [2017]

Why should we care about fairness in machine learning? In today's world, machine learning has become increasingly pervasive, touching more and more domains of our lives. It is used by employers for job applicant screening, by banks for mortgage approval, by courts for predicting recidivism, and by online recommender systems like Netflix and Amazon to personalize content recommendations. So as these systems are integrated in every part of our personal life, it becomes crucial to ensure that their decision-making processes are fair and just.

Although machine learning may appear objective; in reality they will only be as good as the data they are trained on, giving rise to the often cited motto "garbage in, garbage out" [Johnson, 2021]. There are already various examples of cases in real life where this happened: A well-known use case is that of an automated Amazon recruitment system, that we discussed at the beginning of this Chapter, where a machine learning model learned a bias against women in a resume screening task [Dastin, 2022]. Another well-known case is the Gender Shades study by Buolamwini and Gebru [2018]. The authors discovered that commercial facial recognition systems had significant bias, performing poorly on darker skinned women. Finally, in the realm of medicine, AI systems have used health costs as a proxy for health needs and falsely concluded that patients of color are healthier than equally sick white patients, as less money was spent on them in the past [Obermeyer et al., 2019]. Consequently, these algorithms gave higher priority to white patients when treating life-threatening conditions [Norori et al., 2021]. A wealth of such examples can be found, and it is likely that many cases of unnoticed bias persist. These cases underscore the crucial need for identifying and mitigating biases in machine learning models.

A growing awareness of these societal implications, has spurred the development of new regulations that specifically target fairness in these technologies.These regulations seek to address issues of algorithmic bias, discrimination, and unequal treatment by requiring organizations to adopt measures that promote fairness. A recent example is the European Union's AI Act, which establishes a comprehensive framework for AI development and employment, emphasizing fairness and non-discrimination as core principles.

Before we go further, it is important to define some of the key terminology that is often used in the fairness literature. A sensitive attribute refers to a characteristic or feature of an individual that is considered sensitive, often with respect to potential discrimination. This can include attributes such as race, gender, age, religion, sexual orientation, or any other factor that could be the basis for unfair treatment. Consequently, a protected group typically refers to the demographic group that is at risk of being unfairly treated or discriminated against based on their sensitive attribute, while the privileged group is the demographic category that is typically not subject to unfair treatment based on that sensitive attribute. Fairness metrics are quantitative measures used to assess the fairness of AI models, while fairness or bias mitigation strategies are techniques and approaches used to modify AI models to reduce discrimination.

3.3.1 *What is algorithmic bias?*

Bias can be a confusing term, and in Machine Learning it is used to point to any systematic error made during model development. In our daily life, bias can be a prejudice toward or against one person or group based on their characteristics. What are common sources of bias that lead to unfairness in algorithmic systems? Algorithmic biases can emerge in various ways. We discuss some of the most well-known types:

historical bias - Historical bias refers to situations when the target variable is dependent on the sensitive attribute, but in principle no relationship should exist [Baumann et al., 2023]. This is the consequence of certain groups being discriminated against in the past. For example, the difference in average income between genders is generally perceived as reflecting long-lasting social barriers and does not reflect any intrinsic differences among genders [Baumann et al., 2023]. Another example is when past arrest records are used to predict future crime, because minority neighborhoods often experience levels of policing, leading to more arrests [Kleinberg et al., 2018].

measurement bias - Measurement bias occurs when a proxy of some variable is used, because the real value cannot be properly measured, and that proxy is dependent on a sensitive characteristic [Baumann et al., 2023]. For example, IQ is often used to measure intelligence, but it has been shown that it could systematically favor or disfavor specific groups.

representation bias - Representation bias occurs when the training data used to build a model fails to adequately represent all relevant subgroups in the population. In other words, certain groups or categories within the dataset may be underrepresented or over represented, leading to a lack of generalizability to those groups. This bias can result from various factors, such as sampling methods, data collection processes, or historical inequalities. For example, if a facial recognition system is trained primarily on images of individuals from a particular demographic (e.g., a specific ethnicity or gender), it may not perform well for individuals from underrepresented groups.

SELECTION BIAS - Selection bias arises when the data used for training or evaluation is not randomly sampled or is selectively chosen, leading to a skewed representation of the overall population. This bias can result from non-random sampling methods, such as convenience sampling, or from the intentional inclusion or exclusion of certain instances. For instance, in a medical study, if only patients with mild symptoms are included, the model may not generalize well to those with severe symptoms.

This list is in no way exhaustive, as there is an abundance of ways in which bias can seep into the data.

3.3.2 *Fairness definitions*

Simply put, fairness is the absence of any prejudice towards an individual or group. How can we define this in ML systems?

In this thesis, we focus on fairness in binary classification problems with a single sensitive attribute. However, many of these concepts extend to other settings (regression, multiclass classification, multiple sensitive attributes, etc.).

In the computer science community, a plethora of fairness metrics have been proposed [Corbett-Davies et al., 2023]. We will focus on some of the most well-known: Fairness through unawareness, group fairness metrics, individual fairness, and counterfactual fairness.

FAIRNESS THROUGH UNAWARENESS A common starting point for designing a fair algorithm is simply to exclude the sensitive attributes from the model. The limitations of fairness through unawareness have been commonly addressed, with the most fundamental limitation being ‘the proxy problem’. The proxy problem states that the omission of sensitive attributes can lead to the emergence of proxy variables that indirectly encode the information contained in the sensitive attribute and hence still introduce bias into the model’s decision-making process. A classic example of the proxy problem, is the use of zip codes in the United States as a proxy for racial information, as these two attributes tend to be heavily correlated. Furthermore, removing the sensitive attribute makes it more difficult to act on this discrimination. However, some authors still note that while blinding can lead to sub-optimal decisions, the legal, political, and social benefits of, for example, race-blind and gender-blind algorithms may outweigh their costs in certain scenarios [Coots et al., 2023, Corbett-Davies et al., 2023].

group fairness metrics One of the most popular approaches are group fairness metrics, which quantify the fairness of a machine learning model across different demographic or sensitive groups, aiming to identify disparities in the outcome between these groups. The most well-known metrics are presented in Table 3.1. Assume the following terminology:

- We have a sensitive attribute S with values $\{s, ns\}$, where the sensitive value is s and the non-sensitive value is ns.
- We have a target variable Y with outcomes {0,1}
- We have a predicted target variable $\hat{Y}$ with outcomes {0,1}

One of the simplest and most commonly used definitions is *demographic parity* (or *statistical parity*), which states that the positive classification rate must be the same regardless of the protected attributes. In our example of college admissions, this means that a model must admit equal percentages of white and black applicants (if race is the sensitive attribute) or of women and men. Equal opportunity states that the proportion of true positives must be equal, while equalized odds examines whether both the proportion of true positives and trues negatives is approximately equal across groups. Lastly, predictive parity examines the ratio of true positives to predicted positives. Besides these, many other fairness metrics exist, and the issue is that most of them are mutually incompatible [Kleinberg et al., 2016]. Deciding upon a group fairness metric to enforce, thus means already imposing a certain world view. [3]

Fairness Metric	Formula	Description
Demographic parity	$P(\hat{Y}=1 \mid S=s) \approx P(\hat{Y}=1 \mid S=ns)$	Demographic parity examines whether the probability of a positive outcome ($Y=1$) is approximately equal across groups.
Equal Opportunity	$P(\hat{Y}=1 \mid S=s, Y=1) \approx P(\hat{Y}=1 \mid S=ns, Y=1)$	Equal opportunity assesses whether the true positive rate is approximately equal for different groups.
Equalized Odds	$P(\hat{Y}=1 \mid S=s, Y=1) \approx P(\hat{Y}=1 \mid S=ns, Y=1)$, $P(\hat{Y}=1 \mid S=s, Y=0) \approx P(\hat{Y}=1 \mid S=ns, Y=0)$	Equalized odds examines whether the true positive rate and the true negative rate are approximately equal for different groups.
Predictive parity	$P(Y=1 \mid S=s, \hat{Y}) \approx P(Y=1 \mid S=ns, \hat{Y})$	Predictive parity examines whether the positive predictive value (the ratio of true positives to predicted positives) is approximately equal for different groups

Table 3.1: Overview of some of the most used group fairness metrics (but many more exist).

individual fairness Individual fairness, proposed by Dwork et al. [2012], emphasizes the notion that similar individuals should be treated similarly. This perspective shifts the focus from group-based fairness, which concerns itself with ensuring that different demographic groups are treated equally, to an individual-centric approach. They argue that fairness is inherently about how we treat each person rather than how we treat groups of people. In this context, the idea is that if two individuals are alike in relevant aspects, they should receive comparable outcomes from an algorithmic decision-making process. However, it hard to determine an appropriate metric to measure the similarity between 2 inputs, and the choice of similarity metric has a lot of influence on the outcome.

[3] For an overview of true positives etc, we refer to the confusion matrix in Table 2.1

COUNTERFACTUAL FAIRNESS Counterfactual fairness, proposed in Kusner et al. [2017], takes a more causal approach. In line with Pearl's causal model [Pearl et al., 2000], they deem the prediction of a model for an individual as fair if it is the same in the real world as it would be if the individual would belong to a different demographic group [Kusner et al., 2017, Wu et al., 2019]. To measure this, they make explicit assumptions about the causal relationships in the data. One way for a predictor to be counterfactually fair is if it is a function of only non-descendants of the sensitive attribute, so the result of this metric will heavily depend on the chosen causal model. This metric is further discussed in Chapter 6.

3.3.3 *Overview of fairness mitigation strategies*

There are many algorithms that claim to improve fairness. We can divide most of them into three categories: preprocessing, inprocessing and post-processing.

PREPROCESSING These methods will change the representation of the data before the machine learning model is learned. The idea behind *Learning Fair Representations* introduced by Zemel et al. [2013] is that a new representation Z is learned that removes the information correlated with the sensitive attribute, but preserves the other information about X as much as possible. This intermediate representation can be used for other downstream tasks such as regression and classification, and produce results that satisfy demographic parity and individual fairness. Another preprocessing technique is *Reweighing*, that aims to mitigate bias in the training data by adjusting the weights of instances belonging to different groups [Kamiran and Calders, 2012]. The idea is to assign higher weights to instances from underrepresented groups and lower weights to instances from overrepresented groups. *Sampling* uses the same reasoning, but involves creating a balanced dataset by either oversampling instances from underrepresented groups or undersampling instances from overrepresented groups.

INPROCESSING In-processing methods improve fairness during the training process by incorporating various strategies designed to ensure that the model learns to treat individuals or groups more equitably. These methods often involve adding additional constraints or modifying the objective function to balance accuracy with fairness considerations. A popular technique is adversarial debiasing, which combines a classifier that predicts the class label with an adversary that predicts the sensitive attribute [Zhang et al., 2018]. The goal is to maximize the classifier's performance while minimizing that of the adversary.

In-processing methods also include techniques such as constrained optimization and dual learning frameworks, where dual variables or Lagrange multipliers are used to enforce fairness constraints dynamically during training [Komiyama et al., 2018]. These methods allow for a flexible trade-off between fairness and accuracy, as the model can adjust the importance of fairness constraints based on the training data and the specific fairness goals.

POST-PROCESSING These methods attempt to modify the posteriors in a way that satisfies fairness constraints. The most straightforward option is to modify the classification thresholds per sensitive group to enforce a fairness metric. Hardt et al. [2016] demonstrates this for equalized odds. Another option, introduced by Kamiran et al. [2012], is to use reject option classification. This strategy enforces one of the fairness metric by imposing a confidence threshold and flipping all predictions that fall below it.

3.4 PRIVACY

In an era of increasing data availability and advanced machine learning technologies, the topic of privacy has garnered significant attention and concern. Privacy is a fundamental human right, and its preservation becomes a critical consideration when applying machine learning techniques to personal or sensitive data. Personal data boils down to data that can be linked to an individual [Martens, 2022]. More formally, personal data comprises "any information related to an identified or identifiable natural person" (Art 4(1) European Parliament and Council, 2016). Consequently, data that can be used to identify a person, either directly or indirectly (such as by combining an individual data point with another piece of data enabling identification), is classified as personal data [Van Dijck and Poell, 2013].

The privacy landscape is continually evolving with the introduction of regulations like the General Data Protection Regulation (GDPR) in Europe and the California Consumer Privacy Act (CCPA) in the United States [Zaeem and Barber, 2020]. These regulations emphasize individual data rights and place responsibilities on organizations that collect and process data. The European Union's General Data Protection Regulation (GDPR) lays down a legal framework for data protection and privacy in organizations and research activities involving personal data. Article 5 of the GDPR outlines principles for processing personal data (European Parliament and Council, 2016). Significantly, data collection should be limited to what is necessary, emphasizing the concept of "data minimization" to ensure that organizations do not retain more data than required for their intended purposes. Moreover, they are obligated to be transparent about their reasons for collecting personal data and to align their data processing with the expectations of the individuals concerned ("purpose limitation"). Even when they collect and use personal data fairly and lawfully, organizations are prohibited from retaining it longer than necessary ("storage limitation").

3.4.1 *The privacy paradox*

The *privacy paradox* constitutes the apparent contradiction between individuals' concerns about their online privacy and their actual behavior when interacting with digital platforms and services [Barth and De Jong, 2017]. Surveys and studies consistently show that people value and prioritize the protection of their personal data. However, individuals still readily share personal information on social media, use online services that collect extensive data, and engage with apps and platforms that request broad permissions.

3.4.2 *Privacy risks in machine learning*

We list some challenges with regard to data leakage that the field of machine learning is currently facing. With data leakage, we mean the risk of unintentional disclosure of private information during the training or use of machine learning models. Just as removing sensitive information from a machine learning model is insufficient to guarantee fairness, the mere removal of identifiers from a dataset falls short of ensuring privacy. Numerous instances serve as cautionary tales, illustrating how ostensibly anonymous data can still be leveraged to identify individuals and inadvertently result in data breaches.

One well-known case is Netflix's release of movie ratings data in 2016 for a data mining competition aimed at enhancing their movie recommendation algorithm. Although published anonymously, some Netflix users could be identified by comparing the data with another public data set from IMDB. Netflix was forced to cancel the competition following allegations of divulging their customers' movie preferences [Amatriain and Basilico, 2020, Martens, 2022].

Why do we even care about the ability to identify a person in the Netflix dataset? Your movie-watching history might reveal movies you do not want to share publicly (as opposed to the public ratings you share on IMDB), and can lead to the prediction of your political, sexual, and religious preferences [Narayanan and Shmatikov, 2008, Kosinski et al., 2013].

We list some of the most discussed privacy attacks in Machine Learning below (this list is not exhaustive). We will discuss these attacks in more detail in Chatper 8.

MODEL EXTRACTION ATTACK Techniques that attempt to reverse-engineer or replicate the underlying machine learning model by making queries to the model [Rigaki and Garcia, 2023]. This can lead to theft of intellectual property and the model's architecture and parameters.

MEMBERSHIP INFERENCE ATTACKS Methods to determine whether the data of a specific individual was part of a training dataset, revealing their participation in sensitive or confidential activities [Shokri et al., 2017].

ATTRIBUTE INFERENCE ATTACK This kind of attack aims to predict private attributes of an individual based on the output of the machine learning model [Rigaki and Garcia, 2023].

3.4.3 *Privacy-preserving machine learning*

Addressing these challenges while still benefiting from the power of machine learning requires the development of privacy-preserving techniques. These methods aim to strike a balance between predictive accuracy and safeguarding individual privacy. We discuss some common privacy-preserving techniques:

k-anonymity A privacy-preserving technique that makes it difficult to distinguish one individual's data from at least *k*-1 other individuals, by suppressing or generalizing the quasi-identifiers [Sweeney, 2002b]. It is defined by Sweeney [2002b] as: *'A property of a dataset where for each combination of quasi-identifiers in the dataset, there are at least $k - 1$ other instances with the same value combination'*. However, this property does not offer protection against privacy attacks such as the homogeneity attack and the linkage attack. Therefore, more strict properties such as *l*-diversity [Machanavajjhala et al., 2007] and *t*-closeness [Li et al., 2006] were suggested.

differential privacy A mathematical framework that adds noise to the data to make it more challenging to identify individual records while still allowing for useful analysis [Dwork, 2006]. The definition of differential privacy by Dwork [2006] notes: *'A property of an algorithm where the outcome or result will remain essentially the same whether you participate or not in the dataset'*.

federated learning Federated learning is a distributed machine learning approach that aims to train a centralized model while the training data remains distributed over the local devices [Martens, 2022]. Models are trained locally on individual devices or servers, and only model updates, not raw data, are shared [McMahan et al., 2017]. This approach minimizes data exposure.

homomorphic encryption Homomorphic encryption offers a solution to protect privacy while also allowing for cloud computing [Martens, 2022]. This technique allows data to remain encrypted while computations are performed on it, preserving privacy during processing [Yi et al., 2014].

3.5 conclusion

The ethical principles of fairness, transparency, and privacy within the realm of machine learning are not isolated principles. Rather, they are interconnected, often both supporting and hindering one another, creating a complex web of ethical considerations. In this thesis, our aim is to shed light on these intricate relationships and the tensions they give rise to. An example of these tensions can be found in the field of XAI. While XAI plays a pivotal role in enhancing transparency and accountability in machine learning models, it also introduces additional risks, both on the level of privacy as well as manipulation. In this thesis, we will show how it can both benefit decision subjects (by offering transparency and the means to spot discriminatory patterns) as hurt them by allowing for new forms of both manipulation and privacy attacks.

Part II

TRANSPARENCY

4

The Cost of Comprehensibility

A key challenge in Artificial Intelligence (AI) has been the potential trade-off between the accuracy and comprehensibility of machine learning models, as this also relates to their safe and trusted adoption. While there has been a lot of talk about this trade-off, there is no systematic study that assesses to what extent it exists, how often it occurs, and for what types of datasets. Based on the analysis of 90 benchmark classification datasets, we find that this trade-off exists for most (69%) of the datasets, but that somewhat surprisingly for the majority of cases it is rather small while for only a few it is very large. Comprehensibility can be enhanced by adding yet another algorithmic step, that of surrogate modelling using so-called 'explainable' models. Such models can improve the accuracy-comprehensibility trade-off, especially in cases where the black box was initially better. Finally, we find that dataset characteristics related to the complexity required to model the dataset, and the level of noise, can significantly explain this trade-off and thus the cost of comprehensibility. These insights lead to specific guidelines on how and when to apply AI algorithms when comprehensibility is required.

4.1 INTRODUCTION

In 2019, a series of tweets went viral where a tech entrepreneur was complaining about the fact that Apple Card offered him twenty times the credit limit that it offered to his wife, although they had shared assets. After complaining to Apple representatives, he got the reply: "I don't know why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM" [Agrawal, November 12, 2019, Martens, 2022]. Apple co-founder Steve Wozniak replied that the same thing happened to him and his wife and added [Wozniak, November 10, 2019]: "Hard to get to a human for a correction though. It's big tech in 2019." These complaints led to a formal investigation into the potential sexist credit scoring by Apple Card [Agrawal, November 12, 2019, Martens, 2022]. This example shows how predictive modelling is facing major challenges due to its inability to explain its decisions, which often stems from the use of complicated models. But why is everyone using these kinds of models? It is often claimed that they have a higher performance than more simple models, but is this always true? How often is it the case and to what extent?

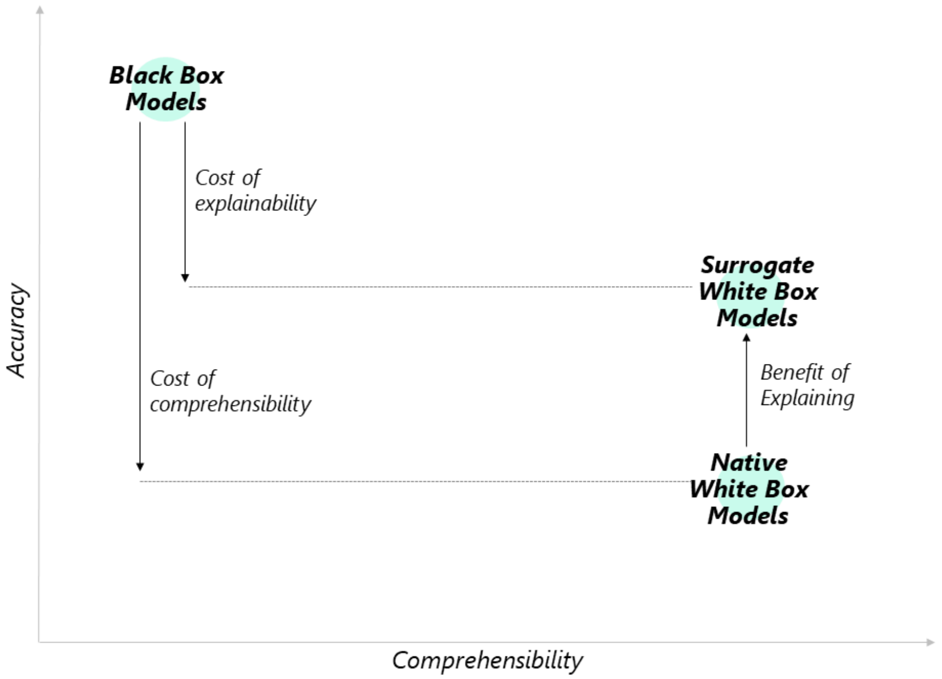


Figure 4.1: Definitions of the Cost of Comprehensibility, the Cost of Explainability and the Benefit of Explaining

This trade-off between accuracy and comprehensibility is arguably one of the important debates in Artificial Intelligence (AI)[1] [Breiman, 2001b, DAR, 2016]. This trade-off can either

[1] We focus on prediction models trained on data using machine learning algorithms.

limit the performance of AI , if accuracy is lost due to comprehensibility restrictions (for example imposed by regulators) [Martens et al., 2007, Wachter et al., 2017a], or hurt AI adoption, if user trust is lost due to opaqueness [Linardatos et al., 2021]. The Apple Card example shows that companies may use black box models to achieve higher predictive performance, but with the risk of being unable to explain their AI decisions to users or regulators. However, while there has been a lot of research mentioning this trade-off, with most claiming there is one [DAR, 2016, Linardatos et al., 2021, Freitas, 2014, Murdoch et al., 2019] and others contradicting this [Rudin and Radin, 2019, Makridakis and Hibon, 2000], there is no systematic study that assesses to what extent there indeed exists a trade-off and for what types of datasets.

The goal of this paper is to provide such a systematic study. We focus on tabular datasets as we believe that for these datasets the trade-off would be less clear - and possibly smaller than expected. Deep learning models, which are models composed of multiple layers to learn representations of data with multiple levels of abstraction [LeCun et al., 2015] and can thus be considered as black box models, perform very well for classification on homogenous data such as image, audio or text but they not necessarily outperform other machine learning techniques on tabular datasets [Borisov et al., 2022, Popov et al., 2019, Arik and Pfister, 2021].

Based on the analysis of 90 benchmark datasets across different domains, we study the nature of the differences between the accuracies among a number of widely used a) opaque ("black box") models, b) comprehensible ("white box") models, and c) surrogate models used to develop a comprehensible surrogate of the opaque ones. We call the difference between (a) and (b) "Cost of Comprehensibility", that between (a) and (c) "Cost of Explainability", and that between (b) and (c) the "Benefit of Explaining" (Figure 4.1).[2] Our main findings are: first, there is indeed a trade-off but somewhat surprisingly it appears to be highly non-linear across datasets. Both costs are relatively small for most datasets, but very large for a few. Second, there are datasets for which the comprehensible models perform as well or better than the black box models, supporting that one should not forgo trying comprehensible models [Zeng et al., 2017]. We call these datasets "comprehensible datasets", as opposed to datasets where the black box is strictly better which we call "opaque datasets". Understanding what makes a dataset "opaque" vs "comprehensible" and more so, given the non-linearities observed, what makes the costs very high (positive or negative) is a challenging question as it relates to understanding the data generation processes themselves (e.g., the "nature" of the data and problem at hand). We discuss initial results indicating that some of the main differences between opaque and comprehensible datasets are about their inherent complexity as well as the level of noise in the data. The results indicate that reporting some simple characteristics of a dataset can provide clues, for example to users or regulators, about the potential accuracy and comprehensibility trade-off. To summarize, the contributions of our paper are threefold:

- A benchmark study comparing state-of-the-art white box and black box algorithms on 90 tabular datasets, and assessing their difference in performance;

[2] We note that the terms "interpretability", "comprehensibility" and "explainability" have also been used in different ways in the literature. We use one of the definitions of comprehensibility, but note that many more exist. For example, comprehensibility can also be enforced in the training process, as was done by Carrizosa et al. [2017].

- An analysis of whether surrogate modelling could improve any trade-off between comprehensibility and accuracy;
- Insights in how dataset properties could predict the nature/size of the trade-off we study.

4.2 BACKGROUND AND SETUP OF THE STUDY

4.2.1 *What is comprehensibility?*

We refer to the discussion in Section 3.2 for an overview of comprehensibility in machine learning. As mentioned, there are two main approaches commonly used: inherently comprehensible models and post-hoc explanations. We will focus on the former, and investigate models that are comprehensible *by nature*. However, comprehensibility is very difficult to measure due to its subjective nature. Some compare the comprehensibility of models using user-based surveys [Huysmans et al., 2011, Allahyari and Lavesson, 2011] while others based on mathematical heuristics [Freitas, 2014], typically the size of the model (e.g., number of rules for a rule learner, number of nodes for a decision tree, or number of variables for a linear model) [Askira-Gelman, 1998, Bibal and Frénay, 2016, Freitas, 2019, Rüping et al., 2006]. Very deep decision trees, for example, can be considered as less comprehensible than a compact neural network [Lipton, 2018]. We use the latter, heuristic approach to measure comprehensibility due to its objectivity and scalability.

4.2.2 *What are intrinsically comprehensible models?*

In line with the literature, we consider small decision trees, rule sets and linear models as comprehensible or "white box" models [Linardatos et al., 2021, Molnar, 2020, Guidotti et al., 2018, Stiglic et al., 2020]. We limit the size of these models during training in order for them to be comprehensible. We opted for seven as the size limit for comprehensibility, based on cognitive load theory [Miller, 1956]. According to this theory, the span of absolute judgement and the span of short-term memory pose severe limitations on the amount of information that humans can receive and process correctly, with seven being the typically considered maximum size in both cases [Miller, 1956]. We consider larger decision trees[3], rule sets and linear models as "black box" ones. We also consider three other machine learning methods in

[3] A decision tree of eight nodes is arguably not a black box model, and may be in a "grey zone" of comprehensibility. For this reason, in our experiments we focus on the very large and small trees, rule sets and linear models, defined as those with size larger than 50 or smaller than 8 (in number of nodes/rules/coefficients) as it is a general assumption in the literature that smaller decision trees are more comprehensible than larger ones due to the cognitive size limit [Freitas, 2014, Huysmans et al., 2011, Confalonieri et al., 2019, Ramon et al., 2021b]. This focus ensures that our findings are applicable to all applications and end users, because of the arbitrariness to consider models with size between 8 and 50 as black box, which actually depends on the application and end user.

the list of black boxes we test: neural networks, random forests and nonlinear support vector machines. It is generally agreed upon that these algorithms are not comprehensible as their line of reasoning cannot be followed by human users. We base this choice of black box models on the results of benchmark studies in the literature, where these often are among the best performing ones, as can be seen in Table 4.1.[4] Comparing all possible models available is of course infeasible, which is a practical limitation of such a study. All the papers mentioned in Table 4.1 compare different machine learning models but none investigate the difference in performance between the best black box model and the best white box model, nor whether this can be linked to any dataset properties. Many papers claim that black box models will always have a better performance, or on the contrary that simpler models work equally well [Rudin and Radin, 2019, Makridakis and Hibon, 2000], but a large-scale study about the difference of performance is missing.[5]

ML algorithms	Count	*Olson et al. (2017)*	*Fernandez-Delgada et al. (2014)*	*Zhang et al. (2017)*	*Lessman et al. (2015)*	*Mayr et al. (2018)*	*Lorena et al. (2011)*	*Macia and Bernado-Mansilla et al. (2014)*
Random Forest	7	✓	✓(1)	✓(1)	✓	✓(3)	✓(1)	✓
Bayesian	7	✓	✓	✓	✓	✓	✓	✓
SVM	7	✓	✓(2)	✓(2)	✓	✓(2)	✓(2)	✓
LR	6	✓	✓	✓	✓		✓	✓
Nearest Neighbor	6	✓	✓	✓	✓	✓	✓	
Neural networks	5		✓(3)	✓(3)	✓	✓(1)	✓	
Decision tree	4	✓	✓(4)	✓				✓
Boosting	3	✓	✓	✓(1)				
Discriminant analysis	2		✓		✓			
Bagging	1		✓					
Rule-based	1				✓			

Table 4.1: Models that are used in other benchmark studies. Symbol ✓indicates that this kind of model was used in the study, and the numbers between brackets indicate the rank of the model (if this was included in the study).

4.2.3 *Surrogate modelling*

A common practice is to mimic the predictions of a black box with a global white box surrogate model, in order to improve the accuracy while remaining comprehensible [Fung et al., 2005, Martens et al., 2008b]. The typical process is to first build a black box model using the available training data, and then build a comprehensible model by training a white box model using the predictions of the black box instead of the original training data. This process

[4] We do not include k-nearest neighbors and Bayesian networks, which are also used frequently in other benchmark studies, as it is debatable whether they can be considered as comprehensible models. K-nearest neighbors lacks global model comprehensibility as there are is no global model structure learned [Molnar, 2020] and in Bayesian networks, it is not easy to interpret the mapping implicit in the network or do other data inference tasks, as the reasoning method is not necessarily aligned with human reasoning [Lacave and Díez, 2002, Chubarian and Turán, 2020].

[5] Besides white box and black box models, some researchers also mention the existence of "grey box" models, which are defined as aiming to develop an ensemble of black and white box models and acquire the benefits of both by being nearly as accurate as black box models but more comprehensible [Pintelas et al., 2020]. As the literature is not conclusive on whether grey boxes are always as comprehensible as white box models [Pintelas et al., 2020, Freitas, 2019, García, 2020], we will focus only on the trade-off between black box and white box models in this study.

is called surrogate modelling [Molnar, 2020], oracle coaching [Johansson et al., 2012, 2014], or rule extraction in case the white box model is a decision tree or rule set [Martens et al., 2007, Craven and Shavlik, 1995]. A key metric of the quality of the surrogate model is *fidelity*, which measures how well the predictions of the surrogate model match those of the black box [Zhou, 2004]. The most common goal of this kind of modelling is to use the surrogate model to explain the black box model, while still using the black box to make predictions. This requires of course that the surrogate model is (1) more comprehensible than the black box model and (2) sufficiently explains the predictions made (high fidelity).

One can also use the surrogate model instead of the black box to make predictions, in order to improve the performance one could achieve using only comprehensible models. A possible reason why this approach can work, instead of just training a white box model directly using the training data, can be that the black box model may filter out noise or anomalies that are present in the original training data [Johansson et al., 2014, Martens et al., 2008a]. In this case, a comprehensible model mimicking a black box may be more accurate than a comprehensible model trained on the original data, as shown in some previous work [Martens et al., 2008b, Johansson et al., 2012, 2014]. Therefore, we also investigate whether surrogate modelling can lead to better performing comprehensible models and, as such, improve the trade-off we study. Specifically, for each dataset we train a white box on the predictions of the *best* performing black box for that dataset. We call this a *surrogate white box model* as opposed to a comprehensible model trained on the training dataset which we call a *native white box model* - see Figure 4.1.[6]

4.2.4 *Dataset properties*

Finally, we study whether there are simple (standard) properties of a dataset that may determine whether it is opaque (the best black box model outperforms the best white box) or comprehensible (the reverse happens). We use a standard toolbox [Alcobaça et al., 2020], which automatically extracts numerous characteristics ("meta-features") for any given dataset. We consider four types of dataset characteristics from this toolbox: general ones, which capture basic information such as the number of instances or the number of attributes [Rivolli et al., 2018]; statistical ones, which capture information about the data distribution such as the number of outliers, variance, skewness, etc. [Rivolli et al., 2018]; information-theoretic ones, which capture characteristics such as the joint entropy, class entropy, class concentration, etc. [Rivolli et al., 2018]; and so-called complexity related ones, which, for example in the case of a classification problem estimate the difficulty in separating the data into their classes [Lorena et al., 2019].[7] We opt for using a standard toolbox and set of dataset characteristics to make this analysis general, easily reproducible and simple to use in practice. A list of the used dataset properties and their meaning can be found in Table A.2.

[6] In this study, we make a distinction between the surrogate models and the native white box models, but surrogate models are in se just one type of white box models.

[7] See Supplementary Information material.

4.3 materials and methods

4.3.1 *Materials*

We use a large benchmark study to compare the algorithms on different tabular datasets. Benchmark comparisons are usually developed over a few, typically standard data sets, as a machine learning method might perform well on some of the datasets but not generalize to a broader range of problems [Olson et al., 2017].
To perform our experiments, we use all the binary classification datasets from the Penn Machine Learning Benchmark (pmlb) suite [Olson et al., 2017]. This is a dataset suite that is publicly available on Github[8], which consists both of real-world and simulated benchmark datasets to evaluate supervised classification methods. It is compiled from a wide range of existing ML benchmark suites such as KEEL, Kaggle, the UCI ML repository and the meta-learning benchmark. At this moment, PMLB consists of 162 classification datasets and 122 regression datasets. We focus on the binary classification datasets which amount to 90 datasets in total.

Some preprocessing was already done by the compilers of this benchmark suite. All the datasets were preprocessed to follow a standard row-column format and all the categorical and features with non-numerical encodings were replaced with numerical equivalents. All datasets with missing data were excluded, to avoid the impact of imposing a specific data imputation method. The used datasets are shown in Table A.1.

4.3.2 *Methods*

Our methodology is shown in Figure 8.3. For each dataset we create a training and test set, using 75% of the data for training and 25% for testing. Both the training and the test set are scaled according to the parameters of the training set with Sklearn's MinMaxScaler.[9] This estimator scales each feature individually so that it is between zero and one on the training set. We also use a stratified split to make sure that enough labels are present for the training phase. GridSearchCV from *Sklearn*[10] is used with its default 5-fold cross validation to tune the hyperparameters of every model. The dataset is divided in five folds, where each time another fold is taken as the validation set. GridSearchCV then performs an exhaustive search over a specified hyperparameter grid (which is reported in Section 4.3.2 and in Section 4.3.2 for each modelling technique) and then checks on the validation set which parameter settings performed best. By doing this five times, instead of just using one validation set, we get a more accurate representation of how the model behaves on unseen data, and we are not reliant on the data we used as the validation set. We select the best hyperparameter values for each modelling technique based on this tuning. Moreover, for each dataset we also select

[8] https://github.com/EpistasisLab/pmlb
[9] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

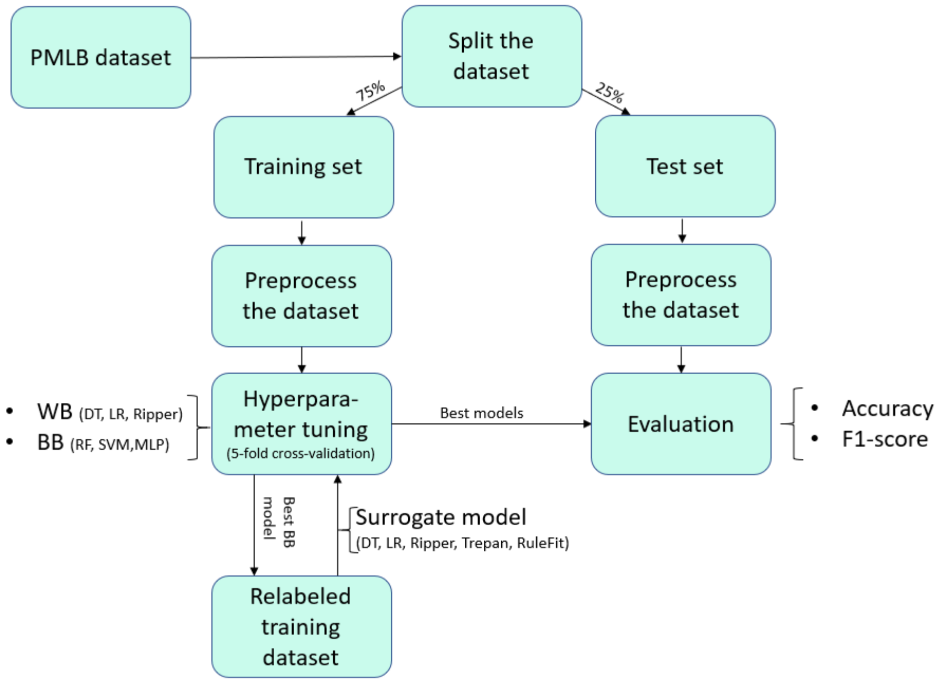


Figure 4.2: Methodology

the best surrogate model. We do this by creating a new training set, which is a copy of the original training set but with as labels the predictions of the best black box model, based on the cross-validation performance. The surrogate model is trained on this relabeled training set and can be any of the original white box models, as well as Trepan or RuleFit. The final performance of all the models (black box, white box and surrogate) is evaluated on the test set based on two metrics: accuracy and f1-score. The difference in the test set performance among the different models is shown in Figure 4.3. For each dataset we select the best black box, the best white box and the best surrogate, based on their performance on the test set.[11] In our aggregate analyses, we compare the test performances of these across all datasets.

[11] Note that using the test data to select the best black and white boxes and then reusing the same data to compare those two across all datasets adds some bias in the results. We opt for this approach (instead of also using, for example, a validation set) as some datasets do not have many observations and we only select among a few (in total six) black boxes and among a few (in total three) white ones, making the bias small. We also verified whether our results are robust when using cross validation to select the best model and note that our results indeed hold (e.g., still for 68.89% of the datasets, the best black box model outperforms the best white box model).

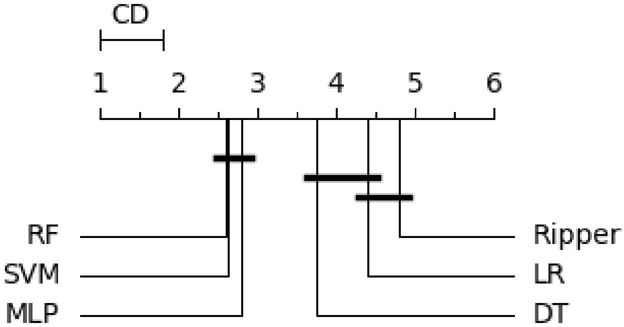


Figure 4.3: Critical difference diagram of the comparison of classifiers. Models that are not connected with a bold line have a significant difference in performance (at a 5% level with the Nemenyi test).

Black Box Models

We use three state-of-the-art black box models: neural networks, random forests and nonlinear support vector machines [Baesens et al., 2003, Singh et al., 2016]. The functioning of each model is described in more detail in Section 2.2. As noted below, we also include in the list of black boxes the three comprehensible models when their size - after training - is very large.

RANDOM FOREST We use the RandomForestClassifier[12] from *Sklearn* and use a grid search to tune the number of trees in the forest with several values between 10 and 2000 and the number of features to consider when looking for the best split with (*'sqrt'*, *'none'*).

SUPPORT VECTOR MACHINE We use the SVC[13] from *Sklearn* and use a grid search to tune the regularization hyperparameter with values between 0.1 and 1000 and the kernel coefficient with several values between 0.0001 and 1. We use the default kernel type of *rbf*.

NEURAL NETWORK We use the MLPClassifier[14] from *Sklearn* and use a grid search to tune the size of the hidden layer. We only test neural networks with one hidden layer. We tune the hidden layer with several sizes between 10 and 1000.

[12] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[13] https://scikit-learn.org/stable/modules/svm.html
[14] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Comprehensible Models

We use three models that are in general considered to be comprehensible, when their size is constrained. As discussed in the main article, we limit the size of these models to 7 (maximum number of nodes for trees, rules for rule based systems, coefficients for logistic regression). We also train these models without constraining their size. In this case, when their size after training is very large, with more than 50 elements,we consider them as part of the black boxes in our analysis. The functioning of each model is described in Section 2.2.

DECISION TREE We use the DecisionTreeClassifier[15] from *Sklearn*. We use a grid search to tune the function to measure the quality of the split (gini, entropy), tune the maximal depth between 2 and 30 and tune the minimum number of samples in a leaf (2,4). We tune the maximal amount of leaf nodes between 2 and 7 for the constrained cases (white boxes) and between 2 and 1000 for the unconstrained ones (black boxes).

LOGISTIC REGRESSION We use the LogisticRegression[16] from *Sklearn*. We use $l2$ regularization and the liblinear solver. We use a grid search to tune the regularization parameter values between 0.0001 and 1000.

RIPPER We use a rule learning algorithm, based on sequential covering. This method repeatedly learns a single rule to create a rule list that covers the entire dataset rule by rule [Molnar, 2020]. RIPPER (Repeated Incremental Pruning to produce Error Reduction), which was introduced by Cohen in 1995 is a variant of this algorithm [Cohen, 1995]. We use the Python implementation of Ripper hosted on Github.[17]

Surrogate Models

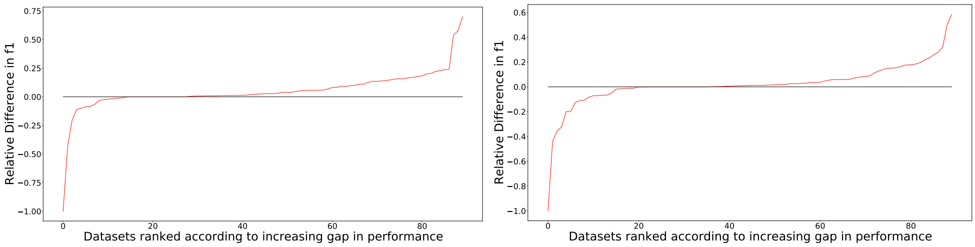
We use the three comprehensible models above but this time we train them on the predictions of the best performing black box instead of using the training data. We also include Trepan [Craven and Shavlik, 1995], which is used for rule extraction based surrogate modeling, and RuleFit [Friedman and Popescu, 2008], which is based on an underlying Random Forest model. Again, we limit the size of the comprehensible models to 7.

TREPAN We use the Python package Skater to implement TreeSurrogates[18], which is based on Craven and Shavlik [1995]. The base estimator (oracle) can be any supervised learning model. The white box model has the form of a decision tree and can be trained on the decision boundaries learned by the oracle. We use the same hyperparameter settings to tune the decision trees from Trepan as for the DecisionTreeClassifier.

[15] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[16] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[17] Imoscovitz. Ripper Python package. url: https://github.com/imoscovitz/wittgenstein
[18] A. Kramer et al Skater Python package. url: https://github.com/oracle/Skater

RULEFIT The RuleFit algorithm learns sparse linear models that include automatically detected interaction effects in the form of decision rules [Friedman and Popescu, 2008]. The interpretation is the same as for normal linear models but now some of the features are derived from decision rules. We use the Python implementation of RuleFit hosted on Github.[19]

4.4 RESULTS



(a) Non-linearity of the cost of comprehensibility (b) Non-linearity of the cost of explainability

Figure 4.4: Comparing black box and white box models. For both plots, the datasets are ordered according to the gap in f1-score between the best black box and the best native (left figure) or surrogate (right) white box model (right). The y-axis measures the relative difference in the f1-score, defined as the ratio of the difference between the black and white box f1-scores divided by that of the best model.

First, we address the cost of comprehensibility, by testing whether native white and black box models have a significant difference in performance. To assess this cost, we use both the models' *f1-score* and *accuracy*.[20] The figures for the latter are reported in the Appendix. We first compare all the classifiers using the Friedman test[21] [Demšar, 2006] to identify whether there are any significant differences between the different models, and then the post-hoc Nemenyi test [Nemenyi, 1963] to identify significant pairwise differences.[22] The null hypothesis of the Friedman test is rejected with a p-value of $2.43 \cdot e^{-25}$ (a value with the same order of magnitude when using accuracy instead of f1-scores). This means that there are significant differences among some groups of algorithms. We use the post-hoc Nemenyi test to perform all possible pairwise comparisons [Trawiński et al., 2012]. The results are shown in the critical difference diagram[23] in Figure 4.3. The performance of the black box models (RF, MLP, SVM) is significantly better than the performance of the white box models (DT, LR, Ripper), already confirming that, overall, the cost of comprehensibility indeed exists.

[19] Molnar. RuleFit Python package. url:https://github.com/christophM/rulefit

[20] We include the results with f1-score to account for imbalance issues that could bias our results.

[21] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.friedmanchisquare.html

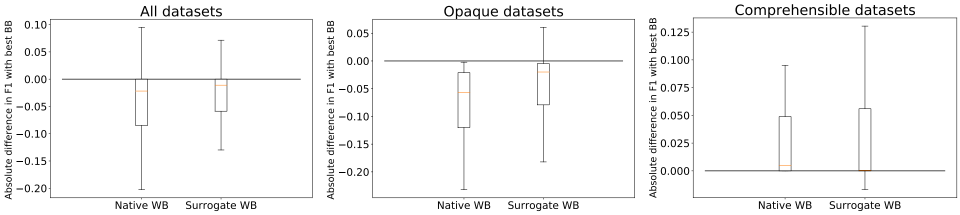
[22] We cannot just use a pairwise comparison because this would inflate the probability of a type I error. The Friedman test is the non-parametric equivalent to the repeated-measures ANOVA [Demšar, 2006].

[23] These diagrams were created with the Orange Data Mining Library [Demšar et al., 2013].

4.4.1 *The Cost of Comprehensibility*

Having established that the cost of comprehensibility exists, we study how large it is across datasets. As discussed, for each dataset we select the best black and white box models and measure their relative difference in performance - namely, the cost of comprehensibility. Figure 4.4a shows the results across all datasets when we order them according to this cost. This figure reveals a somewhat surprising result: this cost is highly non-linear (e.g., the plot is a sigmoid instead of being closer to a straight line). For most datasets the accuracy-comprehensibility trade-off is low, only for a few it is very high (right) and for a few it is very "negative" indicating that comprehensible models largely outperform the black box ones for these datasets (left). Yet, for 68.89% of the datasets the best black box model outperforms the best white box model, reconfirming the overall existence of the cost of comprehensibility.

4.4.2 *Can Surrogate Modeling Improve the Accuracy-Comprehensibility Trade-off?*



(a) Comparison of performance of the native and surrogate white box model across all the datasets.

(b) Comparison of performance of the native and surrogate white box mode but only across all opaque datasets.

(c) Comparison of performance of the native and surrogate white box mode but only across all comprehensible datasets.

Figure 4.5: Comparison across datasets of best black box model for each dataset, surrogate white box model mimicking this best black box, and best native white box model. BB stands for black box and WB for white box. The line at 0 indicates the performance of the best black box model. The y-axis indicates the absolute difference in f1-score from the best black box model.

We next investigate whether surrogate modelling can improve the performance of the (native) comprehensible models. For all datasets we generate the best black box and the best (native) white box trained on the training data, and then we also train a surrogate model mimicking the best black box one - what we previously called a surrogate white box. We compare the performance of these three types of models across all datasets in Figure 4.5. As indicated in Figure 4.5a, surrogate modelling does improve accuracy slightly relative to native white box models, on average across all datasets. We term this improvement the "Benefit of Explaining", a benefit in terms of improved predictive accuracy. Based on the Wilcoxon Signed Rank test[24] [Demšar, 2006], used to compare classifiers across several datasets, we can reject the hypothesis that the native and surrogate white boxes perform equally well (p-value 0.003) –

[24] https://scikit-posthocs.readthedocs.io/en/latest/generated/scikit_posthocs.posthoc_wilcoxon/

the latter performing on average better.

We perform the same analysis, but this time for two different types of datasets: those for which the best performing model is a black box, what we termed opaque datasets, and those for which white boxes perform at least as well as or better than black boxes, what we called comprehensible datasets. The results are shown in Figure 4.5b and Figure 4.5c. Interestingly, in this case the surrogate white box models outperform the native white box models on average across the opaque datasets (Wilcoxon test p-value of $7.72 \cdot e^{-5}$), while the two are not significantly different for the comprehensible datasets (Wilcoxon test p-value of 0.20). For these datasets, there is no need to go through a black box if its performance is not better than that of a native white box [Martens et al., 2008a, de Fortuny and Martens, 2015], as the latter would dominate both in terms of accuracy and comprehensibility. Hence, if one considers only opaque datasets, the use of surrogate modeling can indeed improve the accuracy-comprehensibility trade-off on average.

4.4.3 *The Cost of Explainability*

Next, we investigate the difference in performance between the best black box model for each dataset and the best surrogate white box model from that black box - what we call the cost of explainability. Figure 4.4b shows the results when we sort all datasets based on this cost. The results are similar to what we observe for the cost of comprehensibility: the difference is small for most datasets, but very large for a few. The results are also in agreement with those in Figure 4.5, where we see that the cost of explainability is a bit lower than the cost of comprehensibility.

4.4.4 *Opaque vs Comprehensible Datasets*

Finally, we study whether the cost of comprehensibility relates to some properties of the dataset. To do so, for each dataset we generate a number of standard dataset properties as discussed above (see also Table A.2), and use them to explain the cost of comprehensibility. Specifically, we run a regression analysis using the generated dataset properties as independent variables with the dependent variable being the difference between the performance of the best black box model and the best native white box model. We used all 90 datasets, hence the number of observations used for the regression was also 90. The variables that are significant are shown in Table 4.2. Overall, these results indicate that properties related to the complexity required to model a dataset and the level of noise in a dataset significantly explain the cost. While this is a relatively simple analysis, the results suggest that one may be able to identify or communicate whether there is a potential cost of comprehensibility by simply reporting specific dataset properties.

Table 4.2: The dataset properties that are significant when explaining the cost of comprehensibility using a number of standard dataset properties as independent variables in a regression model where the cost is the dependent variable.

Variable	**MSE**	**P-value**	**Coef**
EqNumAttr	0.508	$4.57 \cdot e^{-10}$	-0.72
NsRatio	0.508	$4.57 \cdot e^{-10}$	-0.72
N3	0.148	0.00191	-0.23
F1v	0.139	0.00267	0.15
L1	0.089	0.0170	0.12

Specifically, the following five properties are found to be significant. *F1v*, which is the directional-vector Maximum Fisher's discriminant ratio that indicates whether a linear hyperplane can separate most of the data, where lower values means that more data can be separated this way [Lorena et al., 2019]. *L1*, which is a linearity measure that quantifies whether the classes can be linearly separated [Rivolli et al., 2018]. Higher values of this attribute indicate more complex problems as they require a non-linear classifier [Lorena et al., 2019]. These properties have a positive coefficient in the regression analysis, which means that all these factors increase the gap between the best black box model and the best white box model. The sign of these coefficients is as expected, namely that for datasets that are more complex to separate linearly, the performance of black box models compared to simple models is on average better.

Two other features, *EqNumAttr* and *NsRatio*, capture information related to the minimum number of attributes necessary to represent the target attribute and the proportion of data that is irrelevant to the problem (level of noise) [Rivolli et al., 2018, Michie et al., 1994]. We see that these dataset properties have a negative relationship with the size of the cost. Note that when we analyze this result at the level of each individual prediction model, we see that these properties negatively affect both the performance of the black box models and the white box models, but more so for the black box ones. This could be because black box models may pick up more of the noise or use a lot of irrelevant features. Perlich et al. [2003] also find that when the signal-to-noise ratio is higher (so the opposite from these features as they measure the amount of noise), complex models perform better. Finally, *N3* [Lorena et al., 2019] is a neighbor-based measure that refers to the error rate of the nearest neighbor classifier. Low values of this dataset property indicate that there is a large gap in the class boundary [Luengo and Herrera, 2015]. We see again that this property negatively affects both the performance of the black box models and the white box models [Luengo and Herrera, 2015], and that the effect on the gap depends on how much it affects the performance of each model.

4.5 DISCUSSION

Understanding the trade-off between comprehensibility and accuracy can have important implications for regulators as well as companies [Adadi and Berrada, 2018]. Our results indicate that most of the time the trade-off is relatively small, indicating that one should consider native white box algorithms as a key benchmark. Indeed, given the non-linearities we observe, one would expect that black boxes are used relatively infrequently, even if for the majority of cases they outperform white boxes, as our study indicates that this outperformance is typically relatively small. Some papers in the literature also confirm that for certain datasets simple models work as well as complex ones [Rudin and Radin, 2019, Makridakis and Hibon, 2000] or that for most datasets the out-performance by black box models will be very small [Schwartzenberg et al., 2020], despite the popular belief that more complex models are always better. Of course it depends on the use case and application domain whether this small difference in performance is worth the loss in comprehensibility. Due to social and ethical pressure, awareness in when one should opt for a comprehensible model could be a competitive differentiator and drive real business value [Adadi and Berrada, 2018]. Insights in this trade-off could lead to specific guidelines from regulators on how and when to apply AI algorithms when comprehensibility is required.

Our results also show that using surrogate modelling could reduce the cost of comprehensibility, especially for opaque datasets. As we discussed, this may be the case because the black box model in between can filter out noise and anomalies [Johansson et al., 2014, Martens et al., 2008a]. We also see that simple properties of a dataset could provide insights (for example to a third party such as a user or regulator) in the nature of the trade-off without requiring knowledge of the algorithms tested or the data used. For example, attributes that measure how difficult it is to linearly separate the data are significantly correlated with the size of the gap. Indeed, one would expect that for these datasets black box models might be better in capturing the non-linearities. This can lead to practical tests of the feasibility of using a native white box – and the potential accuracy loss – in a given use case.

Our general findings suggest the following guidelines:

1. Start with white box models.

2. Train additional black box models if: (a) the application allows for a (possibly small) increase in performance at a cost of comprehensibility, and, (b) the level of noise is high and the data requires complex modeling, as indicated by the listed, easy to calculate dataset metrics.

3. If there is a practically important cost of comprehensibility (hence you are dealing with an opaque dataset), apply additional surrogate modeling algorithms.

Finally, we note that in this study we focused on tabular datasets. For other kinds of datasets, the trade-off we study may be different.

There are also some limitations to this study. First, the results strongly depend on the choice of datasets. To avoid any selection bias from our side, we used all the binary classification datasets from the PMLB repository [Olson et al., 2017], but we do not know whether this sample is representative of all binary classification datasets. Next, our results obviously depend on the choice of models and the used preprocessing steps. We did not take into account feature selection of feature engineering, which could also have an impact on the results.

5

Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem

Artificial Intelligence (AI) systems are increasingly used in high-stakes domains of our life, increasing the need to explain these decisions and to make sure that they are aligned with how we want the decision to be made. The field of Explainable AI (XAI) has emerged in response. However, it faces a significant challenge known as the disagreement problem, where multiple explanations are possible for the same AI decision or prediction. While the existence of the disagreement problem is acknowledged, the potential implications associated with this problem have not yet been widely studied. First, we provide an overview of the different strategies explanation providers could deploy to adapt the returned explanation to their benefit. We make a distinction between strategies that *attack* the machine learning model or underlying data to influence the explanations, and strategies that *leverage* the explanation phase directly. Next, we analyse several objectives and concrete scenarios the providers could have to engage in this behavior, and the potential dangerous consequences this manipulative behavior could have on society. We emphasize that it is crucial to investigate this issue now, before these methods are widely implemented, and propose some mitigation strategies.

5.1 INTRODUCTION

Artificial Intelligence (AI) is used in more and more high-stakes domains of our life such as justice [Berk, 2012], healthcare [Callahan and Shah, 2017], and finance [Lessmann et al., 2015], increasing the need to explain these decisions and to make sure that they are aligned with how we want the decision to be made. However, the complexity of many AI systems makes them challenging to comprehend, posing a significant barrier to their implementation and oversight [Arrieta et al., 2020, Samek et al., 2019]. Legislative initiatives, including the EU General Data Protection Regulation (GDPR), have recognized the 'right for explanation' for individuals affected by algorithmic-decision making, emphasizing the legal necessity of explainability [Goodman and Flaxman, 2017]. In response, the field of Explainable Artificial Intelligence (XAI) has emerged, aimed at developing methods for explaining the decision-making processes of AI models [Adadi and Berrada, 2018, Holzinger et al., 2022, Xu et al., 2019].

Nevertheless, the landscape of post-hoc explanations is diverse, and each method can yield a different explanation. Furthermore, even within a single explanation method, multiple explanations can be generated for the same instance or decision. This phenomenon, known as the *disagreement problem*, has been studied in literature [Brughmans et al., 2023b, Krishna et al., 2022, Neely et al., 2021, Roy et al., 2022]. While the existence of the disagreement problem is acknowledged, the potential implications of this problem have not yet been extensively explored. Barocas et al. [2020] already mention that the power to choose which explanation to return, leaves the providers with significant room to promote their own welfare. Aïvodji et al. [2019] discuss the possibility of fairwashing, where discriminatory practices can be hidden by selecting the right explanations, while Bordt et al. [2022] argue that post-hoc explanations fail to achieve their purpose in adversarial contexts. Finally, Carli et al. [2022] highlight how singular explanations can already be a source of manipulation as they can interfere with the users' natural decision-making process. However, an overview of potential misuses by the explanation provider is still missing from the literature, and we believe it is imperative to study the implications now, before explainability methods are implemented on a wide scale. The main contributions of this paper are:

- Providing a comprehensive framework that outlines the different strategies that could be employed by malicious entities to manipulate the explanations.

- An overview of the different objectives these actors could have to engage in this behavior, and the potential implications.

This paper is structured as follows: We introduce the field of Explainable AI and the disagreement problem in Sections 5.2 and 5.3. In Section 5.4, we explore various strategies that providers could employ to manipulate the explanations according to their preferences. Additionally, in Section 5.5, we present specific objectives and scenarios that may drive providers to engage in such behavior. Finally, in Section 5.6, we offer discussion and potential solutions to address this.

	Sex	Age	Residence time	Home status	Occupation	Job status	Employment time	Other investments	Bank account	Time at bank	Liability	Account reference	Housing expense	Savings account
Instance	2	16	22	1	2	6	7	0	0	0	0	1	1	125
CBR	2	16	0.25	1	2	6	7	0	1	0	0	1	1	125
DiCE	2	16	22	1	2	6	7	24	0	0	0	1	1	125
GeCo														
NICE(none)	2	34	0	3	3	10	1	0	0	0	0	1	2	136
NICE(plaus)	2	34	0	3	3	6	1	0	0	0	0	1	2	136
NICE(prox)	2	34	0	1	2	10	1	0	0	0	0	1	1	136
NICE(sparse)	2	16	0	1	2	10	1	0	0	0	0	1	1	136
SEDC	2	16	22	1	2	6	7	0	1	0	0	1	1	125
WIT	1	278	8	2	1	5	1	6.5	1	1	6	0	0	102

Table 5.1: Illustration of the disagreement problem for an instance of the Australian Credit dataset.

5.2 EXPLAINABLE AI

For an overview of why we need Explainable AI, and the various techniques at our disposal, we refer to Section 3.2.3. Based on this overview, the first question we can ask ourselves is, what actually is the definition of an explanation? How do we decide whether something is an explanation? Can anything count ask an explanation? The fact that no real definition for an explanation exists, leads to a plethora of explanation techniques, and consequently, to the disagreement problem. However, this is not the entire issue, as even within one clearly defined explanation technique, some randomness can exist which can lead to multiple explanations (as we will discuss later). While this paper predominantly uses counterfactual explanations as an example, the findings and discussion presented are applicable to other post-hoc explanation techniques as well. At the moment, we do not see manipulation issues for inherently transparent models but this would be an interesting avenue for future research [Bordt et al., 2022].

In line with Greene et al. [Greene et al., 2023], we define an explanation recipient as a person who requests an explanation for an automated decision, and an explanation provider as the entity who provides the algorithmic explanations to the recipient. For example, in the domain of finance, the explanation provider could be a bank, and the explanation recipient a loan applicant; while in the domain of employment the explanation recipient would be the job applicant, and the explanation provider the hiring agency [Greene et al., 2023]. Not all scenarios described in Section 5.5 assume that there is one actual recipient; the explanation provider can also provide explanations of the model to the public proactively or to comply with regulatory requirements.

5.3 THE DISAGREEMENT PROBLEM

A known issue within Explainable AI is that the results of different explanation techniques do not always agree with each other. Even one explanation technique can generate many different explanations for one instance, which is known as the disagreement problem [Krishna et al., 2022, Neely et al., 2021, Roy et al., 2022]. One of the reasons behind the disagreement problem is that a 'true internal reason' why the machine learning model comes to a certain decision, generally does not exist [Bordt et al., 2022]. For example, for feature importance methods such as SHAP and LIME, there is no mathematically unique way to determine the importance of each feature to the decision of a black-box function [Bordt et al., 2022, Sundararajan and Najmi, 2020]. As a consequence, all feature importance methods rely on their own assumptions to approximate this [Bordt et al., 2022, Sundararajan and Najmi, 2020]. For counterfactual explanations, this issue also exists as the optimization problem to create the explanations can be set up in different ways. Even a single counterfactual explanation method could lead to a large number of explanations, as the choice of parameters (such as the distance metric) has an influence on the explanations that are returned first [Goethals et al., 2023b]. The diversity of multiple counterfactual explanations, generated by the same counterfactual algorithm is also known as the Rashomon effect [Molnar, 2020].[1]

Other authors already showed the level of disagreement between different post-hoc explanation techniques: Roy et al. [2022] show disagreement between LIME and SHAP explanations, Brughmans et al. [2023b] illustrate this for different counterfactual explanation algorithms, and Bordt et al. [2022] demonstrate the disagreement between SHAP, LIME, and counterfactual explanations. We illustrate the disagreement problem between different counterfactual explanation algorithms for one specific instance with an example in Table 5.1, in line with Brughmans et al. [2023b]. This table demonstrates the disagreement problem for one instance from the Australian credit dataset, where the target variable indicates whether a person should be granted a loan or not. The depicted instance was not awarded credit and asks for a counterfactual explanation to know which features to change to receive a positive credit decision. Table 5.1 shows the explanations returned by 10 different counterfactual algorithms, which vary widely. [2] This example illustrates that every feature can be included in the explanation by switching between explanation algorithms. Brughmans et al. [2023b] verify this for multiple datasets and classifiers, and establish the feasibility of both including and excluding specific features across different scenarios. Note that the potential for manipulation of explanations extends beyond switching between different counterfactual explanation algorithms. In Section 5.4, alternative strategies that can be employed for manipulation are explored. Currently, a consensus on how to resolve this ambiguity has not yet been reached. Research indicates that most developers rely on arbitrary heuristics, such as personal preferences, to choose the final explanation [Krishna et al., 2022].

This plurality is not necessarily a bad thing. Bordt et al. [2022] distinguish between a cooperative and an adversarial context. In cooperative contexts, where stakeholders have the same goal, this plurality can be beneficial as it is expected that the explanation provider will choose the explanation that is in both parties' best interest. For example, when data scientists are debugging a model for their own company, this plurality of explanations can be useful. Other researchers suggest combining multiple explanation techniques to provide a more accurate *meta* explanation [Mollas et al., 2023]. However, in adversarial contexts, the interests of the explanation provider and the data subject are not necessarily aligned, and the explanation providers will be incentivized to choose the explanation that best fits their own interests. An example of such an adversarial context is a loan application where the customer was denied the loan and wants to flag the decision as being discriminatory [Bordt et al., 2022]. In this case, the bank might want to conceal this discriminatory practice by returning a different explanation. This phenomenon is known as *fairwashing*, and has received significant

[1] The Rashomon effect means that an event can be explained by multiple causes, and is named after a Japanese movie that tells multiple (contradictory) stories about the death of a samurai [Molnar, 2020].

[2] The counterfactual algorithm GeCo was not able to find a counterfactual explanation for the given instance.

attention [Aïvodji et al., 2019]. While fairwashing is the most extensively studied objective, we will explore additional scenarios for misuse in adversarial contexts in Section 5.5. However, even in adversarial contexts, this plurality can be used in a positive way. For example, Bove et al. [2023] do mention that in settings such as loan applications, the plurality of explanations can benefit the user if they are provided with multiple explanations.

5.4 manipulation strategies: how can explanation providers exploit the disagreement problem?

The manipulation of explanations by explanation providers is not limited to the mentioned example of switching between explanation algorithms, but can occur at various stages throughout the pipeline, as depicted in Figure 5.1. We specifically focus on the manipulation that takes place in the post-processing stage, where the explanations are generated, as we imagine that the explanation provider may not always possess the authority to modify the machine learning model or underlying data (the explanation provider is not necessarily the same entity as the model owner). Nevertheless, it is important to note that manipulations directly to the data or model are still feasible, and we discuss some relevant literature exploring this below.

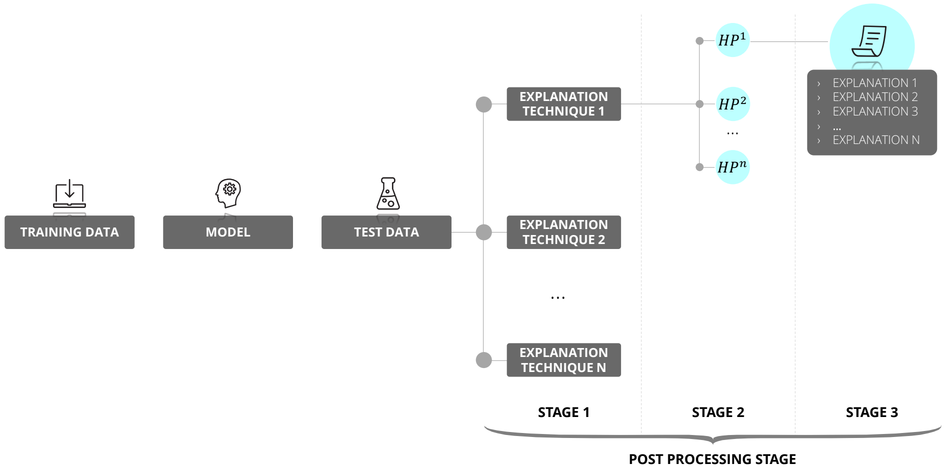


Figure 5.1: Strategies the explanation providers could deploy to manipulate the explanations

Manipulating the training data to result in different explanations, is related to the area of *data poisoning attacks*. Data poisoning attacks usually involve injecting manipulated data into the training set to compromise the performance of the machine learning model, and while the main focus in literature is on model behavior, its goal might also be manipulating the explanations. Baniecki et al. [2023] illustrate that it is possible to attack Partial Dependence plots by poisoning the training data. Bordt et al. [2022] highlight the important role of the reference dataset, and show how changing this set influences the resulting SHAP explanations. With regard to changing the model, Slack et al. [2020] demonstrate the possibility of modifying biased classifiers in such a way that they continue to yield biased predictions, while the explanations generated by LIME and SHAP will appear harmless. Other authors show the possibility of fine-tuning a neural network to conceal discrimination in the model explanations [Dimanov et al., 2020, Heo et al., 2019]. Finally, in the domain of images, Dombrowski et al. [2019] present evidence showcasing the manipulation of explanations through the application of nearly imperceptible perturbations to visual inputs. In this case, the test data, for which the prediction needs to be explained, is altered. These perturbations would not change the output of the machine learning model, but could result in drastic changes in the explanation map. [3] Additionally, Slack et al. [2021] focus on modifying both the model and the test data, such that slight perturbations to the input data can lead to more cost-effective recourse for specific subgroups, while giving the impression of fairness to auditors.

As mentioned, we focus on strategies to alter the explanation in the post-processing stage, without making any alterations to the used data or the underlying machine learning model. We foresee three main strategies the providers could deploy in this stage:

1. **Change the explanation technique**
   Many different post-hoc explanation techniques exist, both local and global, as outlined in Section 5.2. Consequently, a first evident strategy entails switching to a different explanation technique. For example, when the surrogate model reveals patterns the explanation provider wants to conceal, he might switch to using Partial Dependence plots as an alternative if these patterns do not manifest clearly in those plots. However, on a local level, using different explanation techniques between instances may attract greater attention than the strategies described below, as the output could have a significantly different format (e.g., feature importance plot versus a counterfactual explanation).

2. **Change the parameters or used implementation of an explanation technique**
   Even within a single explanation algorithm, significant leeway exists for manipulating the explanations, contingent upon the selected parameter configurations. For example, LIME explanations depend on the number of perturbed instances and the bandwidth [Bordt et al., 2022, Garreau and Luxburg, 2020], while for Shapley values, there is a multitude of ways to implement them and each operationalization yields significantly different results [Sundararajan and Najmi, 2020]. Global methods, such as surrogate modeling, are heavily influenced by the choice of architectural design (e.g., linear models, decision trees, etc.) and the complexity of the surrogate model. In the case of counterfactual explanations, as shown in Table 5.1, the used implementation exerts a substantial influence on the returned explanations, with the number of potential implementations proliferating at a rapid pace. Additionally, even within one counterfactual algorithm, there often exist many modifiable parameters that influence the results.

3. **Exploit the non-deterministic component of some explanation algorithms**
   Some explanation algorithms such as DICE [Mothilal et al., 2020] inherently provide multiple possible explanations for one instance. In such cases, the explanation provider can simply select an explanation from the available options without requiring any modifications. Furthermore, certain explanation algorithms are not designed in a

[3] One could argue that altering the test data in an imperceptible way will be mostly applicable to image data, as in tabular data these changes may be more noticeable.

deterministic way and may return different explanations across runs. For example, when using LIME, the randomness introduced during the sampling and perturbation process can lead to variations in the generated explanations for each execution [Lee et al., 2019, Zhang et al., 2019]. Additionally, de Oliveira and Martens [2021] show that multiple counterfactual algorithms do not generate consistent results over multiple runs, when the same model, input data and parameters are used. In this scenario, the explanation providers can repeatedly execute the explanation algorithm with the same parameters until an explanation that aligns with their preferences is returned.

In the scenario we describe, we assume explanation providers deliberately choose the explanation out of all the possible explanations that best aligns with their interests. The returned explanation will still be technically correct, it will just not necessarily be the explanation that will be in the best interest of the user. It is important to note that we are not referring to situations where explanations chosen by the explanation provider are not in the best interest of the user 'by accident' due to differences in knowledge background or a lack of awareness of the user's preferences [Bove et al., 2023, Gilpin et al., 2022]. Instead, we are concerned with cases where the explanation provider knowingly opts for an explanation that serves their own agenda, despite knowing that it may not be the optimal explanation for the end user. Note that in described strategies, the providers maintain a partial ethical stance by delivering explanations that retain technical correctness. However, providers have the potential to further exploit the situation by offering spam explanations, containing superfluous features [Greene et al., 2023], or by deliberately presenting entirely false explanations that are fabricated. The complexity of the pipeline depicted in Figure 5.1 demonstrates the extensive potential for manipulation and, consequently, the fragility of explanations.

5.5 manipulation objectives: why would explanation providers want to exploit the disagreement problem?

Which objectives could the providers have to engage in this behavior? We outline them in Figure 5.2, and discuss various scenarios for each objective in the subsections below. At the moment, we see mitigating liability, implementing their beliefs and maximizing their profits as the main objectives. This list may not be exhaustive yet as the way that technology is used in society is constantly evolving and new objectives may emerge.

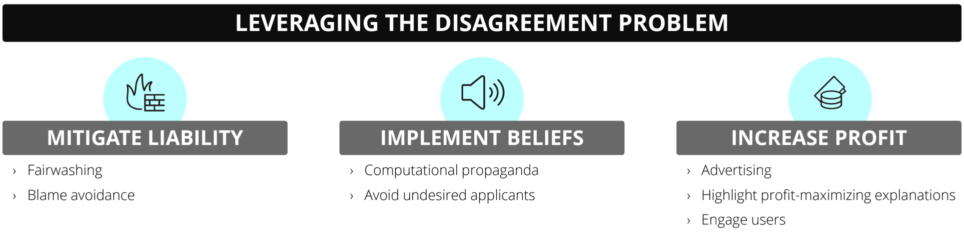

LEVERAGING THE DISAGREEMENT PROBLEM
MITIGATE LIABILITY
› Fairwashing
› Blame avoidance
IMPLEMENT BELIEFS
› Computational propaganda
› Avoid undesired applicants
INCREASE PROFIT
› Advertising
› Highlight profit-maximizing explanations
› Engage users

Figure 5.2: Main objectives to leverage the disagreement problem

5.5.1 *Mitigate liability*

The model could be unethical or suboptimal in several ways and model explanations could reveal this. Explanation providers could manipulate the explanations to avoid these issues coming to light.

5.5.1.1 *Fairwashing*

The first, and most studied, reason for explanation providers to engage in this behavior, is *fairwashing* [Aïvodji et al., 2019, 2021, Shahin Shamsabadi et al., 2022]. Fairwashing is defined as '*promoting the false perception that a machine learning model used by the company is fair while this might not be so*' [Aïvodji et al., 2019]. In a fairwashing attack, the explanation provider will manipulate the explanations to under report the unfairness of the machine learning model. This has a significant impact on the individuals that received a negative decision based on unfair grounds, as this will deprive them of the possibility to contest it [Aïvodji et al., 2021]. The relative easiness with which fairwashing can be executed has already been shown in the literature. [Aïvodji et al., 2021, Shahin Shamsabadi et al., 2022]. Imagine a bank that decides it prefers people from a certain demographic group, and predominantly gives out loans to this group (without a justified reason to do so). It could easily mask this behavior by choosing a different explanation. For example, instead of returning the explanation '*If you would have belonged to a different demographic group, you would have received the loan*', it could return as explanation '*If your income would be double as high, you would have received the loan*', even if the latter explanation is less plausible. Some counterfactual algorithms such as DICE [Mothilal et al., 2020] even have as an input parameter the features that can be part of the explanation, so if sensitive features such as demographic attributes are removed from this list, counterfactual explanations will never flag discrimination. We use counterfactual explanations as illustration here, but this objective extends to other explanation techniques as well. All the mentioned techniques in Section 5.2 have the potential to reveal bias within a model (for example a feature importance ranking where the sensitive attribute has a very high score). This misleading practice undermines the core principles of algorithmic fairness and hampers efforts towards achieving equitable and just outcomes.

5.5.1.2 *Blame avoidance*

Explanation providers can also take advantage of the plurality of explanations to shift blame or evade responsibility for controversial or erroneous decisions made by Artificial Intelligence (AI) systems. Nissenbaum [1996] already mention that placing accountability in a computerized system can be a very obscure process due to the '*problem of many hands*' (many actors and factors contribute to the process, and is not clear which factor ultimately led to the decision). This issue is reflected in the explanations, where different explanations can point to different actors or circumstances. For example, in the case of autonomous vehicles, AI systems make critical decisions that impact passenger safety. Malicious model owners, such as manufacturers or operators, may downplay system failures or accidents caused by their vehicles. They could selectively present an explanation that attributes the fault to external factors or human error, and as such divert attention from potential design flaws or inadequate safety measures. Similarly, in the field of healthcare, this exploitative behavior can manifest when mistakes by surgeons or flaws in operating machines are concealed to avoid accountability. Explanation providers, which could include medical professionals, institutions, or even the manufacturers of medical devices, may withhold or manipulate explanations to protect their reputations or evade legal consequences. Such practices can have severe consequences, as critical flaws in life-critical systems may go unnoticed, posing a threat to the safety and well-being of future users. These actions not only endanger lives but also run contrary to our ethical values. Placing the entire blame on parties that are only partially responsible for an incident contradicts the principles of fairness and accountability. The appropriate distribution of responsibility is crucial for ensuring that the errors are properly addressed and the necessary improvements are made.

5.5.2 *Implement beliefs*

Explanation providers may use the explanations to promote their belief system, either by influencing people through propaganda or by excluding applicants that they deem unworthy, despite the machine learning model not sharing this perspective.

5.5.2.1 *Computational propaganda*

The power to choose an explanation that best fits its interest, can be used to exert an influence on the public opinion. Propaganda itself is defined as '*the expression of opinion or action by individuals or groups deliberately designed to influence opinion or actions of other individuals or groups with reference to predetermined ends*', while computational propaganda is defined as '*propaganda created or disseminated using computational (technical) means*' [Martino et al., 2020]. Note that propaganda does not necessarily have to lie; it could simply cherry-pick the facts, which is exactly the option explanation providers have to their disposal [Martino et al., 2020]. By selectively presenting explanations that align with their preferred ideology or desired narrative, explanation providers can amplify certain perspectives while downplaying or ignoring others. For example, in the realm of political campaigns, AI systems are used to analyze public sentiment, create targeted messaging, and influence voter behavior. Imagine an entity with access to an AI model that predicts the likelihood of successful integration for immigrants based on various factors like employment, language proficiency, and government support. The entity firmly believes in the principle of stricter requirements for immigrants, and they could selectively highlight specific factors such as language proficiency or employment history, while downplaying or omitting other important factors such as government support and community involvement. By presenting the AI model's predictions as mainly being driven by these selected factors, they could frame the narrative that successful integration is mainly due to language proficiency, and engaging in employment. The goal is to shape public opinion regarding immigration policy and generate support for stricter language and employment requirements for immigrants. Evidently, machine learning models cannot perfectly mimic the actual world, so even if a machine learning model could be perfectly explained, such an explanation would not constitute a perfect explanation of the real world. However, the concern here lies in the fact that people may still perceive machine-generated explanations as accurate depictions of the actual world, and consequently, the cherry-picked explanations have the potential to influence and shape their understanding of the world at large. Additionally, if the power to generate the explanations would be in the hands of a few actors, they would have the potential to wield significant influence over a large number of people. In this context, the manipulation of explanations can have far-reaching consequences for public opinion and democratic decision-making, and could promote the spread of misinformation.

5.5.2.2 *Avoid undesired applicants*

In this scenario, the explanation provider, who is using a machine learning model, has the ability to engage in discriminatory practices without directly manipulating the model itself. Instead, they alter the quality of the explanations given to certain population groups, thereby introducing discrimination. In algorithmic decision-making, explanations are often provided to users (the explanation recipients) to help them understand the factors that influenced the decision and potentially take corrective actions (*algorithmic recourse*). Counterfactual explanations are most often used here, as they guide users in modifying their input data to achieve a desired outcome.

In this case, the explanation provider treats different population groups unequally by manipulating the quality of the explanations provided to them. The preferred population group is given explanations that are concise, actionable, and easily implementable. For example, they might receive suggestions such as adjusting the loan amount slightly or making small changes to their reported income. These explanations empower the preferred group to take specific actions that could potentially improve their chances of receiving a positive outcome. On the other hand, the disadvantaged demographic group is given explanations of lower quality. These explanations are designed to be difficult or even impossible to act on. They might involve suggesting large changes to their income or modifying their age, which are factors that applicants typically have limited or no control over. By providing such explanations, the explanation provider creates a significant imbalance in the recourse options available to different society groups. These population groups are not solely confined to traditionally protected characteristics such as race or gender. They can extend to any characteristic that the explanation provider deems undesirable. For example, in the hiring domain, the hiring company (and explanation provider) may deliberately offer lower-quality explanations to older individuals or individuals with certain health conditions, as they perceive them as less desirable for future employment. For some cases, this could also lead to an increase in profit which shows that the multiple objectives can be pursued in parallel and may not always require mutual exclusion. Note that the discriminatory practices described in this scenario are not related to the machine learning model itself, but to the post-processing stage where explanations are generated and shared with applicants. This issue is related to fairness in algorithmic recourse, where fairness is assessed by measuring the distance between the factual and the counterfactual instance [von Kügelgen et al., 2022, Sharma et al., 2020], and highlights the need for fairness assessments not only during the modeling stage but throughout the entire decision-making pipeline, including the provision of explanations.

5.5.3 *Increase profit*

Explanation providers might feel incentivized to capitalize on the explanations. They could return the explanation that would be the most profitable for them, and for this we envisage several scenarios.

5.5.3.1 *Advertising*

One possibility discussed in previous work, is the integration of algorithmic explanations with advertising opportunities, creating an '*explanation platform*' where advertisements are served alongside the explanation [Greene et al., 2023]. An example of this could be, that during a job application you receive the following explanation: '*If your CV would have included Python, you would have been invited for the next round*'. This explanation would then be accompanied by an advertisement for an online Python course, which would be a convenient solution for users to reach their goal [Greene et al., 2023]. This approach allows the explanation provider to select the explanations that have the potential to generate the highest revenue in the advertising market.

5.5.3.2 *Highlight profit-maximizing explanations*

However, monetization avenues can go beyond advertising. Explanation providers can also exploit the plurality of explanations to direct users towards actions that would maximize their own profits directly. This is related to the advertising scenario, but in this case the actions of the decision subject would directly lead to an increase in profit for the provider. For example, in the domain of healthcare diagnostics, AI systems are increasingly used for the identification of diseases and treatment recommendations. Malicious explanation providers, such as healthcare providers or insurance companies, may strategically choose explanations that prioritize certain measures or specific treatments. In this context, the goals of healthcare providers and insurance companies may diverge. Healthcare providers may have incentives to promote more expensive treatments, while insurance companies may prefer cost-saving measures and cheaper treatment options. However, by favoring explanations that are not necessarily the best or most appropriate, these providers can exert influence over medical decisions and potentially compromise patient care. This scenario could also happen in other domains than healthcare: for example, in the realm of credit scoring, AI systems are employed to evaluate an individual's creditworthiness. Barocas et al. [2020] already mention that decisions (and therefore explanations) in this scenario are not simply binary. The provider gives the decision subject a counterfactual that results in a *specific* interest rate, and as such it can choose the interest rate that is likely to maximize its profit [Barocas et al., 2020].

5.5.3.3 *Engage users*

In line with *Computational Propaganda*, discussed in Section 5.5.2.1, providers could also choose to return the explanations that reinforce the ideologies of the data subject itself. In this case, the explanation provider would be a platform, and the goal would be to maximize the revenues of the platform by keeping users as engaged and satisfied as possible (for many platforms daily/monthly active users is an important objective in their financial reports). An example of an explanation in this case, could be the same as in the scenario of propaganda, but in this case different society groups would receive very different explanations, depending on their beliefs. It is known that presenting them with content and information that is likely to resonate with their interests is a way to achieve this (in line with filter bubbles in content recommendation systems). However, this could lead to different groups in society receiving vastly different explanations for the same phenomenon, and consequently to *epistemic fragmentation* [Milano et al., 2021].[4] . By reinforcing filter bubbles and echo chambers, these platforms exacerbate polarization and hinder constructive dialogue between different groups in society.

Introducing a profit motive into the generation of explanations at all seems contradictory to the initial goals of Explainable AI. An explanation recipient should not have to wonder whether the selected explanation was chosen for its profit-making potential rather than for its ability to accurately explain the situation [Greene et al., 2023].

5.6 DISCUSSION

The examples discussed in Section 5.5 shed light on potential ethical concerns, even though they may not necessarily involve illegal activities. In these scenarios, the generated explanations remain factually correct but are selectively hand-picked by the explanation provider to serve their own interests. At the moment, this process is completely unregulated, but could have very serious consequences, as outlined in the scenarios above. In scenarios listed in Section 5.5, we assumed the explanation providers had malicious incentives, but obviously, this will not always be the case. In fact, some providers may be motivated to manipulate the explanations for the social good. For example, they might explicitly avoid providing explanations that reinforce biased stereotypes, in an attempt to promote fairness and equity. Nevertheless, even though their motives might be aligned with societal goals, it remains questionable whether unregulated entities without the required authority should have the power to make this call.

As we are at the forefront of the XAI revolution, it is crucial to address this issue now, before these methodologies are implemented on an even wider scale. Currently, a substantial portion of AI power is concentrated among a few tech giants. If we also grant them the authority to control the explanations generated by AI models, they would possess yet another means to exert significant influence over society. To mitigate this concentration and potential misuse of power, it becomes imperative for government institutions to collaborate and establish agreed-upon standards and tools for XAI. In particular, in adversarial contexts where interests

[4] Epistemic fragmentation refers to the tendency for different people to have different sources of knowledge and different, often conflicting, understandings

may clash, it should not be left solely to the explanation providers to create and choose the explanations. Instead, we argue that governments and policy makers should take the matter into their own hands, and agree on a framework that should be used as soon as possible. The key question here is "*What should be the process to make this decision, and what tools are needed to support this process?*". Similar to the no free lunch theorem, that indicates that there is no algorithm that always outperforms all others, there likely will also not be an universally superior explanation method. An agreement on which method to use in which scenario should be established, and this should be done democratically by allowing those affected by XAI to voice their opinion [Kuźba and Biecek, 2020, Vermeire et al., 2022b], in line with the 'democratic principles of affected interests' [Fung and Wright, 2001]. Another line of research has also investigated the possibility of grouping the explanations [Carrizosa et al., 2024a] which could also counter the issue of having multiple explanations.

To address the disagreement problem effectively, it is essential to first establish a clear and precise definition of what constitutes an explanation Different objective functions naturally lead to varying explanations, so first one should agree on what a desirable explanation should entail. However, this does not constitute the entire problem, because even with a clear definition, sources of randomness can lead to multiple explanations. We also want to emphasize again that in many cases multiple explanations is not necessarily a bad thing. In many cases, these explanations do not necessarily conflict but rather provide complementary perspectives. The real issue arises when providers operate under the assumption that there is only one true explanation.

It will take some time to reach a global consensus on the procedures that should be used, and therefore as a short-term solution, regulation should demand full **transparency** in the used explainability method, and settings. This would remove some flexibility for the explanation provider to change the explanation technique continuously, but not remove all potential for manipulation as the providers could still exploit the non-deterministic component of some explanation algorithms or simply lie about the used parameters. Therefore, to ensure adherence to ethical values, we also foresee that it would be mandatory to have external auditors conducting audits of AI systems, explanations, and decision-making processes. These auditors should be independent entities without a vested interest in the outcomes, similar to how audits are conducted in other industries. We argue that it would be better to create good ways to detect whether someone is gaming the system than to create yet another explanation method. Furthermore, in high-stakes contexts, where transparency is of paramount importance, we argue that the the use of white-box models needs more attention [Goethals et al., 2022], given the manipulation risks surrounding explanations. To conclude, we believe that implementing these measures can ensure that AI systems are developed and deployed in a manner that aligns with societal values, and foster a more transparent and ethical XAI ecosystem.

Part III

FAIRNESS

6 Predictive Counterfactual Fairness

This study investigates how counterfactual explanations can be used to assess the fairness of a model. Using machine learning for high-stakes decisions is a threat to fairness as these models can amplify bias present in the dataset, and there is no consensus on a universal metric to detect this. The appropriate metric and method to tackle the bias in a dataset will be case-dependent, and it requires insight into the nature of the bias first. We aim to provide this insight by integrating explainable AI (XAI) research with the fairness domain. More specifically, apart from being able to use (Predictive) Counterfactual Explanations to detect *explicit bias* when the model is directly using the sensitive attribute, we show that it can also be used to detect *implicit bias* when the model does not use the sensitive attribute directly but does use other correlated attributes leading to a substantial disadvantage for a protected group. We call this metric *PreCoF*, or Predictive Counterfactual Fairness. Our experimental results show that our metric succeeds in detecting occurrences of *implicit bias* in the model by assessing which attributes are more present in the explanations of the protected group compared to the privileged group. These results could help policymakers decide on whether this discrimination is *justified* or not.

6.1 INTRODUCTION

As the influence and scope of decisions made by AI models is increasing, there are growing concerns that the models making these decisions might unintentionally encode and even amplify human bias [Corbett-Davies et al., 2023]. This is why it is of huge importance to understand the decisions models are making and to ensure they are fair. We focus on fairness in classification, where the goal is to prevent discrimination against people based on their membership of a sensitive group, without compromising the utility of the classifier [Caton and Haas, 2020, Dwork et al., 2012].

Different automatic methods to deal with discrimination, however, make different implicit assumptions about the nature of bias in the data and the right method to apply will be case-dependent and often policy-related [Wachter et al., 2021]). Arguably, the data scientist is not the right person to make this call. The necessity for the involvement of policymakers and legal scholars enlarges the need for an automated, data-driven procedure that can detect and assess the source of automated discrimination in predictive models to support decision making [Wachter et al., 2021]. As other authors already argue [Rudin et al., 2020], it is misguided to focus on fairness while not obtaining transparency first as it is not fair that life-changing decisions would be made without entitlement to an explanation.

In this paper we answer the call for more transparency in the fairness domain [Rudin et al., 2020, Wachter et al., 2021] by linking Explainable AI with fairness, using Counterfactual Explanations. Counterfactual explanations form the basis of an important class of explainable AI methods [Adadi and Berrada, 2018], and a counterfactual explanation of a data instance is defined as the smallest change to the instance so that it ends up with a different classification outcome. We name our metric *PreCoF*, which stands for *Predictive Counterfactual Fairness*. *PreCoF* finds counterfactual explanations for all individuals in each sensitive group by assessing for each of the attributes whether changing it to one of the default values would result in a class change.[1] It identifies the attributes that are proportionally more present in the explanations of the protected group compared to the unprotected (or privileged) group. This term is not to be confused with *counterfactual fairness* as we will explain in Section 6.2.3.1. The goal of *PreCoF* will not be to provide yet another calculation on the output of a decision making system but to shed light on underlying patterns for the discrimination in the model, so that policymakers can decide how to handle this appropriately.

A first example of something our metric is able to detect can be seen in the Adult Income dataset: the attribute *marital status* is the attribute that is proportionally the most present in the explanations of women compared to men. This offers additional insights into the model so that policymakers can decide whether this is a pattern that can be kept in the model or if the model should be modified. The results of the other datasets are also in line with patterns that we know to be present based on literature or through further analysis of the datasets.

It is important to highlight that our metric will make statements about the model but not about the underlying data. We expect them to reflect underlying patterns in the data but it is possible that two different machine learning models trained on the same data will give very different results.

[1] We will clarify the calculation of these default values in the methodology.

6.2 background

We refer to Section 3.3 for a more general background about fairness in machine learning, but we will discuss some aspects in more detail here. As mentioned, legislation is often attempting to achieve fairness by using a 'colorblind' approach that ignores socially-sensitive features, which is misguided to begin with [Johnson, 2021]. The idea here is that you remove the bias from the dataset by removing the discriminatory attributes from it. However, in any sufficiently rich dataset, proxy variables will likely exist that closely correlate with the sensitive attributes [Kim, 2017] so just removing them will not work. Removing all the attributes that are correlated with the sensitive attribute is not a good solution either [Kamiran and Žliobaitė, 2013]; in some cases, all attributes will be correlated with the sensitive attribute, or some of the correlated attributes are too informative to remove (e.g., field of study is correlated with gender but too important to remove in hiring decisions).

We make the distinction between *explicit bias*, when the model involves direct use of the sensitive attribute, and *implicit bias*, when there is a neutral attribute that substantially disadvantages the protected group. These are also called *direct* and *indirect discrimination* respectively. *Indirect discrimination* is arguably the most likely type of discrimination to arise from automated decision making due to the reliance of these system on inference and proxies of target variables and protected attributes [Wachter et al., 2020].

Many scholars see value in judging discrimination with common sense [Doyle, 2007], however, this is often ineffective in cases of *indirect discrimination*, especially when the relation between the protected attribute and the neutral attribute is not straightforward [Wachter et al., 2021]. Intuition might fail us because it cannot be assumed that automated systems will discriminate in ways similar to humans or follow their patterns of discrimination: new and counterintuitive proxies for traditionally protected attributes can emerge but will not necessarily be detected [Wachter et al., 2021]. If such an attribute is found that substantially disadvantages the protected group, this is not necessarily a problem: some attributes can be justified, depending on the context of the case and the relevant legislation. *Justified indirect discrimination* occurs when the 'proportionality test' is passed, meaning that the attribute is both legally necessary and proportionate [Wachter et al., 2020]. *PreCoF* is developed to fit in this mindset: can we find the attributes that explain why some sensitive groups are more often predicted with a negative outcome? This can then lead to a discussion about these attributes being justified or not.

There are three main responses when such a bias is detected: First, one can do nothing and allow the bias to be amplified; second, fix the technical bias but maintain the society status quo and make sure that the machine learning does not make the society more biased which is called a *bias preserving* approach [Wachter et al., 2020]. A third option is what are called *bias transforming* metrics and these aim to actively account for historical inequalities [Wachter et al., 2020]. The adequate response will depend on the situation at hand, but doing nothing will in our opinion never be the right call.

6.2.1 *Fairness metrics*

There is no universal definition of fairness, which greatly complicates our research question. A more complete overview of the fairness field can be found in Section 3.3, but for clarity, we will shortly discussed some of the used concepts again. Some define fairness as *fairness through unawareness* [Pedreshi et al., 2008], which establishes fairness through removing the sensitive attributes from the dataset. However, this is not always possible as sometimes sensitive attributes are needed to make predictions. Even when the sensitive attribute is not directly relevant to the prediction task, correlated variables (e.g., race from zip code in the United States) make such a "blind" approach less efficient to counter discrimination [Fryer Jr et al., 2008]. Other often-used fairness metrics include *individual fairness* [Dwork et al., 2012], which states that similar individuals should be treated similarly, *demographic parity* [Calders et al., 2009] (which is also called *disparate impact* [Feldman et al., 2015] or *statistical parity* [Dwork et al., 2012]) which minimizes the absolute difference in outcome distributions of all groups, *equalized opportunities* [Hardt et al., 2016], which optimizes towards equal positive rate conditional on the target outcome and *equalized odds* [Hardt et al., 2016], which optimizes towards equal positive and negative rate conditional on the target outcome.

Demographic parity, *equalized odds* and *equal opportunity* are all group-based criteria, which are more suited to statistical analysis [Ritov et al., 2017] but can be very unfair from the point of the individual [Dwork et al., 2012]: it provides protection for groups but not for specific individuals in those groups and we tend to care more about protection for individuals [Fleisher, 2021]. It also does not provide protection against phenomena like cherry-picking.[2] Even more problematic, many of the group fairness metrics are mutually incompatible, which means it is impossible to satisfy all of them at the same time [Kleinberg et al., 2016, Verma and Rubin, 2018]. This has as a consequence that the detection of discrimination can be 'gamed' through choosing the right fairness metric [Wachter et al., 2021]. It has been shown that all these metrics suffer from deep statistical limitations and that they can even negatively impact the well-being of the groups they are trying to protect [Corbett-Davies et al., 2023]. *Individual Fairness* is more strict than any group-notion fairness as it imposes a restriction on the decision for each pair of individuals. It also forbids a variety of discriminatory practices like explicit discrimination, implicit discrimination, redlining and tokenism [Fleisher, 2021]. It can also detect cases of discrimination that various group fairness criteria miss like cherry-picking. However it is hard to define a metric function to measure the similarity of two inputs [Fleisher, 2021, Kim et al., 2018]. A last metric is *Counterfactual Fairness* [Kusner et al., 2017], which is more related to our metric and will be discussed in Section 6.2.3.1.

All the metrics that are conditional on the target outcome such as *equalized odds* and *equal opportunity* are *bias preserving*, which means that they will preserve historical biases and just ensure that the machine learning model will not amplify these biases or insert new bias into the system [Wachter et al., 2020]. They share the idea that the bias present in the target labels is meant to be there [Wachter et al., 2020]. *Demographic Parity*, *Individual Fairness* and

[2] Cherry picking refers to members of sensitive groups being randomly chosen, or chosen for malicious reasons as a way to undermine members of those groups [Dwork et al., 2012, Fleisher, 2021]. An example of this in college applications could be when the majority group is carefully screened, and the same number of applicants is randomly selected from the minority group. This is not fair for hard-working members of the minority group that will not get admitted, but would be compatible with a variety of group fairness criteria [Fleisher, 2021].

Counterfactual Fairness are *bias transforming metrics*. *PreCoF* is aimed to be a *bias transforming metric*, but it offers the transparency and flexibility for policy makers to decide this for each situation at hand. Choosing an appropriate metric can have political, legal and ethical implications and should be subject to more consideration and justification than is currently the case [Wachter et al., 2020]. The previously discussed fairness metrics are not well suited to answer normative and legal questions on how the discrimination in the model should be handled and might ultimately prove to be irrelevant in court [Wachter et al., 2021].

6.2.2 *Conditional fairness metrics*

In practice, there often exists a certain set of attributes on which we deem it fair to discriminate [Xu et al., 2020]. An example of such an attribute is the department choice in the Berkeley's graduate admission problem, where there allegedly was a bias against female applicants as they had a lower admission rate then male applicants [Xu et al., 2020, Pearl, 2009]. After conditioning on department choice, this was no longer the case [Pearl, 2009]. Conditional fairness is a more sound fairness metric where the outcome variables should be independent of sensitive variables conditional on these fair attributes [Xu et al., 2020]. There exist various methods to implement conditional fairness such as explainable discrimination [Kamiran et al., 2013, Wachter et al., 2021] or conditional demographic disparity [Wachter et al., 2021]. They have the point of view that some differences in decisions across sensitive groups can be explainable and hence tolerable [Kamiran et al., 2013]. For example, in job applications the education level of a candidate can be such an explainable attribute [Kamiran et al., 2013].

The underlying fairness metrics in explainable discrimination and conditional demographic disparity are a bit different but they are based on the same principle [Kamiran et al., 2013, Wachter et al., 2021]: Kamiran et al. [2013] measure the discrimination as the difference in positive rates between two sensitive groups: the discrimination that remains after subtracting the discrimination that can be explained by using the conditional attribute (*explainable discrimination*) is the *illegal discrimination*. Wachter et al. [2021] define *demographic disparity* as the difference in proportion of people from the protected group with a favorable and an unfavorable outcome. *Conditional demographic disparity (CDD)* follows the same principle but adds a conditional attribute: a decision-making system has no conditional discrimination if, after conditioning on this attribute, the decisions are statistically independent of the sensitive attribute [Wachter et al., 2021]. However, in both methods, it is not clear how the attributes on which conditional fairness is calculated are chosen: searching over all combinations of attributes would be prone to finding false positives [Wachter et al., 2021]. Developers can be inclined to choose favorable conditions [Wachter et al., 2020] and it should not be up to them to choose these variables, but this should be fixed externally by law or domain experts [Kamiran et al., 2013]. The selection of these conditional attributes becomes confusing and debatable as people might not agree about which combinations are reasonable [Kamiran et al., 2013]. Furthermore, the conditional attributes are not necessarily the attributes that are used by the model. In large datasets, conditional variables might exist such that the data can be stratified in groups in such a way that there is no conditional demographic disparity, while that conditional variable is not even a factor used in the model. We will show this in Section 6.4.3.1.

We agree with the point of view that part of the discrimination can be explainable by other attributes, but our goal is to shed light on which attributes are making up the discrimination in the model so that policymakers can decide whether these are justified or not. Is it fair to use GPA in law admissions schools even though it is often biased against ethnic minorities? Is it desired to trade accuracy for fairness in crime recidivism prediction as this can result in a higher crime rate overall? Which biases are socially acceptable and can be maintained? Which actions are appropriate for a specific case? These are all questions that should be answered case by case in an open and transparent debate.

6.2.3 *Related metrics*

6.2.3.1 *Counterfactual fairness*

In recent years, fairness-aware machine learning has been studied from the causal perspective using causal modelling [Pearl et al., 2000]. In line with this research, Kusner et al. [2017] define counterfactual fairness as a notion of fairness derived from Pearl's causal model [Pearl et al., 2000] where for an individual the prediction of the model is considered as fair if it is the same in the real world as it would be if the individual would belong to a different demographic group [Kusner et al., 2017, Wu et al., 2019]. To measure this, they make explicit assumptions about the causal relationships in the data. One way for a predictor to be counterfactually fair is if it is a function of only non-descendants of the sensitive attribute, so this will be different depending on the chosen causal model. The biggest drawbacks of this methodology are that you need to make some untestable assumptions for such a causal model and that it is not scalable [Xu et al., 2020]. It assumes that the causal relations between variables in a dataset are known, while in reality this is not the case. Furthermore, the legal frameworks that govern discrimination in multiple countries do not require a causal relationship with the protected attribute, so counterfactual fairness may fail to identify occurrences of legally actionable discrimination [Black et al., 2020]. Several other authors also propose a causal approach to detect various forms of discrimination in a dataset [Bonchi et al., 2017, Kilbertus et al., 2017] but they suffer from the same drawbacks.

6.2.3.2 *Counterfactual fairness* (bis)

Sokol et al. [2019] already showed how counterfactual explanations can be used to check individual fairness. They consider an instance to be treated unfairly if that instance received the undesirable label and there exists a counterfactual explanation for that instance that includes at least one protected attribute change [Sokol et al., 2019]. We follow this approach when we use counterfactual explanations to identify explicit bias for an individual. On top of that, we also show that aggregating these counterfactual explanations can give more insights about the patterns of explicit bias in the algorithm.

6.2.3.3 *CERTIFAI*

CERTIFAI [Sharma et al., 2020] is a tool that can be applied to any black-box model to assess its fairness. It uses a custom genetic algorithm to generate counterfactuals and examines the explanations to assess the model's fairness, both on an individual and on a group level. The fitness of an individual is defined as the inverse distance between the input instance and its counterfactual. For an individual, if we allow the sensitive attributes to change, and the fitness goes up (distance to the counterfactual becomes smaller: desired outcome is more easily achieved), then the individual could claim the model is treating them unfairly. This tool can also be used to audit fairness on a group level: if the average fitness values of generated counterfactuals are lower for women than for men, this could be used as evidence that the model is not treating women fairly [Sharma et al., 2020]. This tool is different from how we use counterfactual explanations as we will focus on the specific attributes and attribute values that occur in the explanations of both groups and not on the distance to the counterfactual instance.

6.2.3.4 *Fairness in algorithmic recourse*

The literature on algorithmic recourse has focused on finding "an actionable set of changes a person can undertake in order to improve their outcome" [Joshi et al., 2019, Karimi et al., 2021]. Algorithmic recourse poses its own fairness criteria, where the effort to reach the required outcome is taken into account. If individuals from the protected group have to work harder than similar individuals from another group to achieve the desired outcome, then the concept of equal opportunity is violated [von Kügelgen et al., 2022]. This notion of unfairness is not captured by predictive notions and is in line with CERTIFAI, as they both focus on the difference in effort different individuals have to make. To be able to find an 'actionable' set of changes, most authors assume, at least partial, causal knowledge. However, as in Section 6.2.3.1, the reliance on causal information creates practical issues that may limit its applicability [Black et al., 2020]. As we are not necessarily interested in *actionable* counterfactuals, our method will not rely on causal assumptions about the data-generating process. We explain this further in Section 6.4.5.

6.2.3.5 *FlipTest*

FlipTest is a fairness testing approach, that also does not rely on causal information, but instead uses an optimal transport mapping to detect whether a model's behavior is sensitive to changes in the protected status [Black et al., 2020]. Simply changing the protected attribute is not sufficient due to correlations in the data. Therefore, a transport map transports one probability distribution into another, for example women into men, in order to have a pair of inputs with which to query the model. An optimal transport map is used to minimize the sum of distances between a woman and the man she is mapped to (her *counterpart*), where the distance quantifies the difference between them. FlipTest analyzes the cases where the classifiers' output is different between the woman and her *counterpart*, as these are individuals that might be harmed because of their group membership. Like FlipTest, *PreCoF* also aims to shed light on *why* the model is treating a certain subgroup differently but it uses a different method: it does not require to approximate an optimal transport mapping and does not depend on the distance function that is used to construct the mapping.

6.3 COUNTERFACTUAL EXPLANATIONS AS THE SOLUTION

PreCoF aims to explain the discrimination in a predictive classification model, and create transparency regarding which attributes are the most discriminatory between different sensitive groups. This insight can then be used for subsequent discussions and decisions by law or domain experts on which attributes are justified and which attributes will just behave as proxies for the sensitive attribute. An example of Wachter et al. [2021] shows how some attributes can be valid in one case but not in another: when reviewing résumés for a firemen position, height can be deemed a valid discriminator but it seems highly unlikely that this will be the case when reviewing résumés for a CEO position (there it will just serve as a proxy for gender).

We agree with Wachter et al. [2021] that fairness is contextual: it is not possible to create a system that automatically detects and corrects discriminatory models as each case should be handled differently. What is needed is an 'early warning system' that provides transparency in automated discrimination [Wachter et al., 2021] which is what we aim to supply.

As Rudin et al. [2020] also state: it is arguably unfair to have life-changing decisions being made by a system without having any insights into the decisions, which brings us to the field of Explainable AI (XAI). XAI research aims at explaining how an AI system reached its decision [Gohel et al., 2021]. XAI can enhance transparency as well as fairness as it provides explanations that can be understood and as such show bias that is present [Gohel et al., 2021, Sokol and Flach, 2021]. There exist different sorts of explanation procedures for understanding predictive models, both on the global level as on the instance-level. For an overview of these methods, we refer you to Section 3.2.3. We want to assess fairness on the individual level so we will look at instance-based explanation methods, and we will focus on Counterfactual Explanations as they are better suited for our task than LIME or SHAP: the latter explain a prediction score rather than a decision so if we talk about unfair decisions, Counterfactual Explanations are better suited as they focus on the *treatment* an individual received [Fernandez et al., 2020]. We focus on fair decision making, but in the case we want to assess fair scoring, SHAP values can be used in the same set-up. We present the results when using SHAP values instead of counterfactual explanations in Section B.1. Our main argument that more insight is needed in the nature of the bias before deciding on a method to handle it, remains valid for both XAI techniques.

Assume we have a dataset D that consists of n instances and m attributes, where the attribute value of attribute j of an instance i is denoted by x_{ij} with $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., m\}$. The model M will make a decision, which is either a favorable (+, e.g. hired, credit granted) or a unfavorable (-, e.g. not hired, credit rejected) outcome.

$$M(\mathbf{x}_i) \in \{+, -\}$$

A counterfactual $\mathbf{c}$ of a factual instance $\mathbf{x}_i$ is an instance for which:

$$M(\mathbf{x}_i) \neq M(\mathbf{c})$$

and

$$d(\mathbf{x}_i, \mathbf{c}) \text{ is minimal}$$

So the counterfactual is another instance, while the counterfactual *explanation* is the difference between the two: $|\mathbf{c} - \mathbf{x}_i|$. As mentioned in Section 6.2.3, other metrics also use counterfactual explanations to assess fairness. However, our metric will be different as it does not need to assume a causal graph [Kusner et al., 2017], and does not use the distance to the counterfactual like Sharma et al. [2020], but will look at the actual explanations of decisions instead. Furthermore, we will use counterfactual explanations not only to show *explicit bias*, as done by Sokol et al. [2019], but also to get insights into the *implicit bias*, which is arguably the more challenging problem.

An advantage of also looking at *implicit bias* over *explicit bias* is that it deals with rules or patterns of behaviour, and as such can reveal underlying social inequalities and uncover structural unfairness in an algorithm [Wachter et al., 2021]. Direct discrimination is simpler to detect: the action that is alleged to be discriminatory must explicitly refer to a protected characteristic while for indirect discrimination it is more difficult: a neutral attribute or criterion must be shown to substantially disadvantage the protected group, despite not explicitly addressing it [Wachter et al., 2021, Zliobaite, 2015].

6.4 METHODOLOGY

6.4.1 *Materials*

In this study, we focus on tabular datasets, mostly used in fairness-aware machine learning research [Le Quy et al., 2022]. We use datasets from the financial (Adult Income dataset), criminological (Catalonia Juvenile dataset, Crimes and Communities dataset) and the educational (Student performance dataset, Law admission dataset) domain. All the datasets in this study are publicly available.

6.4.1.1 *UCI Adult dataset*

The Adult Income dataset[3], or 'Census Income' dataset contains information extracted from the 1994 census data with as target variable whether the income of a person exceeds $50,000 a year or not. We use it to assess whether there are gender or race inequalities present in people's annual incomes [Asuncion and Newman, 2007]. The Adult dataset contains 48,842 instances with 14 features. As is common in literature, we drop the features *Fnlwgt* as it does not convey a meaning to its values, *EducationNum* as it has the same meaning as *Education* and

[3] https://github.com/EpistasisLab/pmlb/tree/master/datasets/adult

NativeCountry as it has a lot of missing values. We use the features *Age*, *Workclass*, *Education*, *Marital-status*, *Occupation*, *Relationship*, *Race*, *Sex*, *CapitalGain*, *CapitalLoss* and *HoursPerWeek*. The sensitive attributes in this dataset are *Race* and *Sex*. For our experiments we use *Sex* as the protected attribute. The favorable outcome in this dataset is having an income that exceeds $50,000 a year, the unfavorable outcome is having a yearly income below $50,000 .

6.4.1.2 *Catalonian Juvenile Dataset*

This dataset[4] consists of juvenile offenders who were incarcerated in the juvenile justice system of Catalonia and who were released in 2010 [Miron et al., 2021]. Their recidivism behavior was observed between 2010 and 2015. SAVRY is a tool developed in 2003 which predicts recidivism [Miron et al., 2021]. We build a model on most of the individual and criminological variables as in Miron et al. [2021] [5], but we also include the variables that are used in the SAVRY risk scores such as *History of self harm*, *Delinquent peer group*,.. . Our dataset contains 855 instances with 22 attributes. The target variable in this dataset is *Recid*, which is whether the offender has re-offended or not. The favorable outcome here is that there is no recidivism, the unfavorable outcome that there is. The sensitive attributes in this dataset are *Foreigner*, *Sex* and *National Group* of the offenders, but for our experiments we use *Foreigner* as protected attribute.

6.4.1.3 *Crime and communities dataset*

This dataset[6] contains 1994 samples of socio-economic data from the United States. There are 127 attributes in this dataset, but we delete all attributes related to state, race or crime, except for the target variable, so that 91 attributes remain. The target variable is whether the attribute *ViolentCrimesPerPop* is above a certain treshold, which then constitutes a violent community. In line with literature, we also add the attribute *Black* in order to divide the communities in black and non-black communities when the attribute *racepctblack* is above a certain threshold [Kamiran et al., 2013, Le Quy et al., 2022]. The protected attribute here is *Black*.

6.4.1.4 *Student performance dataset*

This dataset[7] consists of 649 students and 30 attributes from a Portuguese high school [Cortez and Silva, 2008]. The attributes of the dataset contain information about the background of the students and their social activities. As commonly done [Hamoud, 2016], we delete the results from the first and the second grade (*G1*,*G2*) as they are very heavily correlated with the final grade (*G3*). The target variable is scoring above average on their final exam of Portuguese,

[4] https://github.com/nkundiushuti/savry/blob/master/dat/reincidenciaJusticiaMenors.csv

[5] https://github.com/nkundiushuti/savry/blob/master/Savry_Fair.ipynb

[6] https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/communities_crime.csv

[7] https://archive.ics.uci.edu/ml/datasets/student+performance

where the favorable outcome is that you score above average and the unfavorable outcome that you score below average. The protected attribute in this dataset is *Sex*.

6.4.1.5 *Law admission dataset*

This dataset[8] contains a Law School Admission Council (LSAC) survey conducted across 163 law schools in the United States in 1991 [Wightman, 1998]. The dataset consists of 20,798 students and the following attributes: *decile1b,decile3b*, *lsat*, *ugpa*, *zfygpa*, *zgpa*, *fulltime*, *fam_inc*, *male*, *tier*, *race* and *pass_bar*. The target variable is whether the student will pass the bar exam or not. The protected attribute in this dataset is *Race*: 92.1% of white students pass the bar exam, while this ratio in non-white students is only 72.3%.

6.4.2 *Explicit bias*

As already highlighted by Sokol et al. [2019], counterfactual explanations can be used to highlight explicit bias in a decision-making model, by searching for explanations that contain the sensitive attribute. We detect *explicit bias* by searching for counterfactual explanations that consist only of the sensitive attribute.

Assume we have a dataset D with sensitive attribute S, where the sensitive value is s, and the non sensitive value is ns. The group with sensitive value s is also called the *protected group* and the group with sensitive value ns is also called the *privileged group* or *unprotected group*. The dataset consists of n instances $\mathbf{x}_i$, with m attributes, where the attribute value of attribute j for instance i is denoted by x_{ij} with $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., m\}$. The index of the sensitive attribute is z. The model M will make a decision, where we denote $+$ as the favorable outcome and $-$ as the unfavorable outcome.

A decision for the factual instance $\mathbf{x}_i$ that has a negative predicted outcome: $M(\mathbf{x}_i) = -$, is deemed to be unfair (*explicit bias*) if there exists a counterfactual instance $\mathbf{c}$, for the instance $\mathbf{x}_i$ that satisfies:

$$x_{iz} \neq c_z \quad \textit{(the counterfactual instance has a different value for the sensitive attribute)}$$

$$x_{ij} = c_j \qquad \forall j \in \{1, 2, ..., m\} \setminus z \quad \textit{(except for the sensitive attribute, the factual and counterfactual instance are identical)}$$

This means that the instance $\mathbf{c}$ that only differs from $\mathbf{x}$ with respect to the sensitive attribute receives a different classification from our prediction model M. An example of such an unfair explanation could be: *"If you would not have been a woman, you would have received the loan."*

This analysis on the individual level could also be aggregated and as such, show patterns in the model. We aggregate the explanations by calculating how many people of each group

[8] https://github.com/tailequy/fairness_dataset/blob/main/experiments/data/law_school_clean.csv

receive such an explanation. How many negatively predicted persons from each sensitive group would have received a positive outcome, simply by changing their sensitive attribute? Which categories of the sensitive group experience *explicit bias* the most?

Machine learning models can also suffer from *fairness gerrymandering*; when there are different sensitive groups, the classifier can be fair for each individual group but can discriminate against structured subgroups [Kearns et al., 2018]. Imagine we have two sensitive attributes: *race* and *gender*. When analyzing the explicit bias in the model, it is possible that no explanations are found with gender or race, but only with a combination of the two attributes (e.g., *"If you would not have been a black woman, you would have received the loan."*). Our method can take this into account by searching for all explanations that contain a combination of the sensitive attributes.

6.4.3 *Implicit bias*

We will use the same terminology as in Section 6.4.2, but now we will remove the sensitive attribute from the dataset before training the model. We will name this new dataset D'. This dataset will consist of n instances $\mathbf{x'}_i$ with $m-1$ attributes, where the attribute value of attribute j for instance i is denoted by x'_{ij} with $i \in \{1, 2, ..., n\}$, $j \in \{1, 2, ..., m\} \setminus z$. We also have access to the original dataset D, where the sensitive attribute for each instance is still available under index z.

What our metric aims to measure, is how much more often a certain attribute is responsible (part of the counterfactual explanation) for a negative outcome decision for members of the protected group, compared to members of the privileged group. So if changing height from short to tall is 100 times part of the counterfactual explanation for a non-hire decision for 100 women ('if your height would have been tall instead of short, you would have been hired') (100%), and only 10 times of the counterfactual explanation for a non-hire decision for 100 men (10%), *PreCoF* will output 90% for the attribute height. We then show the features (and feature values) with the highest *PreCoF*.[9]

More formally, we test for every instance $\mathbf{x}_i$ with an unfavorable outcome for every attribute j whether changing them to one of the default values results in a counterfactual explanation E. We use a set of default values as we will not test every possible attribute value: for numerical attributes or very sparse categorical attributes, this will not be feasible. We select a set of default values, which for numerical attributes are the values of each decile. For categorical attributes, we take the most frequent (max 10) values that are at least present in 1 percent of the training set. If no attribute value is present in more than 1 percent of the training set, we will just take the 10 most occurring values.

Afterwards, we look at all negatively affected members of the protected group, and see how relatively often we can find a counterfactual explanation that consists only of attribute j. This relative number, we call $CoF(j, s)$. Similarly, we measure how often this attribute is part of the

[9] Like explained in Section 6.4.2, the protected group can also be a combination of multiple sensitive attributes. *PreCoF* can take this into account by comparing the explanations of this subgroup (e.g., black women) with the rest of the population.

explanation for the privileged members with negative outcome: $CoF(j,ns)$. Our final metric $PreCoF(j)$ simply calculates the difference between these two.

The mathematical definition for *PreCoF* is thus as follows (where the counterfactual explanation that leads to counterfactual instance **c** can only consist of a single attribute j [10]):

$$\begin{aligned}
CoF(j,s) =& \frac{\mid \{i \mid \exists \mathbf{c} : x_{iz} = s, M(\mathbf{x'}_i) = -, M(\mathbf{c}) = +\mathbf{c}, x'_{ij} \neq c_j, \forall h \in \{1,2,...,m\} \setminus [j,z] : x'_{ih} = c_h\} \mid}{\mid \{i \mid x_{iz} = s, M(\mathbf{x'}_i) = -\} \mid} \\
CoF(j,ns) =& \frac{\mid \{i \mid \exists \mathbf{c} : x_{iz} = ns, M(\mathbf{x'}_i) = -, M(\mathbf{c}) = +\mathbf{c}, x'_{ij} \neq c_j, \forall h \in \{1,2,...,m\} \setminus [j,z] : x'_{ih} = c_h\} \mid}{\mid \{i \mid x_{iz} = ns, M(\mathbf{x'}_i) = -\} \mid} \\
PreCoF(j) =& CoF(j,s) \ - \ CoF(j,ns) \\
PreCoF_1 =& \text{ Attribute } j \text{ such that j} \in \underset{\forall j \in \{1,2,...,m\} \setminus z}{\text{argmax}} \ PreCoF(j)
\end{aligned}$$

Our metric also allows us to look at the exact feature values of the factual and counterfactual instances. A difference here is that we only compare the categorical values as the numerical values are often too sparse to give us insights about the patterns in values. We define $PreCoF_f$ and $PreCoF_c$:

These are calculated in the same way as *CoF*, but CoF_f will output how often each attribute value is present as part of the factual instance and CoF_c will output how often each attribute value is present as part of the counterfactual instance. $PreCoF_f$ and $PreCoF_c$ again calculate the difference for CoF_f and CoF_c between the protected and the privileged group, and $PreCoF_{f1}$ and $PreCoF_{c1}$ will be the attribute values for which respectively $PreCoF_f$ and $PreCoF_c$ are maximal out of all possible attribute values.

By also looking at the specific feature values in the factual and counterfactual instances, we can get more insights into the social patterns in the model. Examples of this can be seen in the results in Sections 6.5.1, 6.5.2, and 6.5.4. Our metric is thus able to give us insights into the implicit bias of a prediction model, without the prediction model even having access to the sensitive attribute.

6.4.3.1 *Toy example*

We will illustrate the use of this metric with a simple toy example.

A machine learning model is trained on this toy dataset in Table 6.1 after removing the sensitive attribute (gender). Assume the following simple rule-based model:

[10] More formally, a counterfactual explanation e that only consists of attribute j means that the counterfactual explanation c satisfies:

$$\begin{aligned}
& M(\mathbf{c}) = + \\
& x'_{ij} \neq c_j \\
& \forall h \in \{1,2,...,m\} \setminus [j,z] : x'_{ih} = c_h
\end{aligned}$$

Row	Gender	School	Hobby	IQ	True Grade	Predicted Grade
1	*M*	*School1*	*Basket*	*High*	*Pass*	*Pass*
2	*M*	*School1*	*Football*	*High*	*Pass*	*Pass*
3	*M*	*School1*	*Football*	*Low*	*Fail*	*Fail*
4	*M*	*School2*	*Football*	*High*	*Fail*	*Fail*
5	*M*	*School2*	*Basket*	*Low*	*Fail*	*Fail*
6	*F*	*School2*	*Dance*	*High*	*Fail*	*Fail*
7	*F*	*School2*	*Dance*	*High*	*Fail*	*Fail*
8	*F*	*School2*	*Music*	*High*	*Pass*	*Fail*
9	*F*	*School2*	*Dance*	*High*	*Pass*	*Fail*
10	*F*	*School1*	*Music*	*High*	*Pass*	*Pass*

Table 6.1: A toy example

If School = School2 or IQ = low, predict Fail; else predict Pass

The predicted outcome by this model can be seen in the last column of the table. This model scores an accuracy of 80 % but predicts more girls to fail than boys, even though in the dataset there are less girls that fail than boys.

We calculate the demographic disparity of our simple rule-based classifier:

$$\text{Demographic disparity} = P(\hat{y} = + \mid M) - P(\hat{y} = + \mid F) = 2/5 - 1/5 = 1/5$$

This metric just tells us that there is a difference in predicted outcome between boys and girls, but tells us nothing about *why* discrimination occurs and gives policymakers no clues on how to handle this. If the reason for this difference in predicted outcome is that the rejected girls have on average a lower IQ, and this is used by the model to predict that they will fail more often, then this could be a *justified* reason for a difference in positive rate, while for other attributes this will not be the case. This shows that group fairness metrics in general are not well suited to answer legal or normative questions as they will not provide any reasoning behind the metric.

In this small example, inspired by the Student Performance dataset, it is straightforward to see which attribute is inducing this bias. The model has learned that *School2* is associated with bad grades which disproportionally affects the female students. We will use this toy example to show that the *PreCoF* metric is able to detect this variable and as such give insights into why the discrimination occurred.

When using the *PreCoF* metric, we get the following results:

$$\begin{array}{rclrclrcl} CoF(School, F) & = & 4/4, & CoF(IQ, F) & = & 0, & CoF(Hobby, F) & = & 0, \\ CoF(School, M) & = & 1/3, & CoF(IQ, M) & = & 1/3, & CoF(Hobby, M) & = & 0 \end{array}$$

We calculate the attribute for which the difference between the protected (F) and the privileged group (M) is the largest:

$$PreCoF_1 = School \qquad (CoF(School, F) - CoF(School, M) = 2/3, \\ \text{which is larger than 1/3 and 0)}$$

We then use the *PreCoF* metric to also detect the feature values causing the differences:

$$\begin{array}{rcl} PreCoF_{1f} & = & School2 \\ PreCoF_{1c} & = & School1 \end{array}$$

PreCoF$_1$ will be *School* as this is the attribute that is proportionally the most present in the explanations of the protected group (girls), compared to the privileged group (boys). As will be discussed in Section 6.5.4, this will also be the case in the real dataset and could have implications in various areas such as college admissions, where girls could be incorrectly rejected because of the school they went to.

This toy example also shows that this metric will not necessarily point to the variables that are the most correlated with the sensitive attribute. Hobby is the most correlated with gender here, but it will not come out of the *PreCoF* metric as the model is not using this variable.

This toy example also allows us to highlight the difference of our metric with conditional fairness metrics; we show the difference by using the formulas of discrimination of Kamiran et al. [2013]. For an explainable attribute E, which could in theory be any attribute from the dataset, Kamiran et al. [2013] consider dividing the database according to the possible values $e_1, \ldots, e_k$ of E. For each of the values e_i they compute a theoretical probability $P^*(\hat{y} = + \mid e_i)$ of being in the positive class by taking the mean $\frac{P(\hat{y}=+ \mid e_i, s) + P(\hat{y}=+ \mid e_i, ns)}{2}$, assuming that if this probability of being in the positive class differs between the protected and privileged group, the truth must be in the middle. Based on this per-group estimate, they compute what would be the unbiased positive class probability for the protected group as follows: $\sum_{i=1}^{k} P(e_i \mid s) P^*(\hat{y} = + \mid e_i)$. The formula for the privileged group is the same. Hence, the explainable difference between the two communities then becomes:

$$\begin{aligned} D_{explainable}(E) &= \sum_{i=1}^{k} P(e_i \mid s) P^*(\hat{y} = + \mid e_i) - \sum_{i=1}^{k} P(e_i \mid ns) P^*(\hat{y} = + \mid e_i) \\ &= \sum_{i=1}^{k} \left(P(e_i \mid s) - P(e_i \mid ns)\right) P^*(\hat{y} = + \mid e_i) \end{aligned}$$

The illegal discrimination then becomes the part of the discrimination that cannot be explained by the attribute E:

$$D_{illegal}(E) = D_{all} - D_{explainable}(E) \ ,$$

where D_{all} is equal to the demographic disparity:

$$D_{all} = P(\hat{y} = + \mid ns) - P(\hat{y} = + \mid s) \ ,$$

which is 1/5 for our toy dataset as calculated above.

With these formulas we get:

$$\begin{aligned} D_{explainable}(Hobby) =&(2/5-0)\times 1/2+(3/5-0)\times 1/3+(0-3/5)\times 0\\ &+(0-2/5)\times 1/2\\ =&1/5\ \text{, giving } D_{illegal}(Hobby)=0. \end{aligned}$$

Similarly we can compute $D_{illegal}(School) = -2/15$, and $D_{illegal}(IQ) = 28/75$.

This example shows that according to the explainable discrimination measure of Kamiran et al. [2013], variable (*Hobby*) could justify the discrimination, while the model is not even using this attribute. This shows the key difference with conditional fairness and our metric: we look at the factors that could change the decision of the model and where these factors differ the most between sensitive groups, while conditional fairness will search for a way to create stratified groups that satisfy a fairness metric.

6.4.4 *Machine learning model*

The machine learning model used for our experiments is a Random Forest model, tuned through five-fold cross-validation. We use a OneHotEncoder to handle the categorical features. The parameter grid that is used is $\{10, 50, 100, 500, 1000, 5000\}$ for the number of trees and $[10, 100, 500, n]$ for the maximum number of leaf nodes.

To measure the *explicit* and *implicit bias* we split each dataset in a training and test set, train the machine learning model on the training set, and then assess the accuracy and fairness on the test set. We generate all the counterfactuals to assess the *explicit bias* as well as the *implicit bias* on the test set. For each dataset we compare three situations: the accuracy and fairness of the model trained with the sensitive attribute (1), the accuracy and fairness of the model trained without the sensitive attribute (2) and the accuracy and fairness of the model without the sensitive attribute and $PreCoF_1$ (3). We expect the accuracy to go down and the fairness to go up going from situation 1 to situation 3 but the trade-off may be different for each dataset. We calculate the fairness by measuring the demographic disparity, which is also equal to D_{all}.

6.4.5 *Counterfactual methodology*

As described in Sections 6.4.2 and 6.4.3, we do not use an existing counterfactual explanation method but develop one ourselves to check for every attribute whether it results in a class change. We use this approach instead of an existing counterfactual explanation method to constrain our method to check every attribute, and hence we have a guarantee that any attribute that more often results in a class change for one group than for another is found.

There exist plenty of counterfactual explanation methods already, and they can lead to different explanations as the optimization problem is set up in a different way [Bordt et al., 2022]. Even a single counterfactual explanation method could lead to a large number of explanations, where the choice of parameters (e.g., the distance metric) could determine which explanations are returned first. This abundance of explanations is not desirable in an adversarial context, as the adversary (in this case the model developer) has considerable freedom to choose which explanation it would return and as such hide bias [Barocas et al., 2020, Bordt et al., 2022]. This is why we use our own counterfactual explanation method: it will not rely on any input parameter that can be manipulated, and neither it will depend on which explanations are returned first as it will check all the attributes, even after several possible explanations are already found. This approach is needed to make tangible statements about whether there is explicit bias, or whether attributes are more often present in the explanations of one group than the other. A drawback of our method is that we limit ourselves to explanations with one feature only, as we do a complete search.

Note that in spite of this reasoning, we did also compare the results found with our counterfactual explanation method with the results when using NICE [Brughmans et al., 2023a] as counterfactual explanation method. We see that in general the same patterns are found, i.e. the same direction of explicit bias and the same *PreCoF* attributes, but that our method is better to detect all cases of *explicit bias* and is better suited to make robust statements about the occurrence of each attribute.

Several works list *actionability* and *plausibility* (adherence to data manifold) as desirable properties of counterfactual explanations [Guidotti, 2022, Karimi et al., 2021, Verma et al., 2020, 2021]. These are two distinct concepts where the former restricts actions to those that are *possible to do*, and the latter requires that the resulting counterfactual instance is *realistic* or in line with the data manifold [Karimi et al., 2021]. We will not take these two properties into account, which is out of line with the *algorithmic recourse* literature: focusing on *actionability* and *plausibility* can actually decrease the ability of our metric to detect bias. After all, our goal is not to look for realistic and actionable advice but to show how the model might be discriminating. For example, the counterfactual explanations to change your race or gender are not *actionable*, however, they are valuable to show explicit bias in the model. Wachter [2022] shows that when immutable characteristics form the basis for decision-making, the decision is likely to be based on undue stereotyping and protection should be offered. That is exactly what we seek to find, while allowing both actionable and immutable features to occur in the explanations. Likewise, imagine a dataset for hiring decisions where all the men are tall and all the women are small: if we want *plausible* counterfactual explanations, women cannot receive the explanation that they should be taller because this will be out of the data manifold. However, in our case, this is, once more, exactly what we are interested in to detect implicit bias. [11]

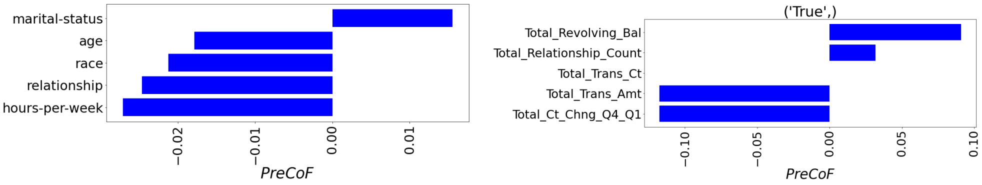
6.5 results

6.5.1 *Adult Income dataset*

When looking at the positive rate of both men and women in Table 6.2, we see that men have a higher positive rate both before and after removing the sensitive attribute. When we

[11] The Python implementation of the proposed metric is available through: https://github.com/ADMAntwerp/PreCoF.

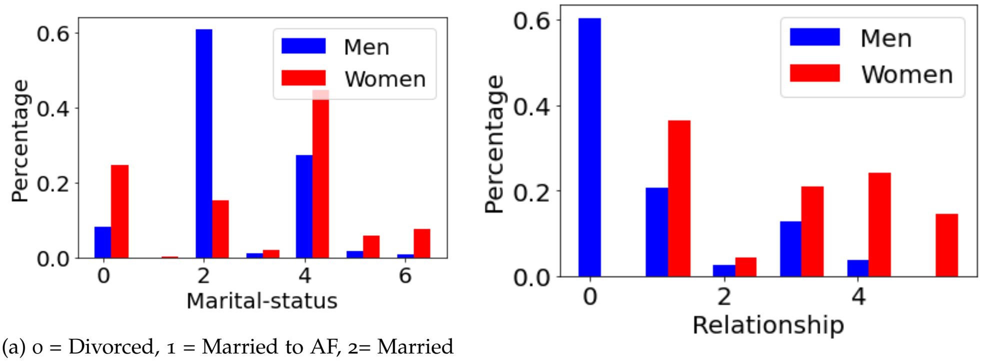
investigate the *explicit bias* of the model (and train the model with the sensitive attribute), we see that the explanation: *'If you would have been a man, you would have been predicted to have a high income'* is present 13 times, while the reverse explanation (*'If you would have been a woman, you would have been predicted to have a high income'*) is only present once. Afterwards, we investigate the implicit bias of the model trained without the sensitive attribute. When we compare the explanations between men and women in Figure 6.1a, we see that women more often receive the explanation *marital-status*. When we look at the exact feature values of the explanations received in Figure 6.1b, so the value of that feature they should change to in order to receive a favorable outcome, we see that the explanations *Marital status: Married to a civilian spouse* and *Relationship status: Husband* are much more prevalent for women than for men. The latter clearly is a proxy, as we see in Figure 6.2b, that this value is only present for men. As we see in Figure 6.2a, the value *Marital status: Married to a civilian spouse* is also present more often for men than for women. Whether marital status is a reasonable attribute to explain the difference in income between men and women, is not for us to decide, but it is valuable to show this pattern so this can be evaluated.



(a) *PreCoF*: attributes in the counterfactual explanations

(b) *PreCoF$_c$*: attribute values in the counterfactual explanations

Figure 6.1: Difference in PreCoF for men and women in the Adult Income dataset



(a) 0 = Divorced, 1 = Married to AF, 2= Married to Civ. Spouse, 3= Married to Abs. Spouse, 4 = Never married, 5 = Separated, 6 = Widowed

(b) 0 = Husband, 1 = Not in a family, 2 = Other relatives, 3 = Own children, 4= Unmarried, 5 = Wife

Figure 6.2: Relationship between sex and the attributes marital status/relationship

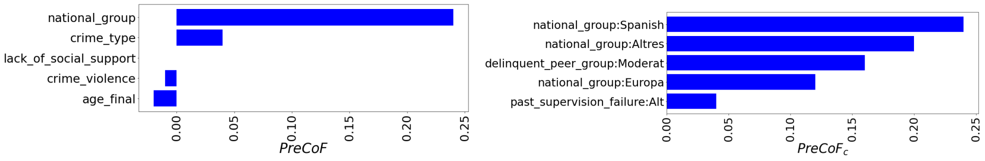
We see in Table 6.2 that the demographic disparity becomes even larger when we remove the sensitive attribute, which is an example of one of the cases where removing the sensitive attribute hurts the *protected group*. When we also remove *PreCoF$_1$* (*marital status*) it decreases slightly but still remains very large.

	Situation 1 *Model with sensitive attribute*	**Situation 2** *Model without sensitive attribute*	**Situation 3** *Model without sensitive attribute and PreCoF$_1$*
Demographic disparity *(Positive rate privileged group - positive rate protected group)*	0.170 *(0.242-0.073)*	0.171 *(0.242-0.071)*	0.168 *(0.236-0.068)*
Accuracy of the model	86.28%	86.23%	86.30%

Table 6.2: Accuracy and fairness metrics for the model trained on on the Adult Income dataset

6.5.2 *Catalonia Juvenile dataset*

We first use our metric to detect *explicit bias* in the model trained with the sensitive attribute. There are 7 foreigners (out of 28) that receive the explanation: *'If you would have been a local, you would have been predicted to not reoffend'* and the reverse case never happens. We also see in Table 6.3, that there is a large demographic disparity in Situation 1 (the model trained with the sensitive attribute). When we remove the sensitive attribute, the demographic disparity goes down but foreigners (*Estrangers*) are still disadvantaged as they are more likely to be predicted to reoffend by our model, compared to locals (*Espagnols*). When we look at the explanations in Figure 6.3a, we see that *national group* is much more present in the explanations of foreigners than in the explanations of locals. As can be seen in Figure 6.4b, this is a clear proxy for foreign status and should also be deleted when race attributes are not allowed. When we zoom in on the feature values in the explanations in Figure 6.3b, we also see which values of national group occur most in the explanations. We see that foreigners are proportionally most likely to receive the explanation to change to *national group: Spanish* in comparison with locals, as it is a proxy for being local. Other national groups that often occur are *Altres* and *Europa*. When we look at the values occurring most often in the factual instances that receive such a class change in Figure 6.4a, the national groups *Central and South America* and *Magrib* are among the most present. Hence, in this case, *PreCoF* succeeds in flagging proxy attributes which could be very helpful for deciding which attributes should be omitted from models.

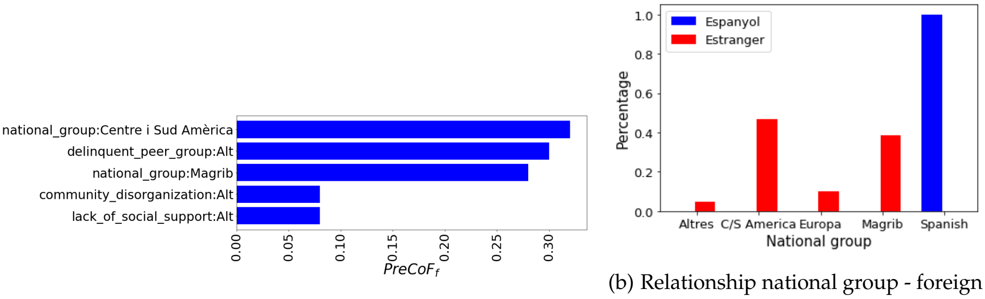


(a) *PreCoF*: attributes in the counterfactual explanations

(b) *PreCoF$_c$*: attribute values in the counterfactual explanations

Figure 6.3: Difference in *PreCoF* for foreigners and locals in the Catalonia Juvenile dataset

We see in Table 6.3 that the demographic disparity goes down when removing the sensitive attribute, but nevertheless still remains quite large. When we also remove *PreCoF$_1$* (*national group*), the demographic disparity almost disappears. The accuracy also goes down when removing this attribute but only slightly.



(a) $PreCoF_f$: attribute values of the factual instances

(b) Relationship national group - foreign status

Figure 6.4: Catalonia Juvenile dataset: analysis

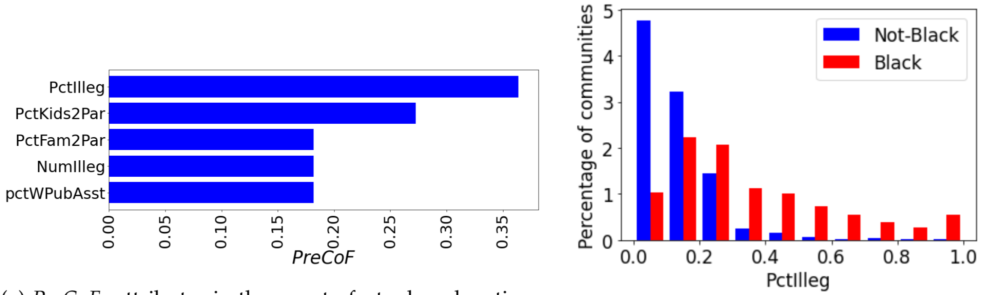
	Situation 1 *Model with sensitive attribute*	**Situation 2** *Model without sensitive attribute*	**Situation 3** *Model without sensitive attribute and $PreCoF_1$*
Demographic disparity *(Positive rate privileged group - positive rate protected group)*	0.175 *(0.897 - 0.723)*	0.119 *(0.812-0.752)*	0.010 *(0.782-0.772)*
Accuracy of the model	71.98%	72.37%	70.82%

Table 6.3: Accuracy and fairness metrics for the model trained on the Catalonia juvenile dataset

6.5.3 *Crime and communities dataset*

We find no cases of *explicit bias* in the model trained with the sensitive attributes. Next, we train a model without the sensitive attribute and assess the implicit bias. We see in Table 6.4 that the not-black communities in the test set are never predicted to be a violent community so their positive rate is 100 %. Black communities are predicted to be violent in around 4.5% of the cases. We hence have only explanations for the protected group, so we will just see which explanations were the most present for this group. In Figure 6.5a, we observe that the attribute *PctIlleg*, which is the percentage of kids born to people who were never married, is the most present. When we look at the distribution of this attribute for black and non-black communities in Figure 6.5b, we indeed see that this percentage tends to be higher for black communities. Research on other models trained on this dataset also find this to be an important predictor of both the target value (violent community) as well as the sensitive attribute (black community) [Le Quy et al., 2022]. When we assess the other top attributes in *PreCoF*, we notice that the four first are related to families with both parents being present, or being married. Earlier research already argued that marriage is linked to a reduction in crime [Sampson et al., 2006].

We also see in Table 6.4 that the demographic disparity goes down when we remove the sensitive attribute. It does not go down when we remove $PreCoF_1$, which makes sense as the $PreCoF_1$ attribute here (*PctIlleg*) is very correlated with other attributes of the dataset such as *NumIlleg*.



(a) *PreCoF*: attributes in the counterfactual explanations (all the attributes in this dataset are numerical so no need to investigate *PreCoF$_c$*)

(b) Relationship between *PctIlleg* and Black in the Crime and Communities dataset

Figure 6.5: Crime and Communities dataset: analysis

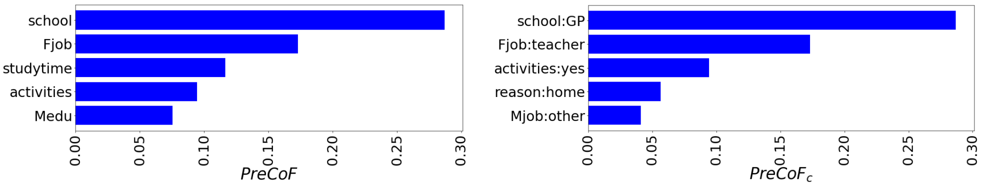
	Situation 1 *Model with sensitive attribute*	**Situation 2** *Model without sensitive attribute*	**Situation 3** *Model without sensitive attribute and PreCoF$_1$*
Demographic disparity *(Positive rate privileged group - positive rate protected group)*	0.045 *(1-0.955)*	0.035 *(1-0.965)*	0.035 *(1-0.965)*
Accuracy of the model	84.97%	85.14%	84.81%

Table 6.4: Accuracy and fairness metrics for the model trained on the Crime and Communities dataset

6.5.4 *Student performance dataset*

We see in Table 6.5 that our classifier predicts girls to be less likely to have a positive label compared to boys. Although they have on average a higher score than boys, they are more often predicted to fail in every situation. We might get some insights into this phenomenon by looking at how the explanations differ for both groups. We see in Figure 6.6a that the attribute *school* is present more often in the explanations for girls and in Figure 6.6b that they receive the explanation to change to *school GP* more often. Depending on what the machine learning model is used for, this kind of analysis could give very important insights. If this model would be used, for example, to determine whether the students would be successful in university and should be admitted, this analysis shows that girls could be disadvantaged compared to boys because of the school they went to. When we look at the *explicit bias* in the model trained with the sensitive attribute, boys are biased against: there are three boys that receive the explanation: *'If you would have been a girl, you would have been predicted as scoring above average instead of below'* and the reverse does not happen. This example shows that *explicit bias* and *implicit bias* can work in opposite ways.

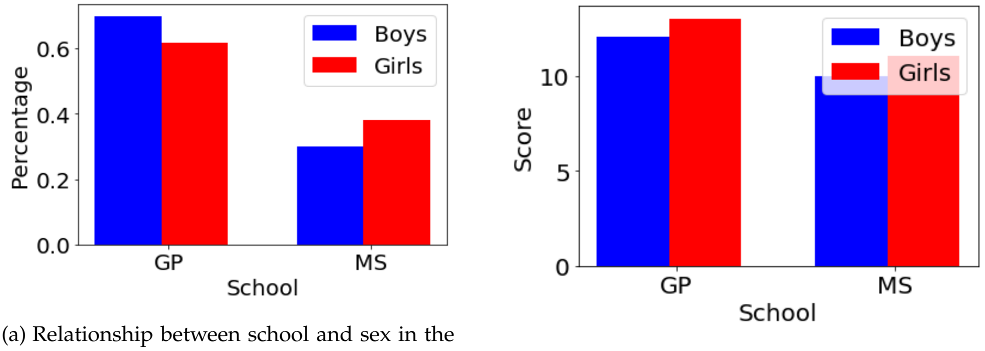
We analyse the relations of the attribute *school*. We see in Figure 6.7b that for both boys and girls, their average score is higher if they went to school GP: for girls their average score on school GP is 13 and on school MS 11.03, while for boys the average score on school GP is 12.03



(a) *PreCoF*: attributes in the counterfactual explanations

(b) *PreCoF$_c$*: attribute values in the counterfactual explanations

Figure 6.6: Difference in explanations for boys and girls in the Student performance dataset



(a) Relationship between school and sex in the Student Performance dataset: percentage of this sex that goes to this school

(b) Relationship between school and sex in the Student Performance dataset: average grade

Figure 6.7: Student performance dataset: analysis

and on school MS 9.95. The average score of girls is also higher independent of school: on average girls have a score of 12.25 and boys of 11.41. When researching this attribute, we see in Figure 6.7a that girls more often go to *school MS* which has a lower average score, so they receive the explanation to change to *school GP*, which has a higher average score, more often. So due to the importance of the attribute school in the machine learning model, they are predicted to fail more often than boys while their true outcome is to fail less. The importance of the school you go to in a machine learning model to predict grades reminds of a recent case in England in 2020, where an algorithm designed to predict grades for A-level exams amidst COVID-19 increased the predicted grades at small private schools but lowered the grades at larger, state-run schools that have a larger proportion of minority students [Wachter et al., 2020]. In terms of accuracy, this model performed well but as a result high performing students from 'good schools' received high marks, whereas highly performing students from 'bad schools' had their marks capped by the lower performance of classmates and got a lower mark than deserved [Wachter et al., 2020]. This system was not biased on purpose: it was the ignorance of the social bias that led to the technical bias in this system [Wachter et al., 2020].

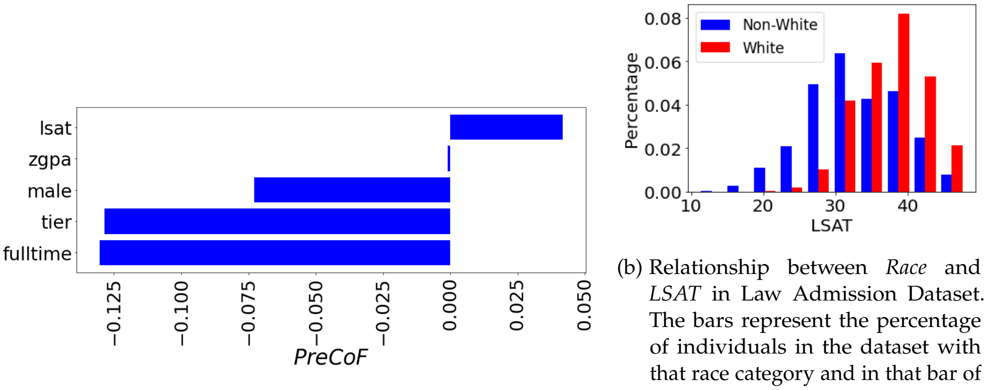
We compare the accuracy and fairness of the three situations in Table 6.5: We see that the accuracy of the model goes down after removing attributes, however only slightly. We see that the demographic disparity increases after removing the gender attribute, which makes sense as girls on average scored better but are disadvantaged by the school they go to: this effect

	Situation 1 *Model with sensitive attribute*	**Situation 2** *Model without sensitive attribute*	**Situation 3** *Model without sensitive attribute and PreCoF$_1$*
Demographic disparity *(Positive rate privileged group - positive rate protected group)*	0.043 *(0.610-0.566)*	0.115 *(0.646-0.531)*	0.066 *(0.659-0.593)*
Accuracy of the model	73.85%	71.28%	70.26%

Table 6.5: Accuracy and fairness metrics for the model trained on the Student Performance Dataset

will become even larger if gender information is removed. There is *explicit bias* against boys, but *implicit bias* against girls through the neutral attribute *school*. If we remove *PreCoF$_1$ School*, the demographic disparity will decrease again but not until the first level. This situation shows that as mentioned in literature already [Corbett-Davies et al., 2023] and as seen in other datasets, removing the sensitive attribute can increase the discrimination in the dataset.

6.5.5 *Law admission dataset*



(a) *PreCoF*: attributes in the counterfactual explanations

(b) Relationship between *Race* and *LSAT* in Law Admission Dataset. The bars represent the percentage of individuals in the dataset with that race category and in that bar of LSAT scores.

Figure 6.8: Law Admission dataset: analysis

When we look at the *explicit bias*, we see that there are 45 instances in the test set that receive the explanation: *'If you would have been white, you would have been predicted as admitted to pass the bar'* and only 3 the other way around, which shows that the model that is trained with the sensitive attribute exhibits *explicit bias*. This also shows that the model is non-linear and both parties can receive such explanations. When we train the model without the sensitive attribute here, we see in Figure 6.8a that the only attribute that is relatively more present in the explanations of *Non-Whites* compared to *Whites*, is the *lsat score*. The fact that almost all the attributes are relatively more present in the explanations of the privileged group means that the rejected individuals in this group are closer to the decision boundary: Changing only one attribute more often leads to a change in outcome, while for the protected group more attribute changes are necessary. It is not surprising that *lsat scores* pop up as $PreCoF_1$ as it is often said that test scores such as GPA and LSAT are racially biased: white test-takers consistently score higher than minority test-takers [White, 2000] and there have been calls for law school admission committees to deemphasize reliance on LSAT scores and to develop new methodologies to assess the skills of each applicant [Hill, 2019]. When we look at Figure 6.8b, we indeed see that the average score of the LSAT is higher for *Whites* compared to *Non-Whites*.

	Situation 1 *Model with sensitive attribute*	**Situation 2** *Model without sensitive attribute*	**Situation 3** *Model without sensitive attribute and* $PreCoF_1$
Demographic disparity *(Positive rate privileged group - positive rate protected group)*	0.159 *(0.994-0.835)*	0.143 *(0.990-0.847)*	0.075 *(0.987-0.912)*
Accuracy of the model	89.94%	89.82%	89.63%

Table 6.6: Accuracy and fairness metrics of the model trained on the Law Admission dataset

When we compare the accuracy and fairness of the three situations in Table 6.6, we see that the accuracy decreases very slightly when removing the sensitive attribute and $PreCoF_1$. When removing the sensitive attribute, the demographic disparity decreases slightly but after removing $PreCoF_1$, it decreases substantially. The question can be asked here whether we deem it fair that there is a difference in positive rate based on LSAT scores: are these objective scores or are they already biased in se?

6.6 DISCUSSION

In this study, we use counterfactual explanations to shed light on which discrimination occurred in models trained on some well-known datasets, both in terms of *explicit* and *implicit bias*. Our experiments reveal that removing $PreCoF_1$, will decrease the demographic disparity in a model, but we want to highlight that this is not the main purpose of our metric. It is possible that removing other attributes will decrease the demographic disparity even more as it is not the goal of the *PreCoF* metric to find that variable that would make the demographic disparity the smallest. Our purpose is not to give members of a protected group an advantage by giving them a better outcome [Wachter et al., 2020], but rather to shed light on which attributes resulted in a different outcome and jump-start a discussion on whether they are based on historical inequalities or are justified discriminators. The fairness results (i.e., the decrease in demographic disparity) simply show that removing the $PreCoF_1$ variable will result in a smaller difference in positive rate between the protected and the privileged group, which can be a desirable outcome in some cases.

What does our technique add compared to other fairness metrics?
Fairness will depend on context-dependent judgements, so it is dangerous to treat the quantitative fairness metrics discussed in Section 6.2.1 as black-box fairness measures [Corbett-Davies et al., 2023]. Using group metrics for fairness can abstract away more subtle issues that are too difficult to operationalize or to decide upon algorithmically [Yeom and Tschantz,

2021]. There is not one criterion that can ensure fairness in all cases, and when a model fails on a fairness metric, this should lead to an investigation as to why this happens [Yeom and Tschantz, 2021]. We also confirmed that just removing the sensitive attribute is not a viable approach as it can even amplify the discrimination of the model, and thus harm the group it was supposed to protect [Corbett-Davies et al., 2023]. *Demographic Parity* can detect whether the model is treating the sensitive groups differently when the model does not directly use the protected attribute but correlated one(s), but it does not consider whether there is sufficient justification for a disparity of outcomes [Yeom and Tschantz, 2021]. Other tests that do take the ground truth into account such as *equalized odds* also just examine the disparities but not how they were reached [Yeom and Tschantz, 2021].

We do not state that removing $PreCoF_1$ to decrease the *demographic disparity* will be a universal solution to tackle the discrimination in a dataset. We just showcased that it is a possible approach. Our point of view is that this should be decided case by case: is this attribute a justified discriminator? Does it just behave as a proxy? Is it warranted to sacrifice accuracy for extra fairness? Is a difference in positive rate a problem when the true outcomes also differ per sensitive group or an accepted consequence? Do the observed outcomes accurately reflect the real world? This last question is related to the two worldviews that Friedler et al. [2021] suggested, namely the *'We are all equal'* worldview and the *'What you see is what you get'* worldviews. These are all questions that should be answered for each case individually, and our metric can help to decide upon them. The benefits of building more fair models could be very large, as fair machine learning models could dramatically improve the equality of consequential decisions [Corbett-Davies et al., 2023].

6.7 future research and limitations

There are limitations to our metric, which at the same time pose opportunities for future research. The patterns detected by this metric will only be trustworthy if both groups in the test set are large enough. Therefore, we do not include the German Credit dataset into our experiments, as this is a very small dataset. The number of individuals with a bad outcome in each sensitive group in the test set will be so small that it is not possible to draw conclusions from them.

Another limitation is that we do not take into account the type of bias that is present in the dataset. If we assume the labels are biased, and that that is why fairness corrections are needed, measuring the performance of the machine learning model on those biased labels also leads to incorrect estimates.

Furthermore, in the COMPAS Juvenile dataset we detect an interesting pattern; every attribute is relatively more present in the explanations of the not African-American group than in the African-American group. This pattern occurs because the 'rejected' individuals (individuals which are predicted the unfavorable outcome by the machine learning model) in the former group are on average closer to the decision boundary than the individuals in the latter group: for the latter, one attribute change will less often be enough to result in a class change. This is related to the fairness notion of CERTIFAI [Sharma et al., 2020] and algorithmic recourse [von Kügelgen et al., 2022], where the effort of both groups to reach the desired target outcome is taken into account. Our metric only looks at univariate changes for now but this could be expanded to changes of two or more attributes in future research.

In our experiments, we focus on the rejected individuals. Another interesting research avenue would be to focus on the *misclassified* rejected individuals and see what are the most occurring explanations for both sensitive groups. This could be a possible avenue to improve the model and reduce misclassifications.

Lastly, this study only takes tabular datasets into account but it will be valuable to analyze this on text and behavioral datasets, as they are very sparse. For some tabular datasets, we know what we can expect as proxies, however for behavioral datasets like Facebook likes, this might not be very intuitive. This will be the focus of our next research.

6.8 CONCLUSION

Fairness literature in AI has already revealed that AI creates new challenges for detecting discrimination: automated discrimination is less intuitive, subtle and intangible [Wachter et al., 2021]. As the algorithmic world will make complex decisions without any reasoning behind them, it will be challenging to detect whether you are treated fairly. It is misguided to focus on fairness while not obtaining transparency first [Rudin et al., 2020]. We aim to provide this transparency by providing a tool that can shed light on: how often *explicit bias* in the decision making model occurs for each subgroup, and which factors are a cause of the *implicit bias* in the decision making model in each subgroup.

7

Reranking individuals: The effect of fair classification within-groups

Artificial Intelligence (AI) finds widespread application across various domains, but it sparks concerns about fairness in its deployment. The prevailing discourse in classification often emphasizes outcome-based metrics comparing sensitive subgroups without a nuanced consideration of the differential impacts *within* subgroups. Bias mitigation techniques not only affect the ranking of pairs of instances across sensitive groups, but often also significantly affect the ranking of instances within these groups. Such changes are hard to explain and raise concerns regarding the validity of the intervention. Unfortunately, these effects remain under the radar in the accuracy-fairness evaluation framework that is usually applied. Additionally, we illustrate the effect of several popular bias mitigation methods, and how their output often does not reflect real-world scenarios.

7.1 INTRODUCTION

In the rapidly evolving landscape of Artificial Intelligence (AI) and machine learning, the pursuit of fairness in algorithmic decision-making has emerged as a central concern. As the influence and scope of the decisions made by AI systems are increasing, there are growing concerns that the models making these decisions might unintentionally encode and even amplify human bias [Corbett-Davies et al., 2023]. *Algorithmic bias* describes situations where sensitive groups are substantially disadvantaged by an algorithm or model. One of the ways bias can seep into a model is when it is trained on biased data, following the famous *garbage in, garbage out* principle which emphasizing that flawed input data results in flawed output [Geiger et al., 2020]. Examples of biased AI models are everywhere, with cases in almost every domain. In the context of hiring, a well-known case is that of an automated Amazon recruitment system that had to be pulled because it was biased against female applicants Dastin [2022]. Much earlier already, St George's Hospital Medical School's Commission for Racial Equality discovered that a computer program used for initial screenings of applicants "written after careful analysis of the way in which the staff were making these choices" unfairly rejected women and individuals with non-European sounding names Johnson [2021], Lowry and Macpherson [1988]. There is an abundance of examples akin to these ones.

In this paper, we focus on *fair classification*, which ensures algorithms make unbiased decisions across groups. Many bias detection and mitigation methods exist, but most of them focus on "between-group fairness" where the primary objective is to rectify disparities in model predictions between distinct demographic groups. This endeavor is undeniably critical, as it aims to rectify long-standing inequalities. However, it is equally essential to acknowledge and scrutinize the complexities that exist "within" these groups, giving rise to a concept that is commonly referred to as "within-group fairness". Speicher et al. [2018] already note that many approaches to group fairness tackle only between-group issues, worsening within-group fairness as a consequence. Krco et al. [2023] highlight that the blind optimization of commonly used fairness metrics does not show who is impacted *within* each group, while Mittelstadt et al. [2023] emphasize that many of the currently used bias mitigation methods can make every group worse off. These issues illustrate why solely looking at fairness by measuring disparities between groups is not adequate.

While various benchmarking studies attempt to evaluate the performance of bias mitigation methods, they often fall short, comparing what can be described as 'apples to oranges'. This issue arises because different bias mitigation methods can significantly vary the number of positive instances, and by comparing them as-is, the actual situation faced by practitioners is not taken into account. Scantamburlo et al. [2024] also argue that the ultimate decision of an automated system is informed by the prediction model, but in nearly all cases is also influenced by additional parameters such as quota or business rules. They make a distinction between the prediction model and the decision-making system, and discuss how the field of fair machine learning tends to blur the boundaries between the two concepts [Scantamburlo et al., 2024]. Kwegyir-Aggrey et al. [2023] confirm that when deploying a classifier in the real world, practitioners typically need to tinker with the threshold to make sure the model predictions meet their domain-specific needs.

This is why directly comparing these methods on prediction labels as-is is not a good idea. However, this is the current state of the art in benchmark studies of bias mitigation [Chen et al., 2023, Hort et al., 2023, Hufthammer et al., 2020, Reddy, 2022, Janssen and Sadowski, 2021, Menon and Williamson, 2018]. The current metrics fall short in acknowledging the inherent differences in the situations they compare, rendering them insufficient for determining the superior performance of one method over another. More importantly, they fail to take into account the constraints in positive decision rate that are faced in the real world. For example, our results demonstrate that for the Adult Income dataset one bias mitigation method yields a positive decision rate of 0.5% while another results in a positive decision rate of 39.3%, whereas the actual positive rate of the dataset is 23.9%. If we compare the accuracy of these two mitigation methods, we are comparing two entirely distinct points of the ROC curve, both of which may be inapplicable in real-world settings. Moreover, the majority of bias mitigation methods do not result in the satisfaction of the fairness metrics, necessitating additional post-hoc interventions after their application.

We argue that the focus should shift to fairness within groups, and the crucial question to be answered should be: *how can we select the best individuals within each group?* Machine learning models not only return prediction labels, they also return prediction scores, which creates an inherent ranking of instances. Given that the prediction labels from many mitigation methods are often impractical due to their unrealistic positive decision rates and non-compliance with the selected fairness metric, it is more logical to concentrate on the prediction scores produced by each mitigation method. The objective should be to have prediction scores that reflect the true target label as accurately as possible. Afterwards, post-hoc fairness methods can be deployed to select the individuals with the highest scores within each group, aligning with both industry constraints and fairness metrics. This separates the function of the prediction model from the decision-making context [Scantamburlo et al., 2024]. However, note that using distinct thresholds for different sensitive groups may be illegal in some contexts as it requires the explicit use of the sensitive attribute in the decision-making process.[1]

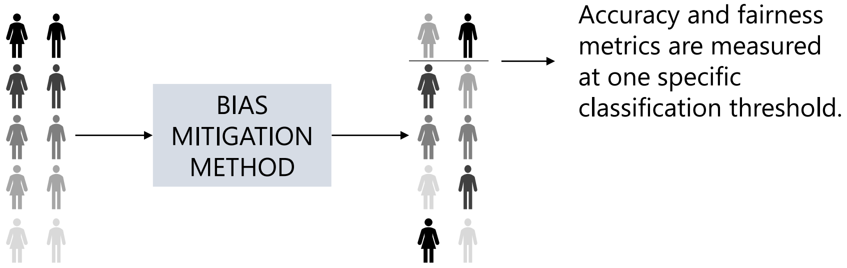

BIAS
MITIGATION
METHOD
Accuracy and fairness metrics are measured at one specific classification threshold.

Figure 7.1: The ranking within each group is shifted by most preprocessing and inprocessing bias mitigation methods. The current state-of-the-art is to measure the accuracy and fairness metrics at one specific classification threshold, and not taking into account the swaps within each group.

When striving to establish the most optimal ranking, it is crucial to differentiate between pre- and inprocessing bias mitigation methods on the one hand, and postprocessing methods on the other. In general, only pre- and inprocessing methods will alter this ranking by either

[1] This framework thus operates under the assumption of having access to the sensitive attribute(s) and the legality of using it.

fine-tuning the training data or incorporating fairness constraints during model training, as this will change the prediction scores of the machine learning model. In contrast, most postprocessing methods merely flip prediction labels based on the scores of the initial machine learning model, without altering the inherent ranking of instances. Figure 7.1 visualizes how the inherent ranking is shifted by pre- and inprocessing mitigation methods, emphasizing the common practice of assessing performance and fairness metrics at a single classification threshold. Furthermore, it underscores the tendency to overlook the impact of fairness interventions on the rankings within each subgroup. This seemingly arbitrary reshuffling within each group that occurs as a side effect of fair classification, is currently not studied. In this process, certain individuals (from both the protected and the privileged group) who were initially assigned a positive prediction label will now be labelled negatively, while some group members that initially received a negative label will be switched to a positive one. This reranking process is not per definition negative, as we will discuss later, but deserves more attention.

This paper serves two purposes:

- To provide insights into the operational dynamics of bias mitigation methods and illustrate how some methods to achieve fair classification will significantly impact the ranking within each group, while others will not.

- To criticise the current approach of benchmarking bias mitigation methods, as it compares situations that are significantly distinct and not applicable in real-world settings. We demonstrate that the different bias mitigation methods lead to very different positive decision rates, and argue that no distinction is made between the prediction model and the decision-making context.

7.2 BACKGROUND

Before we go further, it is important to define some of the key terminology that is often used in the fairness literature. A *sensitive attribute* refers to a characteristic or feature of an individual that is considered sensitive, often with respect to potential discrimination. This can include attributes such as race, gender, age, religion, sexual orientation, or any other factor that could be the basis of unfair treatment. Consequently, a *protected group* typically refers to the demographic group that is at risk of being unfairly treated or discriminated against based on their sensitive attribute, while the *privileged group* is the demographic category that is typically not subject to unfair treatment based on that sensitive attribute. In this paper, we operate under the assumption of a single binary sensitive attribute, implying the existence of a protected group and a privileged group. *Fairness metrics* are quantitative measures used to assess the fairness of AI models, while *fairness (or bias) mitigation strategies* are techniques used to either learn an AI model that is fair by design or modify AI models to reduce bias.

In the computer science community, a plethora of fairness metrics have been proposed [Corbett-Davies et al., 2023]. One of the most popular approaches is the *group fairness metric*, which quantifies the fairness of a machine learning model across different demographic or sensitive groups, aiming to identify disparities in the outcome between these groups. One of the simplest and most commonly used definitions in this category is *demographic parity* (or *statistical parity*), which states that the positive decision rate must be the same regardless of the value of the protected attribute. In our example of hiring, this means that a model must invite equal percentages of white and black applicants for an interview (if race is the sensitive attribute) or of male and female applicants (if gender is the sensitive attribute). Other commonly used metrics include *equalized opportunity*, which states that there should be an equal proportion of true positives in both groups, and *equalized odds*, which examines whether both the proportion of true positives and true negatives is equal across groups. Besides these, many other fairness metrics exist, and the issue is that most of them cannot be optimized at the same time Kleinberg et al. [2016]. Deciding upon a group fairness metric to optimize thus means already imposing a certain world view. Another popular approach to assess fairness in machine learning models is *individual fairness*, which demands that similar individuals receive similar outcomes in a decision-making process, regardless of their group membership [Dwork et al., 2012, Binns, 2020]. Dwork et al. [2012] argue that instead of focusing on a group, we tend to care more about the individual. This notion is related to the contributions of this paper, as we will also argue that satisfying group fairness metrics is not necessarily fair from the viewpoint of the individual. A last important metric is *calibration*. Calibration ensures that the predicted probability of a group of instances reflects the fraction of those individuals that actually have a positive label [Pleiss et al., 2017]. In the context of fairness between groups, we would like this calibration condition to hold simultaneously within these groups as well.

A common starting point for designing a fair algorithm is simply to exclude sensitive attributes from the model, However, the limitations of this approach have been commonly addressed, with the most fundamental limitation being *the proxy problem* [Prince and Schwarcz, 2019]. The proxy problem states that the omission of sensitive attributes can lead the machine learning model to rely on proxy variables that indirectly encode the information contained in the sensitive attribute and hence still introduce bias into the model's decision-making process. A classic example of the proxy problem is the use of zip codes in the United States as a proxy for racial information, as these two attributes tend to be heavily correlated. This has prompted many to argue that proxies should be excluded from the dataset as well, however, this is very difficult to operationalize [Corbett-Davies et al., 2023]. This is because every attribute used in the machine learning model can be at least partially correlated with the sensitive attribute; and often, even strongly correlated attributes may be considered legitimate factors on which to base decisions (for example, education in the case of hiring) [Corbett-Davies et al., 2023].

This illustrates that creating a fair machine learning model is a tedious process. In response, many bias mitigation methods that claim to improve fairness, have been introduced. We can divide most of them into three categories: preprocessing, inprocessing and postprocessing. Each category targets a different stage of the machine learning pipeline to ensure fairness. The idea behind preprocessing methods is that they will change the representation of the data before the machine learning model is learned, and as such neutralize any prejudiced information that could affect the model's decision. Inprocessing methods improve fairness during the training process, by incorporating fairness constraints in the learning algorithm and striving for a balance between accuracy and fairness. Postprocessing methods intervene after the model has made its predictions, by adjusting the outcomes to satisfy fairness criteria. We will discuss the used bias mitigation methods in more detail in the Materials and Methods section. These bias mitigation methods focus on satisfying the aforementioned group fairness metrics that measure disparities between groups Chen et al. [2023].

Another research area that we should consider is the area of fair ranking. Yang and Stoyanovich [2017] measure whether a ranking is fair by comparing the distribution of protected and non-protected candidates on different prefixes of the list, while Zehlike et al. [2017] illustrates how group fairness metrics can be satisfied for different prefixes of the ranking. Yang et al. [2019] demonstrate how adding diversity constraints to ranking algorithms can reduce in-group fairness, a concept related to our measure of within-group fairness. However, it is crucial to highlight the differences with this study, as we study fair classification, which focuses on equal outcomes between groups, and which does not entail that the ranking distribution should be fair. [2] We will show that several methods that are used for fair classification have as side effect that they change the rankings within each group, which in turn has influence on the final prediction labels of each individual. This seemingly arbitrary 'reshuffling' as a consequence of fair classification is currently not studied and deserves more attention.

7.3 MATERIALS AND METHODS

7.3.1 *Materials*

We will use several real world datasets that are common in the domain of fair machine learning [Le Quy et al., 2022]. The **Adult Income** dataset contains information extracted from the 1994 census data with as target variable whether the income of a person exceeds $50,000 a year or not. The **Compas** dataset includes demographic information and criminal histories of defendants from Broward County, and is used to predict whether a defendant will re-offend within two years. The **Dutch Census** dataset represents aggregated groups of people in the Netherlands for the year 2001, and can be used to predict whether a person's occupation can be categorized as a high-level (prestigious) or a low-level profession [Van der Laan, 2000]. The **Law admission** dataset contains a Law School Admission Council (LSAC) survey conducted across 163 law schools in the United States in 1991 [Wightman, 1998] and can be used to predict whether the student will pass the bar exam or not. The **Student Performance** dataset describes the achievements of students in two Portuguese schools [Cortez and Silva, 2008]. The classification task is to predict whether they score above average in mathematics.

Table 7.1: Used datasets

Name	# instances	# attributes	Protected attribute	Protected group	Target attribute	Base rate
Adult	48,842	10	Gender	Female	High income	23.93%
Compas	5,278	7	Race	African-American	Low risk	52.16%
Dutch Census	60,420	11	Gender	Female	High occupation	52.39%
Law admission	20,798	11	Race	Non-White	Pass the bar	88.97%
Student Performance	649	29	Gender	Male	High score in mathematics	53.62%

[2] This can be fixed by postprocessing methods, for example by using different thresholds for each group, for the required positive decision rate. Fair ranking will be more strict, as it requires that the fairness metrics are satisfied for different prefixes of the ranked list.

7.3.2 *Methods*

7.3.2.1 *Machine learning models*

We train fully connected neural networks on each dataset, utilizing binary cross-entropy as the loss metric.[3] Neural networks are chosen because the implementation of one the bias mitigation methods (Adversarial Debiasing) requires this. Although the remaining methods are model-agnostic, we opt for consistency in our approach, employing a neural network across all bias mitigation methods to ensure comparability in our final results.

7.3.2.2 *Bias mitigation methods*

Numerous debiasing strategies currently exist, but we focus on the methods available in the AIF360 package [Bellamy et al., 2019]. For all algorithms, we use the default settings.

As **preprocessing** methods, we use *Learning Fair Representations* (**LFR**) and *Disparate Impact Remover* (**DIR**). The idea behind Learning Fair Representations [Zemel et al., 2013] is that a new representation Z is learned that removes the information correlated with the sensitive attribute, but preserves the other information about X as much as possible. Disparate Impact Remover [Feldman et al., 2015] modifies the training data to reduce the influence of sensitive attributes, but preserves rank-ordering within groups.

We use the **inprocessing** methods of *Adversarial Debiasing* (**ADV**) and the *Meta Fair Classifier* (**MFC**). Adversarial Debiasing [Zhang et al., 2018] combines a classifier that predicts the class label with an adversary that predicts the sensitive attribute. The goal is to maximize the classifier's performance while minimizing that of the adversary. The Meta Fair Classifier [Celis et al., 2019] takes the fairness metric as part of the input and returns a classifier optimized with respect to that fairness metric.

As **postprocessing** methods, we use *Equalized Odds Postprocessing* (**EOP**), *Reject Option Classification* (**ROC**) and *Threshold Optimation* (**TO**). Equalized Odds Postprocessing [Hardt et al., 2016] will solve a linear program to find the probabilities with which to change output labels in order to optimize equalized odds, while Reject Option Classification [Kamiran et al., 2012] will flip the predictions the model is not confident of.[4] Threshold Optimization [Kamiran et al., 2012] is maybe the most straightforward method of mitigating bias as it will optimizes the thresholds for both groups in isolation.[5] ROC and TO are implemented to enforce Demographic Parity, while EOP enforces Equalized Odds by default.

[3] We use a neural network, with one hidden layer with 200 nodes. We use 50 epochs, a batch size of 128 and the Adam optimizer. We do not perform hyperparameter tuning.

[4] Reject option classification identifies instances where the model is uncertain about its prediction and essentially 'rejects' making a definite decision. In this implementation, aiming to enhance fairness, the labels of these instances are flipped to satisfy a fairness criteria.

[5] For this bias mitigation method, we use the implementation available through the Fairlearn toolbox Bird et al. [2020] as this method is not available in the AIF360 toolbox.

7.3.3 *Metrics*

7.3.3.1 *Performance metrics*

Most benchmark studies compare the different mitigation methods on accuracy, which measures how often the prediction label assigned by the machine learning model coincides with the true label [Chen et al., 2023, Hort et al., 2021, Krco et al., 2023]. It is commonly acknowledged that accuracy is not always the best metric to measure the performance of a machine learning model, as it is for example not suitable to deal with imbalanced class distributions (as in this case a model can obtain a high accuracy by just predicting all samples as the majority class) [Chen et al., 2023, Mittelstadt et al., 2023]. This has led some studies to include other metrics such as the F1-score, Precision, or Recall [Chen et al., 2023]. However, another notable drawback of these performance metrics is that they measure the performance at a specific classification threshold, as they use the prediction labels of the machine learning model and not the prediction scores. We can evaluate the performance of the prediction scores by using the Area Under the ROC Curve (AUC). [6]

AUC allows for an objective comparison across classifiers, as it is unaffected by the choice of threshold or the frequency of classes [Hand, 2009]. It measures how well the prediction scores (and thus the ranking) of a machine learning model distinguishes between positive and negative cases. The formula for the **AUC score**, where $S(x_i)$ notes the prediction score of instance x_i:

$$P(S(x_i) > S(x_z) | y_i = 1, y_z = 0)$$

This formula means that the AUC score is equivalent to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Provost et al. [1998] previously advocated for the adoption of AUC as a standard for comparing classifiers in the broader field of machine learning. Despite this, its integration into Fair Machine Learning has been limited.[7] Furthermore, the specific context of Fair Machine Learning presents additional justifications for focusing on the prediction scores generated by machine learning models, rather than solely on their prediction labels, due to unrealistic positive rates and the unsatisfaction of required fairness metrics. In this study, we operate under the assumption that label bias is absent, meaning that the actual labels accurately represent the intended prediction target [Wick et al., 2019, Favier et al., 2023, Lenders and Calders, 2023]. Note that if label bias was present, it would compromise the validity of both the AUC and accuracy metrics, as these measures rely on these labels for their calculations.

Lastly, we will measure the positive decision rate (or positive classification rate) on the whole population. As mentioned earlier, this is important to be able to compare the different bias mitigation methods. We use Y to denote the actual target label, and $\hat{Y}$ to denote the predicted label by the machine learning model. S represents the sensitive attribute, where s represents

[6] Note that this ROC, which stands for Receiver Operating Characteristic is different from the Reject Option Classification, that we also shorten as ROC.

[7] One example is where Fong et al. [2021] investigate how acquiring additional features can improve the AUC of the disadvantaged group.

the protected group and *ns* the privileged group. The formula for the **Positive Decision Rate (PDR)**:

$$P(\hat{Y} = 1)$$

7.3.3.2 *Fairness metrics*

Many possible fairness metrics exist, but we will report two metrics that are commonly used in the fairness domain to measure disparities between groups: Demographic parity (or statistical parity) states that the positive decision rate must be approximately the same in the protected group as in the privileged group. Statistical Parity Difference measures the difference in this positive decision rate between both groups. The formula for the **Statistical Parity Difference** (SPD):

$$P(\hat{Y} = 1 | S = s) - P(\hat{Y} = 1 | S = ns)$$

Equalized opportunity requires the true positive rate to be approximately the same across groups. The formula for **Equal Opportunity Difference** (EOD):

$$P(\hat{Y} = 1 | S = s, Y = 1) - P(\hat{Y} = 1 | S = ns, Y = 1)$$

Larger values of these metrics correspond to a higher level of bias towards one of the sensitive groups [Hort et al., 2021].

7.3.3.3 *Rank correlation*

We can measure the correlation between two ranked lists by using the Kendall Tau metric, denoted as τ. This metric measures the similarity between the orderings of two lists by quantifying the number of pairwise disagreements between them.

The formula for the **Kendall Tau coefficient** (τ) is defined as:

$$\frac{\text{number of concordant pairs} - \text{number of discordant pairs}}{\text{number of pairs}}$$

A concordant pair refers to a pair of observations where the order of the ranks is the same between both lists, while a discordant pair refers to a pair of observations where the order of the ranks is reversed between the two lists. A τ value of 1 indicates perfect agreement between two rankings, whereas a value of -1 indicates perfect disagreement. Values closer to zero suggest little to no correlation, implying a lack of consistency in the ranking order between the two groups being compared.

7.4 RESULTS

7.4.1 *Whom do the bias mitigation strategies affect?*

The lack of transparency in the field of fair machine learning has been acknowledged in literature [Rudin et al., 2020, Wachter et al., 2021, Goethals et al., 2023b], while some have already specifically criticized the opacity in the exact effects of the different bias mitigation methods [Krco et al., 2023, Favier et al., 2023, Holstein et al., 2019]. We aim to shed light on the operational dynamics of the different bias mitigation methods, by comparing the score distributions after deploying each bias mitigation method with the score distributions of the initial machine learning model. We illustrate the prediction scores from the initial ML model on the x-axis and the scores post-application of various bias mitigation methods on the y-axis. Additionally, we categorize the instances based on their affiliation with either the protected group (represented in dark blue) or the privileged group (represented in light blue). The chart is divided into four quadrants, each depicting the classification of instances as either positive or negative by the initial ML model and by each bias mitigation method.[8] We also include a diagonal line that would contain all the instances, if the prediction scores would remain identical. The presented results in Figure 7.2 are calculated using the Compas dataset. The figures for the other datasets can be found in the Appendix (Figures C.6, C.7, C.8 and C.9), with the results being consistent to the results of the Compas dataset.

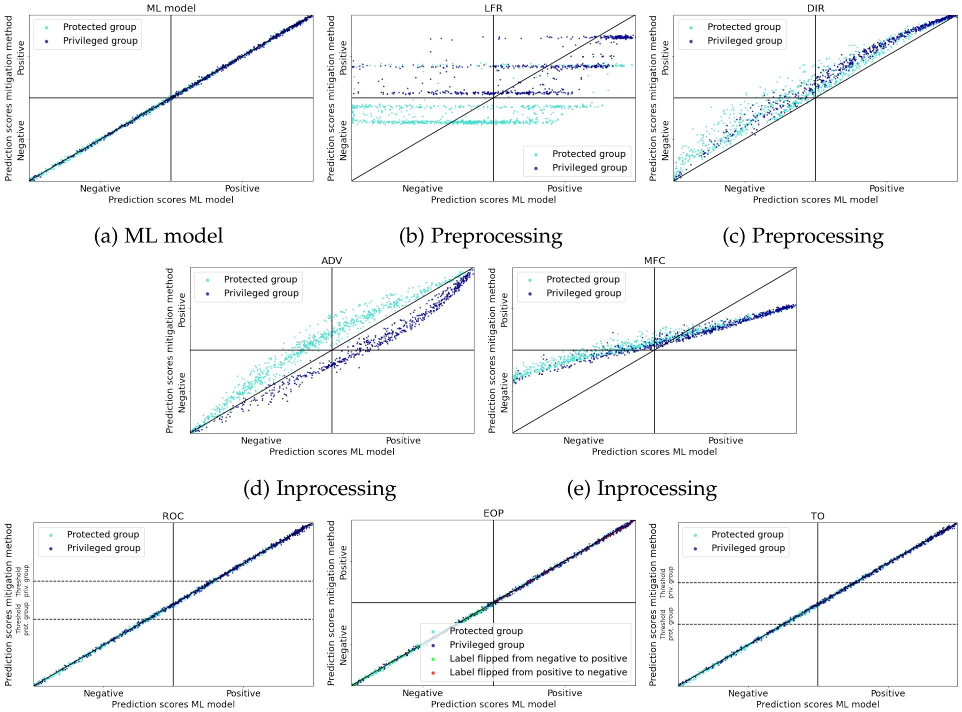
Insights into the operational dynamics of each bias mitigation method

In Figure 7.2a, we observe the prediction scores of the initial ML model. By default, machine learning models use a classification threshold of 0.5, categorizing instances above this threshold as positive and those below it as negative. This behavior remains consistent for both the protected and privileged group.

In Figures 7.2b-7.2e, we can assess the relation between the prediction scores of the initial ML model and the prediction scores after applying preprocessing (LFR and DIR) or inprocessing (ADV and MFC) bias mitigation methods. Figure 7.2b reveals that LFR significantly alters the prediction scores, in a seemingly random way. Instances that receive a positive label post-LFR are notably different from those in the initial score ranking. This suggests a substantial transformation in the feature space due to the preprocessing method. We will assess how this affects the performance and fairness metrics in Table 7.3. As shown in Figures 7.2c, 7.2d, and 7.2e, the methods DIR, ADV, and MFC also introduce significant alterations to the prediction scores, and thus the inherent ranking of instances. However, the altered prediction scores align more closely with the initial prediction scores compared to those obtained using LFR.

Figure 7.2 also provides us with insights into the fundamental distinction among various bias mitigation methods—specifically, whether they impact the prediction scores or only the

[8] Note that we use the custom threshold of 0.5 here to generate the labels, based on the prediction scores. However, as argued by Scantamburlo et al. [2024], this does not take into account the real constraints of decision-making context.



(a) ML model

(b) Preprocessing

(c) Preprocessing

(d) Inprocessing

(e) Inprocessing

(f) Postprocessing

(g) Postprocessing

(h) Postprocessing

Figure 7.2: Score distributions for the Compas dataset. The x-axis represents the prediction scores of the initial ML model, while the y-axis represents the prediction score after applying each bias mitigation method. The second quadrant represents the instances that are ‘upgraded’ by the bias mitigation method (initially predicted as negative, and after using the bias mitigation method predicted as positive), while the fourth quadrant represents the instances that are ‘downgraded’ by the bias mitigation method (initially predicted as positive, and after the bias mitigation method predicted as negative).

prediction labels. Unlike preprocessing and inprocessing methods, the used postprocessing methods do not alter the inherent ranking; instead, they adjust labels based on the prediction scores of the initial machine learning model. Each method implements a unique strategy for label flipping, all with the objective of meeting specified fairness metrics. For instance, ROC (Figure 7.2f) targets the most uncertain instances for label flipping, TO (Figure 7.2h) establishes distinct thresholds for each group, while EOP (Figure 7.2g) determines the ideal quantity of labels to flip within each group and executes these flips randomly. In the scenarios of ROC and TO, group-specific new thresholds are determined (as illustrated in Figures 7.2f and 7.2h). Regarding EOP, label flipping occurs randomly among negatively classified individuals in the protected group and positively classified individuals in the privileged group (as shown in green and red in Figure 7.2g). In this case, the initial ranking impacts the initial labels, but will have no influence on the decision regarding which labels are flipped.

Our examination provides insights into the specific instances affected by each bias mitigation method, in response to the work of Krco et al. [2023]. Take, for instance, the case of ROC: the instances that are flipped from a negative classification to a positive classification, are those instances that were initially classified as negative but that posses the highest prediction score according to the initial machine learning model.

How similar are the score distributions?

Table 7.2: Table with the similarity between the initial ranking produced by the machine learning model and the ranking produced by the model after using a bias mitigation method, measured by the Kendall-Tau statistic. The values between brackets present the similarity in ranking for the protected group and privileged group respectively.

Dataset	LFR	DIR	ADV	MFC	ROC-EOP-TO
Adult	0.318 (0.002, 0.034)	0.770 (0.726, 0.761)	0.729 (0.780,0.812)	0.638 (0.640, 0.663)	1
Compas	0.408 (0.153,0.534)	0.917 (0.883, 0.960)	0.805 (0.926, 0.946)	0.909 (0.900, 0.935)	1
Dutch	0.361 (-0.007, 0-0.004)	0.992 (0.991, 0.991)	0.808 (0.812, 0.913)	0.819 (0.850, 0.786)	1
Law	0.570 (0.622, 0.523)	0.870 (0.882, 0.858)	0.908 (0.896, 0.907)	-0.758 (-0.872, -0.737)	1
Student	0.215 (0.248, 0.216)	0.890 (0.891, 0.891)	0.515 (0.731, 0.678)	0.775 (0.843, 0.794)	1

Instead of visualizing how the prediction scores of the initial ML relate to the prediction scores after each bias mitigation method, we can also calculate the overlap between the two rankings by utilizing the Kendall-Tau statistic [Kendall, 1948]. The outcomes are presented in Table 7.2. Table 7.2 illustrates that the ranking produced by LFR consistently presents the least similarity with the ranking produced by the initial machine learning model. In contrast, the use of DIR leads to a significantly higher degree of overlap in rankings across all datasets. The similarity in rankings when applying ADV and MFC falls into a more moderate category. These observations highlight the varying degrees of impact each method exerts on the prediction scores and thus the ranking of individuals. With postprocessing methods, the ranking remains identical to the ranking of the initial model, resulting in a ranking similarity of 1 (refer to Table 7.2), and the AUC remains consistent with the initial model (see Table 7.3). Our analysis reveals that across all datasets, the prediction scores and consequently the intrinsic rankings are substantially modified by every preprocessing and inprocessing bias mitigation method we evaluated.[9] While such modifications are not necessarily worrisome, it is crucial to assess whether they enhance or worsen the ranking, and

[9] We see that in one case (MFC for the Law dataset), the scores are even completely shuffled in the opposite way, which shows how random some of the effects can be.

to what extent. It also does not feel fair from the viewpoint of the individual that they are suddenly downweighted by a random intervention. [10]

7.4.2 *Can we compare the bias mitigation methods?*

Table 7.3: Results of the bias mitigation strategies on the five datasets. We report: the AUC score overall, and split over the protected group and the privileged group, the accuracy (ACC), statical parity difference (SPD), equal opportunity difference (EOD) and the positive decision rate (PDR). Best values are highlighted in bold.

Dataset	Metric	ML model	LFR	DIR	ADV	MFC	ROC	EOP	TO
Adult	AUC	0.843	0.623	**0.85**	0.847	0.826	0.843	0.843	0.843
	AUC^{pro}, AUC^{pri}	0.811, 0.826	0.501, 0.508	0.824, 0.835	**0.834, 0.843**	0.789, 0.816	0.811, 0.826	0.811, 0.826	0.811, 0.826
	ACC	0.806	0.766	**0.825**	0.82	0.81	0.728	0.77	0.793
	SPD	-0.26	**-0.005**	-0.164	-0.09	-0.139	-0.055	-0.046	**-0.005**
	EOD	-0.139	**0.0**	-0.181	-0.274	-0.278	-0.332	-0.283	-0.368
	PDR	0.247	0.005	0.174	0.161	0.178	0.393	0.151	0.167
Compas	AUC	**0.834**	0.693	0.833	0.808	0.832	0.834	0.834	0.834
	AUC^{pro}, AUC^{pri}	**0.814**, 0.821	0.588, 0.703	0.81, **0.823**	0.809, 0.815	0.812, 0.821	0.814, 0.821	0.814, 0.821	0.814, 0.821
	ACC	**0.758**	0.645	0.753	0.737	0.736	0.732	0.683	0.727
	SPD	-0.376	-0.866	-0.332	-0.064	-0.211	-0.031	-0.07	**0.014**
	EOD	-0.13	**-0.055**	-0.131	-0.231	-0.175	-0.233	-0.25	-0.262
	PDR	0.518	0.49	0.617	0.58	0.696	0.528	0.545	0.569
Dutch	AUC	**0.887**	0.657	**0.887**	0.874	0.883	0.887	0.887	0.887
	AUC^{pro}, AUC^{pri}	0.884, 0.848	0.499, 0.498	0.884, 0.849	0.881, 0.847	**0.89, 0.852**	0.884, 0.848	0.884, 0.848	0.884, 0.848
	ACC	**0.812**	0.476	0.786	0.768	0.695	0.776	0.754	0.763
	SPD	-0.318	**0.0**	-0.432	-0.171	-0.394	-0.066	-0.159	-0.02
	EOD	-0.026	-0.315	-0.047	-0.073	**-0.024**	-0.217	-0.243	-0.254
	PDR	0.416	1.0	0.586	0.303	0.203	0.395	0.45	0.396
Law	AUC	0.882	0.83	**0.883**	0.879	0.122	0.882	0.882	0.882
	AUC^{pro}, AUC^{pri}	0.848, **0.864**	0.803, 0.792	**0.853, 0.864**	0.843, 0.862	0.146, 0.142	0.848, 0.864	0.848, 0.864	0.848, 0.864
	ACC	**0.903**	0.897	**0.903**	0.901	0.22	0.772	0.879	0.892
	SPD	-0.197	-0.207	-0.184	-0.141	0.494	-0.044	-0.021	**-0.002**
	EOD	-0.111	-0.128	-0.122	-0.143	-0.151	**-0.104**	-0.197	-0.21
	PDR	0.954	0.967	0.961	0.965	0.289	0.711	0.961	0.979
Student	AUC	0.803	0.693	**0.817**	0.772	0.762	0.803	0.803	0.803
	AUC^{pro}, AUC^{pri}	**0.819,** 0.788	0.689, 0.642	**0.819, 0.814**	0.757, 0.779	0.781, 0.794	0.819, 0.788	0.819, 0.788	0.819, 0.788
	ACC	**0.759**	0.626	**0.759**	0.703	0.738	0.697	0.728	0.728
	SPD	-0.104	-0.72	-0.065	-0.573	0.031	0.076	**-0.016**	0.071
	EOD	-0.166	-0.063	-0.202	**0.027**	-0.283	-0.191	-0.205	-0.246
	PDR	0.585	0.697	0.574	0.662	0.677	0.492	0.595	0.605

We present the results of the performance and fairness metrics for all bias mitigation methods across five datasets in Table 7.3. Aligning with existing literature, we observe that fairness metrics such as SPD and EOD often yield conflicting results [Kleinberg et al., 2016]. It is rarely the same method that returns the best results for both metrics. Furthermore, no single method consistently outperforms all others for one of the metrics. However, DIR stands out for its excellent AUC performance, aligning with its design to preserve rank-ordering within groups [Feldman et al., 2015].

[10] So far, we emphasized that only preprocessing and inprocessing methods suffer from this reranking process. However, the postprocessing method EOP also significantly suffers from this arbitrariness, as it will execute random flips within each group until a fairness metric is satisfied. This means that the impacted individuals can be different in each run, and can be individuals with a very high prediction score can be 'downweighted' and individuals with a very low prediction score can be 'uplifted' (as the prediction scores are not taken into account to determine who should be flipped).

Can we use the prediction labels to compare the methods?

When comparing bias mitigation methods, conventional benchmarking studies often focus on accuracy and fairness [Chen et al., 2023, Zemel et al., 2013]. However, relying solely on accuracy, which uses prediction labels at a specific threshold, may lead to inappropriate comparisons. As mentioned, there are two reasons why using the prediction labels is not a suitable approach: First of all, the fairness metrics are not yet satisfied, and secondly, different mitigation methods can result in varying positive rates. Both reasons would lead to an additional altering of the prediction labels post-hoc, so comparing them at this stage does not seem sensible. A more comprehensive approach involves assessing the performance of the prediction scores, or comparing the prediction labels when the thresholds have been modified to address both industry constraints and fairness considerations.

Addressing the first concern, we see that in the large majority of cases, bias mitigation methods fail to satisfy a fairness metric, which is confirmed in literature [Chen et al., 2023]. Particularly, only postprocessing strategies tend to show a high success rate in achieving the optimized fairness metric, in contrast to the preprocessing and inprocessing methods. As highlighted in other benchmark studies [Chen et al., 2023], we also note that employing bias mitigation methods can even lead to situations that are **more** unfair in terms of disparities between groups.

Regarding the second concern, Table 7.3 reveals significant positive decision rate disparities among different strategies. This inconsistency poses challenges in real-world applications, where a fixed or reasonably bounded positive decision rate is typically expected [Kwegyir-Aggrey et al., 2023]. For example, when attempting to satisfy one of the fairness metrics, practitioners might consider LFR in the Adult Income dataset, despite a slight accuracy loss. However, this choice results in an unexpectedly low positive decision rate (0.5%), deviating significantly from other methods. This method just predicts almost every instance as negative, which results in a satisfaction of the fairness metrics, but is not realistic in the real-word. Similarly, for the Dutch dataset, MFC leads to the best value for EOD, but it only has a positive decision rate of 20.3%, while the initial model has a positive decision rate of 41.6%. DIR leads to the best AUC, but has a positive decision rate of 58.6%. We argue against treating methods with significantly different positive decision rates as comparable situations. In practice, most real-world applications will have a relatively fixed positive decision rate, and bias mitigation methods must be adapted accordingly.

Evaluate by using the prediction scores

Both these concerns can effectively be addressed by adjusting the classification threshold(s) of the predictive model. Consequently, it makes more sense to evaluate these mitigation methods based on their prediction scores, rather than on their prediction labels, which are still subject to change. We advocate to compare these mitigation methods based on the AUC score, to assess how well the individuals are ranked within each group, and to generate the prediction labels post-hoc, based on the chosen fairness metric and practical constraints.

It is important not only to examine the overall AUC score across the entire population, but also the AUC scores disaggregated by different groups. Certain bias mitigation methods may yield the optimal ranking for one group but not for another. This prompts the need for decision-making: should preference be given to the overall best ranking or to narrowing the gap in rankings between the protected and privileged groups? Alternatively, deploying two separate models could be considered to ensure the best ranking for each subgroup. Given these results, one should evaluate whether a slight improvement in AUC justifies the adoption of distinct models. Additionally, note that for both the Adult and the Dutch datasets, the LFR method reduces the within-group ranking to an essentially random ordering. In any scenario, deploying a bias mitigation method that decreases the AUC score for every subgroup seems counterproductive when we want to optimize for both performance and fairness.[11] Unfortunately, this is a common outcome in practice, as many bias mitigation methods may compromise AUC scores across subgroups in an attempt to adhere to a fairness metric [Mittelstadt et al., 2023].

Note that the goal of this comparison is not to declare the superior performance of one of the mitigation methods. For this, a more comprehensive benchmark study is needed with a larger number of datasets, machine learning models, and extensive tuning of each bias mitigation method. This was already the goal of multiple other benchmark studies [Chen et al., 2023, Hort et al., 2021, Reddy, 2022]. Our primary goal was to emphasize the inadequacy of the current way of benchmarking these methods with each other, based on prediction labels that should still be subject to change, and to highlight the effects within each group.

7.5 DISCUSSION

In this study we want to emphasize two points. First, that bias mitigation methods not only introduce significant changes between-groups but also within-groups and that these changes currently go unnoticed. This reranking process is not necessarily an issue, but deserves more attention. Secondly, we argue that the current approach to comparing bias mitigation methods is insufficient due to the sole focus on the output of the prediction model without taking into account the decision-making system [Scantamburlo et al., 2024]. This will lead to a comparison of disparate scenarios, as we see that the positive decision rate varies widely.

Regarding the first point, is this reranking process necessarily a bad thing? We notice in Table 7.3 that the bias mitigation methods can result in better, worse or approximately the same ranking accuracy (measured by the AUC) as the initial machine learning model. If the ranking improves after using the bias mitigation method, there is no issue, as the method will lead to better rankings that should also be more fair. But what if the ranking accuracy is significantly worse than the the output of the initial machine learning model? Is this always undesirable? Not per se. Up until this point, and in accordance with most papers in fair machine learning, we assumed that we only want to remove bias between groups. However, some of this bias might also seep in the ranking within each group, for example by favoring individuals from the protected group that resemble individuals from the privileged group. If we assume this within-group bias is also present and unwanted, the produced rankings of the

[11] Note that we operate under the assumption that there is no label bias.

mitigation methods might be worse with respect to the target label, but better in avoiding this within-group bias.

If we assume that there is no within-group bias, we propose that for the dual objectives of fairness and accuracy optimization, preprocessing and inprocessing mitigation methods should be adopted only if they improve subgroup rankings. Should they fail to do so, we posit that utilizing the ranking produced by the original machine learning model and then applying fairness interventions post-hoc may be a more effective strategy. However, this practice is not always possible, as this can lead to situations where two otherwise identical entities are treated differently based solely on a sensitive attribute, a practice that may be unlawful in certain contexts. Furthermore, we can not always assume that the decision-making body has access to the sensitive attribute, and thus is able to do these fairness interventions post-hoc. In those settings, it might be better to use preprocessing or inprocessing bias mitigation methods, even if they worsen this trade-off.

What if the performance of the ranking stays approximately the same, but the ranking itself will be significantly different from the ranking of the initial machine learning model (as measured in Table 7.2)? Is this arbitrariness a problem? The opinions on this differ. The literature on 'predictive multiplicity' [Marx et al., 2020] or 'model multiplicity' [Black et al., 2022] discusses the situation where there exist many possible models with similar predictive performance but slightly different decisions on individuals, which is comparable to our situation. They argue that multiplicity should be reduced by removing some of the variance that leads to diverging predictions [Cooper et al., 2023], while Jain et al. [2024] highlight how fairly allocating scarce resources using machine learning could benefit from randomness. As we can see, there is no easy answer to this question, and we look forward to more debate about this topic.

A first limitation of our study is the assumption of no label bias. This leads us to quite a straddle, which is common in fair machine learning. On the one hand, we presume that our labels are correct to ensure the reliability of our metrics. On the other hand, we recognize that models might need adjustments to correct for bias. This make sense for equal opportunity, where the goal is to ensure that all groups have equal true positive rates, thus requiring a classifier that does not make the situation worse. For demographic parity, the issue is more complex: why would we want to deviate from the labels if they are correct? Nonetheless, adjustments may be necessary to align with external policy or legal requirements.

A second limitation of our study is that it operates under the assumption of having access to a static sensitive attribute, and it will face difficulties in scenarios where these assumptions may not hold. For instance, the assumption of the sensitive attribute being static overlooks the evolving nature of certain attributes, such as gender, where people can identify as other categories over time. Another limitation arises when considering the assumption of unrestricted access to the sensitive attribute. In reality, legal and ethical considerations may impose constraints on obtaining or utilizing certain sensitive information [Haeri and Zweig, 2020, Veale and Binns, 2017, Johnson, 2021, Holstein et al., 2019]. For instance, privacy laws and regulations may restrict the collection or use of specific attributes, further complicating research in fairness.

Part IV

PRIVACY

8

The Privacy Issue of Counterfactual Explanations

Black-box machine learning models are used in an increasing number of high-stakes domains, and this creates a growing need for Explainable AI (XAI). However, the use of XAI in machine learning introduces privacy risks, which currently remain largely unnoticed. Therefore, we explore the possibility of an *explanation linkage attack*, which can occur when deploying instance-based strategies to find counterfactual explanations. To counter such an attack, we propose *k*-anonymous counterfactual explanations and introduce *pureness* as a metric to evaluate the *validity* of these *k*-anonymous counterfactual explanations. Our results show that making the explanations, rather than the whole dataset, *k*-anonymous, is beneficial for the quality of the explanations.

8.1 INTRODUCTION

Black-box models are used for decisions in more and more high-stakes domains such as finance, healthcare and justice, increasing the need to explain these decisions and to make sure that they are aligned with how we want the decisions to be made [Molnar, 2020, Goethals et al., 2022]. As a result, the interest in interpretability methods for machine learning and the development of various techniques has soared [Molnar, 2020]. At the moment, however, there is no consensus on which technique is best for which specific use case. Within the field of Explainable AI (XAI), we focus on a popular local explanation technique: counterfactual explanations [Martens and Provost, 2014, Wachter et al., 2017b].

Counterfactual explanations, which are used to explain predictions of individual instances, are defined as a change to the feature values of an instance that alters its prediction [Martens and Provost, 2014, Molnar, 2020].[1] *Factual instances* are the original instances that are explained and the *counterfactual instance* is the original instance with the updated values from the explanation. An example of a *factual instance*, *counterfactual instance* and *counterfactual explanation* for a credit scoring context can be seen in Figure 8.1. *Lisa* is the *factual instance* here, whose credit gets rejected. *Fiona*, a nearby instance in the training set whose credit was accepted, is selected as *counterfactual instance* by the algorithm and based on *Fiona*, *Lisa* receives a *counterfactual explanation* that states which features to change to receive a positive credit decision. These explanations can serve multiple objectives: they can be used for model debugging by data scientists or model experts or to justify decisions to end users [Aïvodji et al., 2020, Molnar, 2020, Martens, 2022]. For a complete overview of potential objectives of counterfactual explanations, we refer you to Section 3.2.3. We will start from the set-up where we assume the counterfactual explanations are used to give actionable recourse. However, other use cases of counterfactual explanations (model debugging, fairness audits, ..) can also lead to privacy issues.

Note that in this set-up, we assume the counterfactual explanations are used to give actionable recourse. However, the use cases of counterfactual explanations are a lot broader as they can also be used to gain general insights into the model, model debugging, fairness audits, etc. I refer to Section 3.2 for a complete overview. Note however that in a lot of these settings, the explanations go to the model owners and not to external parties. We see the setting where the explanations are given to external parties as the ones with the highest privacy risk.

Factual instance						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Lisa	***21***	*F*	***Brussels***	***$50K***	*Single*	*Reject*

Counterfactual explanation=
If you would be **three years older**, lived in **Antwerp** and your income would be **$10K** higher, you would have received a positive credit decision

[1] Depending on the used objective function, counterfactual explanations can be generated to optimize for diversity, proximity, plausibility, actionability, sparsity,... [Verma et al., 2020]. The different optimization emtrics will result in different counterfactual explanations.

Counterfactual instance						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Fiona	***24***	*F*	***Antwerp***	***$60K***	*Single*	*Accept*

Figure 8.1: Example of a counterfactual explanation

At the same time, there is a growing concern about the potential privacy risks of machine learning [Liu et al., 2021]. Privacy is recognized as a human right and defined by Oxford Dictionary as *"a state of being free from the attention of the public"*.[2] In a privacy attack, the goal of an adversary is to gain knowledge that was not intended to be shared [Liu et al., 2021, Rigaki and Garcia, 2023]. Different kinds of privacy attacks exist: both the training data, where the adversary tries to infer membership in a *membership inference attack* or specific attributes of an input sample in an *attribute inference attack*, as well as the model, in a *model extraction attack*, can be the target [Fredrikson et al., 2015, Rigaki and Garcia, 2023, Liu et al., 2022]. These attacks are described in more detail in Section 3.4.

Unfortunately, there exists an inherent tension between explainability and privacy as the usage of Explainable AI can increase these privacy risks [Aïvodji et al., 2020]: model explanations offer users information about how the model made a decision about their data instance. Consequently, they leak information about the model and the data instances that were used to train the model. Earlier research already shows that explanations can provide ground for membership inference attacks, where is determined whether a given instance is part of the training data, [Naretto et al., 2022, Shokri et al., 2020, Quan et al., 2022, Pawelczyk et al., 2023] and model extraction attacks, where information about the functionality of the model is collected through query access [Aïvodji et al., 2020, Quan et al., 2022]. In this paper, we introduce a new kind of privacy attack based on counterfactual explanations and we call this an *explanation linkage attack*. A *linkage attack* attempts to identify anonymized individuals by combining the data with background information. An *explanation* linkage attack attempts to link the counterfactual explanation with background information to identify the counterfactual instance. We illustrate an example of an explanation linkage attack in Section 8.2. Unfortunately, the introduction of these attacks indicates that an attempt to make an AI system safer by making it more transparent can have the opposite effect [Sokol and Flach, 2019]. Other researchers [Shokri et al., 2019, Budig et al., 2021, Patel et al., 2022] also confirm the trade-off between privacy and explainability and emphasize that assessing this trade-off for minority groups is an important direction for future research [Patel et al., 2022].

Our contributions are as follows:

- We introduce a new kind of privacy attack, the *explanation linkage attack*, that are based on real instances.
- As a solution for this problem, we propose k-anonymous counterfactual explanations and develop an algorithm to generate these.

[2] https://www.oxfordlearnersdictionaries.com/definition/american_english/privacy

- We evaluate how k-anonymizing the counterfactual explanations influences the quality of these explanations, and introduce *pureness* as a new metric to evaluate the validity of these explanations.

- We show the trade-off between *transparency*, *fairness* and *privacy* when using k-anonymous explanations: when we add more privacy constraints, the quality of the explanations and therefore the transparency decreases. This effect on the explanation quality is larger for minority groups, as they tend to be harder to anonymize, and this can have an impact on fairness.

8.2 PROBLEM STATEMENT: EXPLANATION LINKAGE ATTACKS

We introduce the privacy problem of counterfactual explanations that are based on real instances, and illustrate this problem by using a simple toy dataset. This dataset contains individuals that are described by a set of identifiers, quasi-identifiers and private attributes [Sweeney, 2002b]. Identifiers are attributes such as name, phone or social security number and need to be suppressed in any case as they often do not have predictive value and can uniquely identify a person. Quasi-identifiers are attributes such as age, zip code or gender that can hold some predictive value. They are assumed to be public information; however, even though they cannot uniquely identify a person, their combination might. It has been shown that 87% of US citizens can be re-identified by the combination of their zip code, gender and date of birth [Sweeney, 2000]. Private attributes are attributes that are not publicly known, and are meant to be kept confidential.

Let us briefly discuss the set-up of this attack: We assume that the adversary has access to the identifiers and quasi-identifiers of everyone in the the dataset. [3] In line with the literature, we look at the following two re-identification scenarios for a single individual [El Emam and Dankar, 2008, Elliot and Dale, 1999, Marsh et al., 1991]:

- Re-identification of a specific individual (*prosecutor re-identification scenario*): The adversary (e.g., a prosecutor) knows that a specific individual is part of the dataset, and wants to infer its private information.

- Re-identification of an arbitrary individual (*journal re-identification scenario*). The adversary (e.g., a journalist) does not care which individual is being re-identified but only wants to prove that it can be done.

If the attacker wants to execute one of the scenarios above and gets access to the private attributes of a user in the dataset, a possible avenue to achieve this is by asking for counterfactual explanations. The counterfactual explanation will never contain identifiers but if it contains a combination of quasi-identifiers that can uniquely identify a person, the attacker can deduce the person's private attributes. We name this kind of attack an *explanation linkage attack*.

[3] People's quasi-identifiers are often rather easy to be obtained by the public as lists like voter records are publicly available [Sweeney, 2000, Machanavajjhala et al., 2007].

Assume the following factual instance *Lisa* in Table 8.1:

Table 8.1: Factual instance *Lisa*

Identifier	**Quasi-identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Lisa	*21*	*F*	*Brussels*	*$50K*	*Single*	*Reject*

Name is the identifier that is deleted from the dataset, but, as mentioned, people can often still be identified by their unique combination of quasi-identifiers. *Age*, *Gender* and *City* are the quasi-identifiers in this dataset that are assumed to be public knowledge for every adversary. A possible reasoning behind this, is that the adversary acquired access to a voter registration list as in Sweeney [2000]. *Salary* and *Relationship* are private attributes that one does not want to be public information, and the target attribute in this dataset is whether the individual will be awarded credit or not. *Lisa* is predicted by the machine learning model as not creditworthy and her credit gets rejected. Logically, *Lisa* wants to know the easiest way to get her credit application accepted, so she asks for a *counterfactual explanation*, the smallest change to her feature values that result in a different prediction outcome.

Table 8.2: Training set

Identifier	**Quasi-identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
Alfred	*25*	*M*	*Brussels*	*$50K*	*Single*	*Reject*
Boris	*23*	*M*	*Antwerp*	*$40K*	*Separated*	*Reject*
Casper	*34*	*M*	*Brussels*	*$30K*	*Cohabiting*	*Reject*
Derek	*47*	*M*	*Antwerp*	*$100K*	*Married*	*Accept*
Edward	*70*	*M*	*Brussels*	*$90K*	*Single*	*Accept*
Fiona	*24*	*F*	*Antwerp*	*$60K*	*Single*	*Accept*
Gina	*27*	*F*	*Antwerp*	*$80K*	*Married*	*Accept*
Hilda	*38*	*F*	*Brussels*	*$60K*	*Widowed*	*Reject*
Ingrid	*26*	*F*	*Antwerp*	*$60K*	*Single*	*Reject*
Jade	*50*	*F*	*Brussels*	*$100K*	*Married*	*Accept*

In our set-up, the counterfactual algorithm looks for the instance in the training set that is nearest to *Lisa* and has a different prediction outcome (the *nearest unlike neighbor)*. The training set, with the nearest unlike neighbor highlighted, is shown in Table 8.2. *Fiona* has similar attribute values as *Lisa*, but is 24 years old instead of 21, lives in Antwerp instead of Brussels and earns $60K instead of $50K. When *Fiona* is used as counterfactual instance by the explanation algorithm, *Lisa* would receive the explanation: *'If you would be 3 years older, lived in Antwerp and your income was $10K higher, then you would have received the loan'*. Based on her combined knowledge of the explanation and her own attributes, *Lisa* can now deduce that *Fiona* is the counterfactual instance, as there is only one person in this dataset with this combination of quasi-identifiers (a 24-year old woman living in Antwerp). Therefore, *Lisa* can deduce the private attributes of *Fiona*, namely *Fiona*'s income and relationship status, which is undesirable.

Obviously, this is just a toy example, but we envision many real-world settings where this situation could occur. For instance, when end users receive a negative decision, made by a *high-risk AI system*: these systems are defined by the EU's AI Act, which categorizes the risk of AI systems usage into four levels [European Commission, 2021]. Among others, they include employment, educational training, law enforcement, migration and essential public services such as credit scoring. Article 13(1) states: *"High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately."* These systems are therefore obliged to provide some form of transparency and guidance to its users, which could be done by providing counterfactual explanations or any other transparency technique. Most of these settings use private attributes as input for their decisions, so it is important to make sure that the used transparency techniques do not reveal private information about other decision subjects. For example, in decisions about educational training or employment, someone's grades could be revealed, or in credit scoring, the income of other decision subjects could be disclosed.

This privacy risk occurs when the counterfactual algorithm uses instance-based strategies to find the counterfactual explanations. These counterfactuals correspond to the *nearest unlike neighbor* and are also called *native counterfactuals* [Brughmans et al., 2023a, Keane and Smyth, 2020]. Other counterfactual algorithms use perturbation where synthetic counterfactuals are generated by perturbing the factual instance and labelling it with the machine learning model, without reference to known cases in the training set [Keane and Smyth, 2020]. We focus on counterfactual algorithms that return real instances: several algorithms do this, as this substantially decreases the run time while also increasing desirable properties of the explanations such as plausibility [Brughmans et al., 2023a]. Plausibility measures how realistic the counterfactual explanation is with respect to the data manifold, which is a desirable property. Guidotti [2022], and Brughmans et al. [2023a] show that the techniques resulting in an actual instance have the best plausibility results. Furthermore, it is argued that counterfactual instances that are plausible, are more robust and therefore are less vulnerable to the uncertainty of the classification model or changes over time [Artelt et al., 2021, Brughmans et al., 2023a, Pawelczyk et al., 2020]. This shows that for some use cases it can be very useful to use real data points as counterfactuals instead of synthetic ones as for the latter the risk of generating implausible counterfactual explanations can be quite high [Laugel et al., 2019]. Algorithms that use these *native* counterfactual explanations include NICE (without optimization setting) [Brughmans et al., 2023a], the WIT tool with NNCE [Wexler et al., 2019], FACE [Poyiadzi et al., 2020] and certain settings of CBR [Keane and Smyth, 2020]. Perturbation-based counterfactual algorithms experience different privacy risks such as membership inference attacks: Pawelczyk et al. [2023] use counterfactual distance-based attacks which leverage algorithmic recourse to determine if an instance belongs to the training data of the underlying model or not. We envisage a different scenario, where the adversary knows which instances are in the training data, but wants to gain access to its private attributes. It is worth emphasizing that some perturbation-based counterfactual algorithms could still have some vulnerability to explanation linkage attacks, although arguably less likely than native counterfactuals. Some perturbation algorithms (such as NICE with optimization settings) start from a real counterfactual instance in the dataset, and it is possible they will return the real instance without perturbations. In many cases, the instance will only be slightly perturbed, so that an ingenious adversary can still have high confidence about the private attribute values of the counterfactual instances. Even when the counterfactual algorithm does not start from a real instance, but uses plausibility with respect to the data manifold to generate its explanations, the explanation algorithm could still return a (or seemingly) real instance. Our algorithm could also offer protection in those cases, by ensuring that the returned instance is at least k-anonymous.

8.3 proposed solution

As a solution, we propose to make the counterfactual explanations *k-anonymous*. *k-anonymity* is a property that captures the protection of released data against possible re-identification by stating that the released data should be indistinguishable between k data subjects [Van Tilborg and Jajodia, 2014].

8.3.1 *What is k-anonymity?*

Before k-anonymity was introduced, data that looked anonymous was often freely shared after removing explicit identifiers such as name and address, incorrectly believing that individuals in those datasets could not be identified. Contrary to these beliefs, we have seen that people can often be identified through their unique combination of quasi-identifiers.

Consider a database that holds private information about individuals, where each individual is described by a set of identifiers, quasi-identifiers, and private attributes. k-anonymity characterises the degree of privacy, where the information for each person in the dataset cannot be distinguished from at least $k-1$ other individuals whose information was also released [Sweeney, 2002a]. A group of individuals that cannot be distinguished from each other and thus have the same values of quasi-identifiers are named an *equivalence class*.

Usually k-anonymity is applied on the whole dataset: the quasi-identifiers of the data records are suppressed or generalised in such a way that one record is not distinguishable from at least $k-1$ other data records in that dataset [Meyerson and Williams, 2004]. In this way, the privacy of individuals is protected to some extent by *"hiding in the crowd"* as private data can now only be linked to a set of individuals of at least size k [Gionis and Tassa, 2008]. However, by generalising or suppressing attribute values, the data becomes less useful, so the problem studied is to make a dataset k-anonymous with minimal loss of information [Gionis and Tassa, 2008, Xu et al., 2006a]. We will measure the loss in information value with the *Normalized Certainty Penalty* (NCP) and explain this metric in Section 8.4.

8.3.2 *Application to our problem*

Our application differs from the original set-up of k-anonymity as it specifically aims to ensure anonymity in counterfactual explanations, rather than anonymizing the entire dataset. While the original application is suitable for situations where the entire dataset is publicly

[1] LeFevre et al. [2006]
[2] Sweeney [2002a]

Table 8.3: Comparison between the original problem setting of *k*-anonymity and our problem setting.

	***k*-anonymity**	
	Dataset	**Counterfactual explanation**
Input	Dataset	Dataset Factual instance Counterfactual explanation Machine learning model
Defined over	Dataset	Counterfactual explanation
Method	Mondrian[4], Datafly[5],..	CF-K
Risk	Identifying instances in the dataset based on their combination of quasi-identifiers and inferring their private attributes	Identifying the counterfactual instance based on its combination of quasi-identifiers and inferring its private attributes
Evaluation metrics	Degree of privacy Information loss	Degree of privacy Information loss Counterfactual validity

accessible. We highlight this difference in Table 8.3. A counterfactual instance is defined as *k*-anonymous if the combination of quasi-identifiers can belong to at least *k* individuals in the training set, and consequently, a counterfactual explanation is defined as *k*-anonymous if the counterfactual instance on which it is based, has a combination of quasi-identifiers that can belong to at least *k* individuals in the training set. We implement this by looking for close neighbours of *Fiona*, that have similar values of quasi-identifiers, and that also have the desired prediction outcome. In this case, the closest neighbor to *Fiona* that has the desired prediction outcome is *Gina*, as can be seen in Table 8.2. Next, we generalise the quasi-identifiers of the counterfactual instance so that they can belong to both the counterfactual instance and the neighbour, resulting in a counterfactual instance that is at least 2-anonymous (see Figure 8.2.) However, by doing so we degrade the quality of the data as we will see in Section 8.4.

The *k*-anonymous counterfactual explanation based on the *k*-anonymous counterfactual instance in Figure 8.2 and factual instance *Lisa* (21, F, Brussels, $50K, Single) is: *'If you would be 3-6 years older, lived in Antwerp and had an income of $60K, you would have received the loan'*. This explanation is 3-anonymous because the combination of quasi-identifiers in the counterfactual instance (24-27, F, Antwerp) could point to at least three instances in the training set in Table 8.2, namely *Fiona*, *Gina* and *Ingrid*.

However, the fact that other instances than the ones explicitly used to generate the *k*-anonymous counterfactual explanation, might also fall in the range of the explanation, introduces a new issue that is specific to *k*-anonymous counterfactual explanations. Counterfactual explanations are defined as the smallest change to the feature values of an instance that alter its prediction outcome, but does this still hold for *k*-anonymous counterfactual explanations? We are no longer sure that all the value combinations in the *k*-anonymous counterfactual instance lead to a change in the prediction outcome and therefore we are not sure whether they are *valid* counterfactual explanations.

In this toy example, the value combination of *Ingrid* in Table 8.2 is also part of the *k*-anonymous counterfactual instance, as *Ingrid* is between 24 and 27 years old, female, single, living in

Counterfactual instance						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
*	*24*	*F*	*Antwerp*	*$60K*	*Single*	*Accept*

+

Neighbor						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
*	*27*	*F*	*Antwerp*	*$80K*	*Married*	*Accept*

⇓

K-**anonymous counterfactual instance**						
Identifier	**Quasi-Identifiers**			**Private attributes**		**Model prediction**
Name	Age	Gender	City	Salary	Relationship status	Credit decision
*	*24-27*	*F*	*Antwerp*	*$60K*	*Single*	*Accept*

Figure 8.2: How to generalize the counterfactual instance. As can be seen, we generalize only the values of the quasi-identifiers. The private attributes are still the same as in the original counterfactual instance as their attribute value is not public and therefore cannot be used to identify someone.

Antwerp and earning \$60K. However, the model predicts *Ingrids* credit decision to be rejected. A possible reasoning behind this could be because the model has learned that for higher age groups a higher income is required to be awarded the credit (or any other pattern). Therefore, if *Lisa* would follow-up the *"advice"* in the counterfactual explanation, it is possible that she would end up in this value combination, which does not result in an altering of the prediction outcome. This is problematic as this is one of the key objectives of counterfactual explanations.

This issue leads us to a new metric: how *valid* is the k-anonymous counterfactual explanation? We discuss the evaluation metrics further in Section 8.4.

8.4 EVALUATION METRICS

We measure the quality of the explanations by using the following metrics:

- The degree of privacy is measured by k: to how many instances from the training set can this counterfactual explanation be linked?
- The validity of the counterfactual explanations is measured by the *pureness*.
- The loss in information value is measured by the *Normalized Certainty Penalty (NCP)*.

We assess how the degree of privacy influences the loss in information value and the validity of the counterfactual explanations in Figure 8.4.

DEGREE OF PRIVACY We measure the degree of privacy by using the definition of k-anonymity. In our toy example, k is 3, as the generalised quasi-identifiers of the k-anonymous counterfactual instance could belong to three people when we look at the training set in Table 8.2 (*Fiona*, *Gina* and *Ingrid*). In our set-up, we will implement the degree of privacy as a minimum constraint in the algorithm.

COUNTERFACTUAL VALIDITY We define a possible value combination as a combination of attribute values that is in the range of the k-anonymous counterfactual instance. Note that we take into account all the attributes here, not only the quasi-identifiers. For a categorical attribute, we look at all the values present in the k-anonymous counterfactual instance. For a numerical attribute, we look at all the values that are in the range of the k-anonymous counterfactual instance *and* are also present in the training set. We illustrate these calculations in Table 8.4. The pureness of a k-anonymous counterfactual explanation can be calculated as follows:

$$\text{Pureness} = \frac{\text{\# of value combinations with desired prediction outcome}}{\text{\# of value combinations}}$$

The theoretical pureness is calculated on all the value combinations, but we will approximate this by querying the model with 100 random combinations[6] and see how many of these combinations lead to the desired prediction outcome. The pureness is the proportion of these value combinations that lead to the desired prediction outcome, which obviously should be as high as possible (preferably 100%).

Table 8.4: Possible value combinations and its model predictions.

Age	Gender	City	Salary	Relationship status	Model prediction
24	*F*	*Antwerp*	*$60K*	*Single*	*Accept*
25	*F*	*Antwerp*	*$60K*	*Single*	*Accept*
26	*F*	*Antwerp*	*$60K*	*Single*	*Reject*
27	*F*	*Antwerp*	*$60K*	*Single*	*Reject*

Table 8.4 shows all possible value combinations of the k-anonymous counterfactual instance, and the prediction outcome to each value combination. The goal of the counterfactual explanation was to alter the prediction outcome from *Reject* to *Accept*, so this is the desired prediction outcome. The k-anonymous counterfactual explanation in our toy example leads to the desired prediction outcome in 50% of the cases ($\frac{2}{4}$). If we sample 100 times out of the value combinations above, we expect this to approximate the theoretical pureness of 50%.

loss in information value When datasets are made k-anonymous, they tend to lose information. In general, excessive anonymization makes the data less useful because some analysis is no longer possible or the analysis provides biased and incorrect results [El Emam and Dankar, 2008].

A variety of metrics to measure information loss have been proposed, and we focus on the metrics discussed in Ghinita et al. [2007], which are Normalized Certainty Penalty (NCP) [Xu et al., 2006b], Discernibility metric (C_DM) [Bayardo and Agrawal, 2005] and the classification metric (CM) [Iyengar, 2002].

NCP penalises attributes for the way they are generalised and captures the uncertainty caused by this generalization [Xu et al., 2006b]. It assigns larger penalties when attribute values are mapped to generalised values that replace many other values [Loukides and Gkoulalas-Divanis, 2012]. An advantage of this metric is that it can give different weights to different attributes, as some attributes can be more important than others for the data analysis process [Xu et al., 2006b]. The NCP for each numerical (*Num*) quasi-identifier A in an equivalence class G is defined as:

$$\text{NCP}_{A_{Num}}(G) = \frac{\max^G_{A_{Num}} - \min^G_{A_{Num}}}{\max_{A_{Num}} - \min_{A_{Num}}}, \tag{8.1}$$

where the numerator and denominator represent the range of attribute A for the equivalence class G and for the whole dataset respectively [Ghinita et al., 2009]. This metric thus measures

[6] We chose for 100 random value combinations instead of trying out all the possibilities as the number of combinations can quickly become very large when there is a lot of generalization. The more random value combinations we test, the more we approximate the theoretical pureness, but the longer the computation time.

which part of the total range of the numerical attribute, is present in the equivalence class. Higher values signify more generalization, and consequently, more information loss. In the case of a categorical (*Cat*) quasi-identifier A, NCP is defined as follows:

$$\text{NCP}_{A_{Cat}}(G) = \begin{cases} 0, & \text{if } |A^G| = 1 \\ \frac{|A^G|}{|A|}, & \text{otherwise} \end{cases} \tag{8.2}$$

where $|A|$ is the number of distinct values of attribute A in the whole dataset, and $|A^G|$ is the number of distinct values of attribute A in equivalence class G [Xu et al., 2006b]. So, for a categorical attribute, this metric will check which proportion of possible unique values is present in the k-anonymous counterfactual instance. The higher this number is, the more generalized this attribute will be and the more information about this attribute is lost. The NCP of equivalence class G over all quasi-identifier attributes is:

$$NCP(G) = \sum_{i=1}^{d} w_i \cdot NCP_{A_i}(G), \tag{8.3}$$

where d is the number of quasi-identifiers in the dataset, A_i is a (numerical or categorical) attribute with weight w_i, where $\sum_i w_i = 1$ [Ghinita et al., 2009]. For our experiments, we assume all attributes have an equal weight but this can easily be altered in future experiments. NCP measures the information loss for a single instance and its equivalence class. This can be aggregated to the information loss in the entire dataset [Ghinita et al., 2009, Xu et al., 2006b] but for our problem setting, we only need to calculate the NCP for each k-anonymous counterfactual explanation, which constitutes one equivalence class. As an illustration, we calculate the NCP of the k-anonymous counterfactual explanation (CE) in our toy example[7]:

$$NCP_{Age}(CE) = \frac{max_{CE^{Age}} - min_{CE^{Age}}}{max_{Age} - min_{Age}} = \frac{27-24}{70-23} = 0.064,$$
$$\text{NCP}_{Gender}(CE) = 0 \qquad (|A^{CE}| = 1), \text{NCP}_{City}(CE) = 0 \qquad (|A^{CE}| = 1),$$
$$\text{NCP}(CE) = \frac{1}{3} \cdot 0.064 + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 0 = 0.021$$

For our experiments, we focus on the metrics NCP and pureness, but for completeness we also report the results with two additional metrics. The discernibility metric assigns a penalty to each tuple, based on how many tuples are indistinguishable from it after anonymizing. The idea is that it is desired to maintain discernibility between tuples as much as is allowed by a given setting of k [Bayardo and Agrawal, 2005]. The discernibility metric for anonymization g, and a degree of privacy k is:

$$C_{DM}(g,k) = \sum_{\forall E \text{ } s.t. |E| \geq k} |E|^2 + \sum_{\forall E \text{ } s.t. |E| < k} |D||E|$$

In this expression, E refers to the equivalence class of the tuple, and D to the dataset. Each successfully anonymized tuple (equivalence class larger than k) gets as penalty the size of the equivalence class, and each suppressed tuple (equivalence class smaller than k) gets as penalty the size of the total dataset. In our set-up, all the counterfactual explanations will be

[7] See Table 8.2 for the range of each attribute in the training set.

successfully anonymized so each anonymized explanation will get as penalty the size of its equivalence class. The discernibility metric for the k-anonymous counterfactual explanation in our example is 3, as this is the number of people belonging to its equivalence class (see Table 8.2.) This metric has been criticized because it does not take into account how much the anonymized data instances approximate the original instances [El Emam et al., 2009]. NCP is a more suitable metric to measure the actual information loss incurred by anonymizing the counterfactual explanations [Ghinita et al., 2007, Pramanik et al., 2016].

The classification metric (CM) is a class-conscious metric that attempts to create equivalence classes that consist of tuples that are uniform with respect to the class label [Iyengar, 2002].

$$CM = \frac{\sum_i^N penalty(tuple_i)}{N}$$

N is the number of anonymized tuples, which can be rows in a dataset or in our case number of the anonymized counterfactual instances. Each tuple receives a penalty of 1 if its class is different from the majority class label of its equivalence class. In the case of our toy example, the *k*-anonymized counterfactual instance does not receive a penalty as its label is the same as the majority class label of its equivalence class (*Accept*). This metric is related to our notion of pureness, but keep in mind that they measure different things. The classification metric looks at the instances in the equivalence class (which are Fiona, Gina and Ingrid in the case of our toy example) and their majority label. For pureness, we take all the attributes into account (so also the private attributes), and not only look at the instances present in the dataset, but at all the possible value combinations in the range of the anonymous explanation (by using sampling). This can be seen in Table 8.4. Pureness is therefore more suitable than the classification metric to measure how often the anonymous counterfactual explanation gives us *correct* advice.

8.5 materials and methods

8.5.1 *Materials*

We choose the datasets described in Table 8.5, as they are all tabular datasets that contain various personal attributes through which individuals could be identified, and are often used in research about privacy-preserving data mining [Kisilevich et al., 2008, Simi et al., 2017, Slijepčević et al., 2021].[14] All these datasets contain private information such as financial and health data that people generally do not want to be made public. In this Table, we list general dataset description properties such as the number of instances and attributes, and the target attribute. We also mention the quasi-identifiers and sensitive attribute (on which discrimination is measured) that we used for our experiments. Additionally, we measure the

[6] https://github.com/EpistasisLab/pmlb/tree/master/datasets/adult
[7] https://archive.ics.uci.edu/ml/machine-learning-databases/cmc/
[8] https://github.com/EpistasisLab/pmlb/tree/master/datasets/german
[9] https://github.com/EpistasisLab/pmlb/tree/master/datasets/heart_c
[10] https://www.opendatanetwork.com/dataset/health.data.ny.gov/82xm-y6g8
[11] https://github.com/kaylode/k-anonymity/tree/main/data/informs
[14] https://github.com/kaylode/k-anonymity

Table 8.5: Description of used datasets with dataset properties

Dataset	**Adult** [8]	**CMC** [9]	**German** [10]	**Heart** [11]	**Hospital** [12]	**Informs** [13]
# instances	48,842	1,473	1,000	303	8,160	5,000
# attributes	11	8	19	12	20	13
QID	*Age, Sex, Race, Relationship, Marital status*	*WifeAge, ChildrenBorn*	*Age, Foreign Personal status, Residence time, Employment, Job, Property, Housing*	*Age, Sex*	*Age Group, Race, Gender, Ethnicity, Zip Code*	*Dobmm, Dobyy, Sex, Marry, Educyear*
Sensitive attribute	*Sex*	*WifeReligion*	*Personal status*	*Sex*	*Gender*	*Race*
Target attribute	*Income*	*Contraceptive method*	*Credit decision*	*Heart disease*	*Costs*	*Income*
Uniquely identifiable (in %)	3.17	4.41	83.7	4.62	6.32	76.18
$\|EQ\| < 10$ **(in %)**	15.39	53.78	100	79.54	37.08	100

privacy risk present in each dataset in two ways: 1) We measure the percentage of people that are uniquely identifiable by their combination of quasi-identifiers, and 2) We measure the percentage of instances that are not protected by *k*-anonymity (with *k* =10). This thus means that we measure the percentage of people that belong to an equivalence class with a size smaller than 10.

8.5.2 *Methods*

On every dataset, we apply the methodology as described in Figure 8.3. We first split the dataset in a training and test set, using a split of 60-40. We fit and tune a Random Forest model through cross-validation on the training set. The following grid is used for tuning:

$$\text{n\_estimators} = [10, 50, 100, 500, 1,000, 5,000]$$
$$\text{max\_leaf\_nodes} = [10, 100, 500, n] \text{ with } n = \infty$$

We use the standard version (no optimization setting) of NICE [Brughmans et al., 2023a] as counterfactual algorithm, as this will return actual instances from the training set, and fit this on the training set and the trained machine learning model. This trained machine learning model is used to make predictions on all the instances in the test set. For all the test instances[15] without the desired prediction outcome, we use NICE to generate a counterfactual explanation. We focus on the test instances without the desired prediction outcome as these are the instances that generally use counterfactual explanations to receive advice on how to change their prediction outcome. As mentioned, when using NICE without any optimization setting, the counterfactual instances are real instances from the training data so they should be anonymized. The final step is to use CF-K to make these explanations *k*-anonymous.

[15] We set a limit at 1,000 instances for the sake of time.

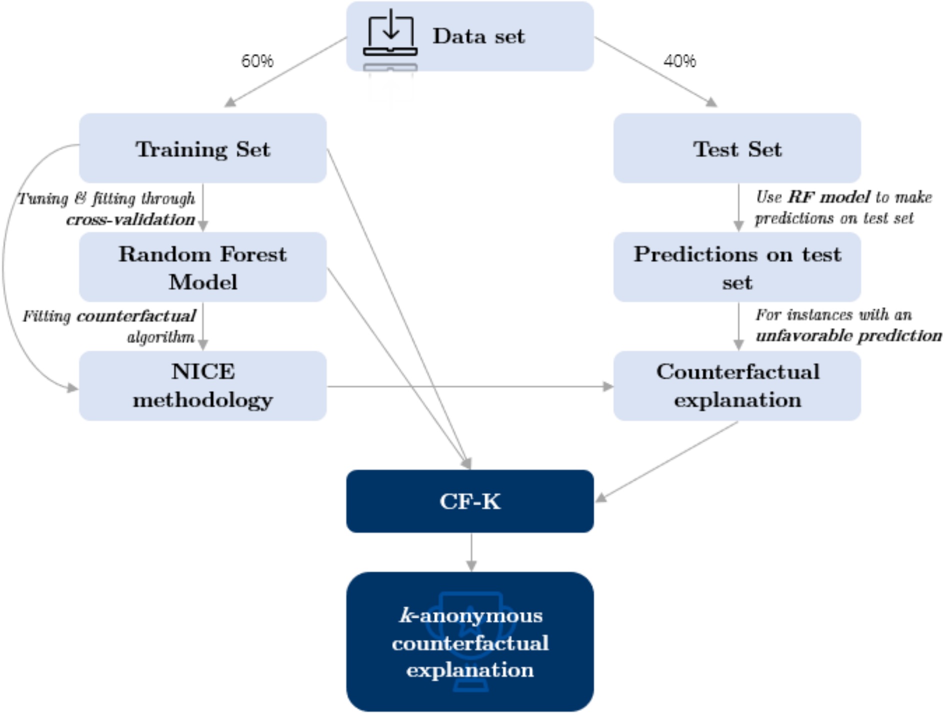

Data set
60%
40%
Training Set
Test Set
Tuning & fitting through cross-validation
Use RF model to make predictions on test set
Random Forest Model
Predictions on test set
Fitting counterfactual algorithm
For instances with an unfavorable prediction
NICE methodology
Counterfactual explanation
CF-K
k-anonymous counterfactual explanation

Figure 8.3: Used methodology to generate *k* anonymous counterfactual explanations from a dataset.

8.5.3 *Novel algorithm CF-K*

In the original application of k-anonymity, the whole dataset is made public and therefore has to be made k-anonymous. The goal is to find an optimal partition, for which both exact algorithms [LeFevre et al., 2005] and heuristics like genetic algorithms [Iyengar, 2002] and greedy algorithms (Mondrian [LeFevre et al., 2006], Datafly [Sweeney, 2002a]) exist.

Our approach differs from the approaches published in the literature, as only the counterfactual explanation is made public and not the whole dataset. Therefore, we search for an equivalence class for each returned counterfactual instance separately. This changes the set-up of the problem, as making the whole dataset k-anonymous can degrade the data more than just making the counterfactual explanations k-anonymous: not every training instance is used as counterfactual explanation and unused training instances do not need to be made k-anonymous or used in the calculation for the best clustering. In the same way that local encoding achieves less information loss than global recoding [Xu et al., 2006b], we hypothesize that only k-anonymizing the counterfactual instances can achieve lower information loss. We verify this claim in Section 8.6.2. Furthermore, specifically for our problem of k-anonymous counterfactual explanations we have to take the *counterfactual validity* of the k-anonymous explanations into account as this is essential for the goal of counterfactual explanations. We use this as additional metric in our algorithm.

We name our algorithm that makes counterfactual explanations k-anonymous CF-K. It is based on the metaheuristic GRASP (Greedy Randomized Adaptive Search Procedure) [Feo and Resende, 1995]. GRASP is a multi-start metaheuristic in which each iteration consists of two phases: construction and local search. After the two phases, the current best solution is updated. The construction phase builds a feasible solution, and the local search phase searches the neighborhood until a local optimum is found [Feo and Resende, 1995]. In this construction phase, GRASP combines greediness with randomness, with the purpose of escaping the myopic behavior of a purely greedy algorithm. We choose a heuristic algorithm, as it is a NP-hard problem and we are not looking for the optimal solution but for the best solution that can be found in limited computing time. Our aim is to provide a method that performs well, but we expect further optimizations to be possible in future research.

8.5.3.1 *Algorithm description*

Our algorithm starts from a counterfactual explanation that is given to one of the instances in the test set with an unfavorable prediction outcome. The counterfactual instance that this explanation is based on is an actual instance in the training set, and we want it to be unidentifiable from at least $k-1$ other instances in the training set. This is the case when at least $k-1$ other instances in the training set have the same values for the quasi-identifiers (these are the attributes that we assume to be publicly known).

PHASE 1: CONSTRUCT GREEDY RANDOMIZED SOLUTION In this phase, we construct a feasible solution. We first check for the current counterfactual instance if its values of quasi-identifiers are present for k individuals in the training set. In this case, a solution is found, the quality of the solution is calculated and the algorithm moves to the next phase. If this is not the case, we generate a list of size α by selecting the closest neighbors of the counterfactual instance in the training set with the required prediction outcome. Then, we *randomly* select a neighbour from this list and create a new *generalized instance* out of this neighbor and the counterfactual instance. The fact that we randomly select a neighbor out of this list and not just select the closest neighbour makes up the probabilistic component of GRASP. We create this *generalized instance* by generalizing the values of the quasi-identifiers so that the generalized instance includes both the values of the quasi-identifiers of the counterfactual instance as well as those of the neighbor. This happens as in Figure 8.2. We check again whether this generalised instance satisfies k-anonymity. If this is the case, a solution is found, the quality of this solution is calculated and the algorithm moves to the next phase. If this is not the case, this loop is repeated until the generalised instance satisfies k-anonymity.

PHASE 2: LOCAL SEARCH The local search algorithm iteratively tries to replace the current solution by a better solution in the neighbourhood. The algorithm terminates when no better solution is found. The neighborhood is defined by checking for every quasi-identifier in the current solution whether slightly changing it, is a feasible solution (satisfies k-anonymity) and improves the solution quality. A slight change in this case is adding a value (if the quasi-identifier is a single value) or removing a value from the list (if the quasi-identifier is already a generalized list).

Algorithm 8.1 GRASP

```
for $i = 1, ..., MaxIter$ do
    $Solution \leftarrow ConstructGreedyRandomizedSolution(Input)$;
    $Solution \leftarrow LocalSearch(Solution)$;
    $BestSolution \leftarrow UpdateSolution(Solution, BestSolution)$;
end for
return $BestSolution$;
```

GRASP We iterate these two phases for a specified number of iterations. After each iteration, we check if the new solution is better than the current best solution and if this is the case, we update the current best solution. After the specified number of iterations, the algorithm terminates and the current best solution is returned.

8.5.3.2 *Choice of parameters*

The input parameters in our algorithm are k, the level of desired privacy, α, the degree of randomness we give to the algorithm and the number of iterations the algorithm can perform. We show the effect of changing the input parameters on the *german* dataset by evaluating the metrics NCP, pureness and execution time.

DEGREE OF PRIVACY k We see that if we increase k, the level of privacy guarantees for each individual, the other metrics deteriorate. The Normalized Certainty Penalty, which

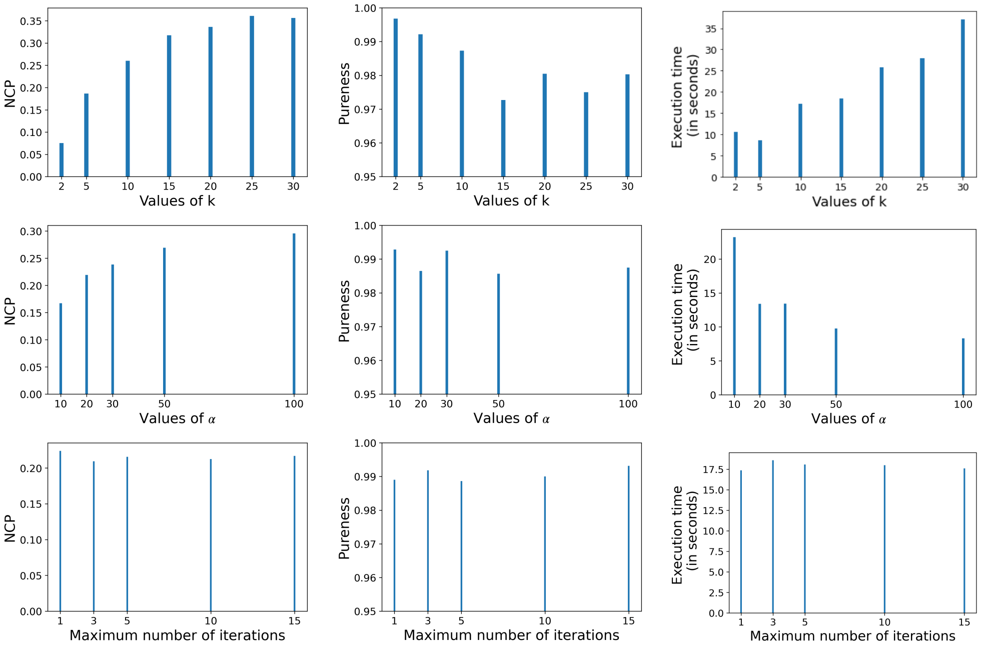


Figure 8.4: Evaluation of parameters

measures how much information value we lose by making the data k-anonymous, increases when we increase the value of k. This makes sense as the data quality degrades more when we add more privacy guarantees and therefore require more instances to be identical. Furthermore, the average pureness, and thus the counterfactual validity, also decreases. The trade-off between privacy (measured by k) and information loss (measured by NCP) has been confirmed by the literature [Ayala-Rivera et al., 2014, Sumana and Hareesha, 2010], but we are the first to show this trade-off between k and counterfactual validity (measured by pureness). Furthermore, the average execution time also increases if we increase k, as more privacy guarantees have to be implemented. For the remainder of the experiments, we use a k of 10 as this is a common number to baseline k-anonymity performance [El Emam and Dankar, 2008, El Emam et al., 2009]. We include the results for other values of k for both our algorithm (CF-K) and Mondrian in the Appendix.

PARAMETER α The parameter α is a measure of the randomness of the algorithm, as it determines the number of closest neighbors from which we randomly select one. We see that increasing α will increase the NCP but will lower the pureness. This is to be expected as we look at further neighbors when α is larger, so this will increase the information loss, but also creates more room to improve the pureness in the local search. Increasing α decreases the execution time, which is reasonable as we will satisfy k-anonymity faster by taking further neighbors The optimal value of α will depend on the dataset and how highly one values the different metrics. To avoid a multiple comparisons problem, we fix α at 20 for the rest of the experiments.

NUMBER OF ITERATIONS Increasing the number of iterations improves both the NCP and the pureness, but also increases the execution time. A trade-off has to be made between solution quality and execution time in determining the optimal number of iterations. We fix the number of iterations at 3 for the rest of the experiments.

8.6 RESULTS

8.6.1 *Results per dataset*

Table 8.6: Results of CF-K over all the datasets ($k = 10$).

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.55%	3.84%	21.41%	2.81%	3.42%	9.97%
Pureness (mean)	99.81%	93.15%	98.52%	100%	91.39%	85.33%
Execution time (mean)	24.78s	16.20s	13.31s	3.93s	17.76s	32.20s
C_{DM}	87,181	5,366	1,010	790	17,115	9,023
$\frac{C_{DM}}{\#explanations}$	110.78	13.2	16.83	14.11	22.94	13.65
CM	0.82	0.28	0.03	0.32	0.77	0.12

When we compare the results of Table 8.6 with the privacy risks of each dataset reported in Table 8.5, we see that explanations of the datasets with the highest privacy risks (*German* and *Informs*) have the highest information loss (in terms of NCP) when they are made anonymous. We measured the privacy risk by calculating the number of people in the dataset that are in equivalence classes smaller than 10 (before anonymizing), and for *German* and *Informs*, this will be the case for every person. For other datasets, such as *Adult*, only around 15% of individuals are in equivalence classes smaller than 10, so only a small portion of counterfactual instances will have to be anonymized. The average information loss (measured by NCP) for the anonymous explanations of this dataset is therefore much lower. The C_{DM} metric is harder to compare across datasets, as the size of the anonymous set has a large influence here. Therefore, we add an extra row where we divide C_{DM} by the number of anonymized explanations. This gives us the average size of the equivalence class for all the anonymized explanations. We see that for *Adult*, some equivalence classes can be really large, but the average NCP is low, which is more important for our problem. This consequently implies that the data did not have to be significantly degraded, but the generalized quasi-identifiers still encompass a substantial number of individuals. We also see that in the *Heart* dataset, the counterfactual validity measured by the pureness is always 100%. We expect this to be the case if the quasi-identifiers, which are *Age* and *Sex* in this case, have a small influence on the outcome of the machine learning model. We verify this by examining the feature importance ranking of the used model, and indeed see that the quasi-identifiers are ranked very low. This could explain why generalizing them has no effect on the counterfactual validity. For all datasets, the pureness is above 85%, which makes the generalized counterfactual explanations pretty valid. We see that although CM and pureness are related, they can give very different results per dataset. CM assesses the majority label of the whole equivalence class, while pureness will evaluate how many value combinations in the k-anonymous counterfactual instance will lead to the desired target outcome. As already said, for our use case, pureness is more relevant as this will actually assess how often the counterfactual explanations points us in the 'right' direction. Also note that for pureness, higher values are better, while for CM, lower values are preferred (less penalties).

We can also assess how the results vary for different values of k in the Appendix. We see that if k increases, in general the information loss becomes higher and the pureness becomes lower. This is in line with the results of Section 8.5.3.2, and again shows the trade-off between privacy and explainability.

With regard to the execution time, we see that it will be fast enough for most applications, and is in line with the order of magnitude of generating counterfactual explanations [de Oliveira and Martens, 2021]. If further speed-ups are necessary, this can be realised by decreasing the number of iterations, further optimization of the algorithm or using a stronger computer. All measurements were taken on a Dell Latitude 7400 laptop with 16GB of RAM and Intel®Core^{TM} i7-8665U CPU.

8.6.2 *Comparison with Mondrian*

Table 8.7: Results of the Mondrian algorithm (k = 10)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	15.97%	7.05%	59.55%	53.01%	26.03%	36.31%
Pureness (mean)	90.30%	69.15%	90.50%	100%	63.77%	72.40%
Execution time (mean)	7.11s	0.87s	0.38s	0.23s	1.19s	1.11s
C_{DM} **(mean)**	120,227	6,318	963	1,044	16,534	9,177
$\frac{C_{DM}}{\#explanations}$	152.77	15.56	16.05	18.64	22.16	13.88
CM (mean)	0.83	0.24	0.17	0.41	0.80	0.40

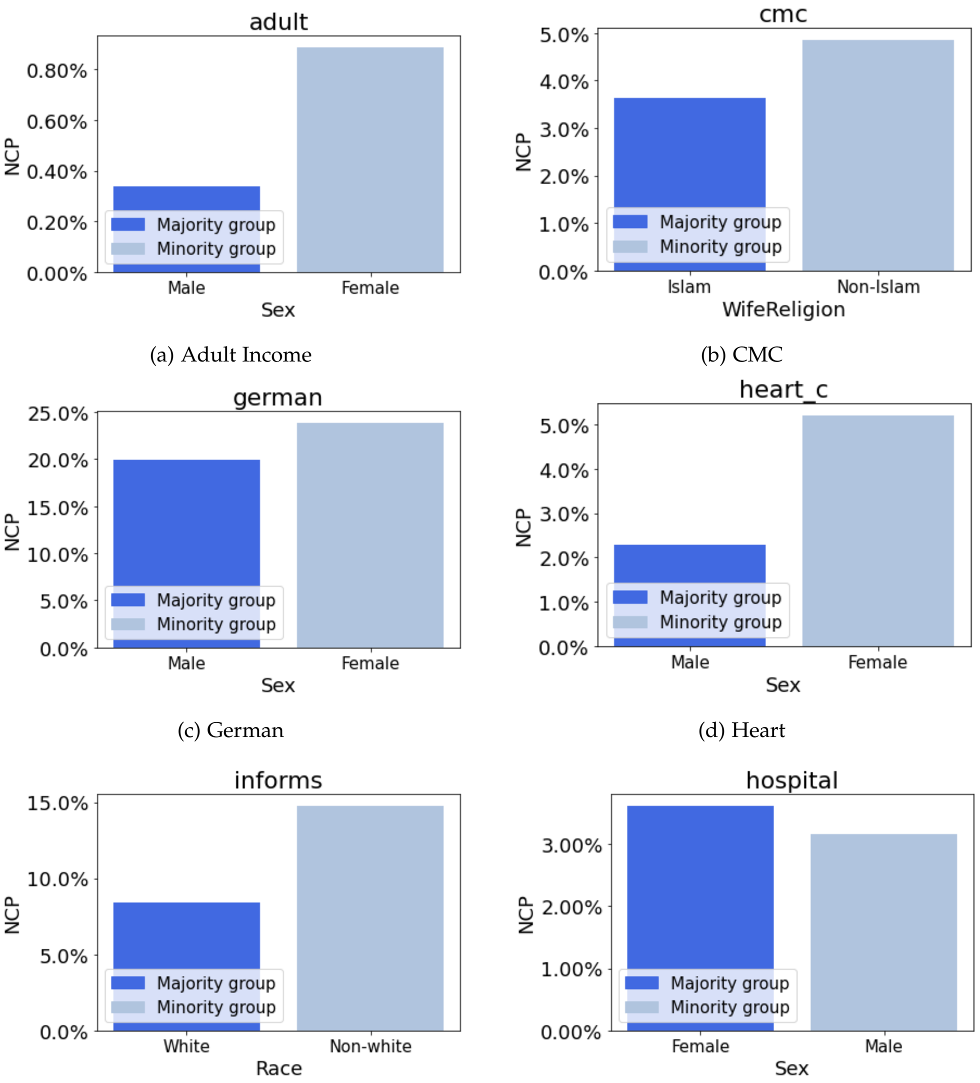
We compare CF-K with an alternative strategy: making the whole dataset k-anonymous, and taking the counterfactual explanations out of this anonymized dataset. This differs from our strategy where we directly make the counterfactual instances and explanations k-anonymous. We use an open source implementation of Mondrian[16] to compare CF-K with. Mondrian is a top-down greedy data anonymization algorithm that has been shown to be one of the best performers [Ayala-Rivera et al., 2014, LeFevre et al., 2006]. For all instances in the test set (max 1,000) with an unfavorable outcome, we compare the k-anonymous counterfactual explanation generated by CF-K with the k-anonymous counterfactual explanation based on an instance selected from the anonymized (by Mondrian) test set. When we compare the results in Table 8.6, with the results in Table 8.7, we see that for all datasets CF-K succeeds in achieving a better (and thus lower) average NCP than the Mondrian implementation on the whole dataset. This is in line with our hypothesis that only k-anonymizing the counterfactual instances can result in lower information loss, as unused training instances do not need to be used in the calculations for the best clustering. Furthermore, the average counterfactual validity (measured by pureness) in all datasets is higher when using k-anonymous explanations than when using an explanation from a k-anonymous dataset (except for the *Heart* dataset, where

[16] https://github.com/danielegiampaoli/Mondrian_K-anonymization

the average counterfactual validity is 100% for both implementations). Counterfactual validity can only be calculated on an explanation, and not on a dataset, so methods to make the dataset *k*-anonymous can not optimize for this metric. Therefore, our methodology to make the explanations *k*-anonymous, was needed to be able to take this metric into account. With regard to the C_{DM} metric: in four out of the six datasets, CF-K results in the smallest classes, while in two out of the six datasets, Mondrian will achieve slightly smaller equivalence classes. However, even in those cases, the average information loss measured by NCP will be lower when using CF-K, and as explained, it makes more sense to focus on this metric. The results for the CM metric show that for most datasets, the CF-K algorithm results in equivalence classes that are a bit more uniform with respect to the class label. However, as mentioned, pureness is more suited to measure the actual validity of the counterfactual explanations. We see in the Appendix, that the results for other values of *k* (5 and 20) are in line. For the Mondrian algorithm, the evaluation metrics also deteriorate when the level of privacy protection (*k*) is increased, and CF-K still outperforms Mondrian in terms of NCP and pureness for all values of *k*.

8.6.3 *Does this have fairness implications?*

A minority group is defined as a group whose characteristics such as race, religion, gender, ... etc. are fewer in numbers than the main group of that classification. Nowadays, it is often used to refer to people that experience a relative disadvantage based on their group membership [Healey et al., 2019]. We define the minority and majority group for each dataset based on the sensitive attribute, mentioned in Table 8.5. The minority group is the category of that sensitive attribute that is the least present in the training set. We see in Figure 8.4 that when we make the explanations more private (increase *k*), the explanation quality decreases and they become less useful. Unfortunately, this effect is larger for minority groups which can lead to potential issues regarding fairness. As can be seen in Figure 8.5, in every examined dataset (except for *Hospital*), the average NCP is higher for the minority group. For the average counterfactual validity, we found no difference between both groups. So we see in Figure 8.5 that the quality of explanations of the minority group has to be reduced more to achieve the same level of privacy. This can be explained by the fact that they often have more unique quasi-identifiers, as there are less people that share their public characteristics (definition of a minority group), so their quasi-identifiers have to be generalised more to be anonymous. For the *Hospital* dataset, the average information loss is slightly higher for the majority group. We hypothesize that this is due to the higher percentage of individuals with the desired target outcome (high income) for the minority group (men) than for the majority group (women), and hence it will be more difficult to find *pure* explanations for the latter. When explanations are used in high-stakes settings, it is undesirable that minority groups are offered lower quality explanations, but also that there is a higher risk of leaking their private information when no precautions are taken [Patel et al., 2022]. Other research showed another possible trade-off between fairness and privacy, as the privacy risks of different demographic groups are disparately affected by fairness-aware machine learning [Chang and Shokri, 2021]. These results show that different ethical objectives can work against each other and that one has to make sure that minority groups are not adversely affected in unexpected ways.



(a) Adult Income

(b) CMC

(c) German

(d) Heart

(e) Informs

(f) Hospital

Figure 8.5: Comparison of the average NCP between the majority and minority group

8.6.4 *Comparison with perturbation-based counterfactual algorithms*

We mentioned before that using *native* counterfactuals increases desirable properties such as plausibility, compared to counterfactual algorithms based on perturbations. CF-K is essentially slightly perturbing the native counterfactuals, so will the returned counterfactual explanations still be more plausible than the explanations from perturbed-based algorithms? Plausibility estimates the closeness of the counterfactual to the data manifold, by measuring the closeness

	NICE (none)	NICE (sparse)	NICE (prox)	NICE (plaus)	CF-K (k=5)	CF-K (k=10)	CF-K (k=20)
1NN	0	2.77	2.94	2.48	0.84	**1.22**	1.32
5NN	2.64	3.73	3.81	3.54	2.72	2.80	2.83

Table 8.8: Plausibility results for various settings of the NICE algorithm and CF-K, lower values are better (closer to the data manifold). NICE (*None*) returns *native* counterfactuals, the other settings of NICE (*sparse*, *prox* and *plaus*) return perturbed counterfactuals, and CF-K returns the anonymized version of the *native* counterfactuals (which will thus be slightly perturbed).

to the nearest instance(s) in the training data [Dandl et al., 2020, Brughmans et al., 2023a]. We report the average distance to the nearest and the 5-nearest neighbors for all settings of NICE (*none*, *proximity*, *sparsity*, *plausibility*). As explained before, only the *None* setting refers to a *native* counterfactual that will be grounded in the dataset, and the other settings will be perturbation-based counterfactual algorithms that aim to optimize for proximity, sparsity, and plausibility [Brughmans et al., 2023a]. We see in the benchmark study of Brughmans et al. [2023a] that the native counterfactual algorithms such as WIT and NICE (*None*) will result in the best plausibility scores, followed by NICE (*plausibility*), which is to be expected as it is designed to optimize for this metric. NICE (*plausibility*) outperformed all other perturbation-based algorithms, so this algorithm is chosen to compare with.

We also calculate the distance to the nearest and the 5-nearest neighbors for the anonymous counterfactual instances generated by CF-K (for different privacy settings).[17] The results for the *German* dataset can be seen in Table 8.8. NICE (*None*) still reports the best results, but CF-K significantly outperforms the other perturbation-based counterfactual algorithms, even NICE (*plausibility*). Furthermore, these other settings of NICE still start from an instance in the training set, so while they are less likely to return real instances, it is still a possibility. This is why for an optimal level of plausibility and a guarantee of privacy, it is better to use CF-K. Furthermore, we are also interested in the relationship between plausibility and privacy. When we increase the level of privacy protection, what is the effect on the plausibility of the *k*-anonymous explanations? We see in Table 8.8 that the plausibility metrics will deteriorate when we increase the level of privacy protection, which shows another side of the privacy-explainability trade-off.

8.7 discussion and future research

Transparency in machine learning has become a major topic, yet there is little research on the resulting potential risks to user privacy [Patel et al., 2022]. Although research has shown that offering model explanations may come at the cost of user privacy [Sokol and Flach, 2019, Shokri et al., 2019], none of the currently offered model explanation technologies offer any

[17] We measure the distance to a generalized counterfactual instance in a conservative way: We sample 100 times a value combination out of the generalized counterfactual instance (as we did to calculate pureness), and calculate its distance to its nearest and 5-nearest neighbors. For one generalized counterfactual instance, we then take the average distance over the 100 samples.

privacy guarantees. Once such explanation systems are deployed on high-stakes data, such as financial transactions or patient health records, a formal investigation of privacy risks is necessary. In this research, we introduce the *explanation linkage attack*, constituting the privacy risk that some counterfactual explanation techniques pose to the privacy of data subjects, because adversaries can infer their private attributes. We are the first to apply k-anonymity on counterfactual explanations instead of on the complete dataset, and show that applying k-anonymity only on the counterfactual explanations can achieve lower *information loss* and higher *counterfactual validity*. Furthermore, we see that if we increase the privacy constraints, the quality of the explanations becomes worse, which demonstrates the trade-off between *privacy* and *transparency*.

Other researchers [Patel et al., 2022, Shokri et al., 2019] have stated that assessing the privacy/explainability trade-off for minority groups is a promising avenue for future exploration, which is what we explored in Section 8.6.3. We noticed that the average information loss tends to be higher for minority groups, and this difference increases with the level of privacy, hereby introducing a new element of unfairness.

A debate on explanation quality could be a promising avenue for future research. For k-anonymous counterfactual explanations that have a pureness of 100%, generalized quasi-identifiers might actually be an advantage instead of a drawback. Think about the following scenario: Would you prefer the explanation *'If you would have been a teacher and would have earned $10K more, then you would have received the loan'* or the explanation *'If you would have been a teacher* or a nurse *and would have earned $10K more, then you would have received the loan'*, if both explanations are valid? While generalizing instances in a dataset means less information value, this trade-off is less clear in counterfactual explanations; generalizing them might give you more options to achieve the required target outcome and thus be *more* valuable. However, this is only the case when the counterfactual explanations are entirely *valid* and have a pureness of 100%. A discussion on explanation quality was not the goal of this study, so we leave this as an avenue for future research.

We also foresee another way to implement privacy constraints in future research, where the explanation technique itself is adapted to have privacy guarantees, instead of enforcing it in post-processing. Other authors propose a methodology where they search for a group of counterfactual explanations for a group of instances [Carrizosa et al., 2024b]. They do not include any privacy guarantees yet, but this kind of set-up could be used to create anonymized explanations as well. This could also have other desired side effects such as more robust explanations.

A last direction for future research we envision is applying other privacy schemes to counterfactual explanations. Beyond k-anonymity, other widely accepted protection schemes include l-diversity [Machanavajjhala et al., 2007], t-closeness [Li et al., 2006] and differential privacy [Dwork, 2006]. K-anonymity can be prone to privacy risks, for example when the attacker has background knowledge, can combine multiple explanations or when there is little diversity in the private attributes. l- diversity tries to solve these issues by requiring that the private attribute(s) should have a minimum of l properly depicted values. T-closeness goes even further and requires that the distance between the distribution of the private attribute in any equivalence class and the distribution in the whole table is less than a threshold t. Differential privacy offers a broader approach that captures the increased risk to one's privacy incurred by participating in a database, and counters this by introducing controlled noise into the data. Up until now, we assumed that all the attributes in the dataset except the quasi-identifiers, are private attributes, so l-diversity and t-closeness might not be that straightforward to implement. It is also important to note that we are explicitly searching for *no* diversity in the target variable, as we want k-anonymous counterfactual explanations that are as *pure* as possible. It will be interesting to see how applying these other privacy schemes (l-diversity, t-closeness and differential privacy) affect the explanations, and whether they will have the same implications regarding the explanation quality and fairness.

9

The Impact of Cloaking Digital Footprints on User Privacy and Personalization

Our online lives generate a wealth of behavioral records—*digital footprints*—which are stored and leveraged by technology platforms. This data can be used to create value for users by personalizing services. At the same time, however, it also poses a threat to people's privacy by offering a highly intimate window into their private traits (e.g., their personality, political ideology, sexual orientation). We explore the concept of *cloaking*: allowing users to hide parts of their digital footprints from predictive algorithms, to prevent unwanted inferences. This paper addresses two open questions: (i) can cloaking be effective in the longer term, as users continue to generate new digital footprints? And (ii) what is the potential impact of cloaking on the accuracy of *desirable* inferences? We introduce a novel strategy focused on cloaking "metafeatures" and compare its efficacy against just cloaking the raw footprints. The main findings are (i) while cloaking effectiveness does indeed diminish over time, using metafeatures slows the degradation; (ii) there is a trade-off between privacy and personalization: cloaking undesired inferences also can inhibit desirable inferences. Furthermore, the metafeature strategy—which yields more stable cloaking—also incurs a larger reduction in desirable inferences.

9.1 INTRODUCTION

A growing portion of our life happens online. We shop on Amazon, entertain ourselves on Netflix or Spotify, and communicate with friends and family via Facebook and Instagram. Whether we like it or not, the digital traces we generate during these interactions provide the mediating platforms with an extensive and comprehensive picture of our personal habits and preferences [Matz et al., 2020, Kosinski et al., 2013]. In fact, research has shown that a person's digital footprints—including their Facebook Likes and status updates, smartphone records or credit card spending—can be used to infer highly intimate characteristics such as sexual or political orientation, personality traits, mental health, or religious views [Kosinski et al., 2013, Matz et al., 2017]. Given that most individuals consider these characteristics deeply private, automated inferences of such traits without individuals' knowledge or consent raises important concerns related to people's rights to privacy and self-determination [Matz et al., 2020]. The act of drawing highly intimate inferences from seemingly innocuous data, for example, can be regarded as an intrusion of privacy, especially when individuals are neither aware of such inferences being made nor able to object to them. Moreover, the psychological insights that platforms (and other third parties) can glean from digital footprints allow them to influence their users' behaviors and decisions through mechanisms of personalization (an approach known as psychological targeting [Matz et al., 2017]).

As a potential remedy, Chen et al. [2017] introduced user *cloaking* of digital footprints. Cloaking first reveals to users which footprints they have to hide from predictive algorithms to avoid certain undesired inferences, and then gives them the option of restricting future inferences about them from using those particular footprints. Cloaking also has a key advantage over simply opting out of each particular offending inference: the exact same inference—that the user is a good target for particular content or a particular ad—may be unlikely to repeat; nonetheless, it may likely that very similar inferences will be made in the future. Cloaking has a substantial advantage over simply opting out of inferences altogether, as it will continue to allow desired inferences, for example for personalized content. However, it is not clear from prior work how well cloaking will perform over time, as individuals leave additional digital footprints.

Digital footprints are typically high-dimensional, sparse, fine-grained behavioral data; models normally draw on combinations of many different features as evidence to support a possible inference Ramon et al. [2021b]. Therefore, as our results show, using a cloaking strategy solely based on the fine-grained features will not be sufficient in the longer-term. People will continue to live their lives and behave similarly in the future, and predictive models will in many cases trigger once again using the new footprints [Chen et al., 2017, De Cnudde et al., 2020]. For example, our analysis of inferences based on Facebook Likes[1] reveals that when we take a snapshot at a later point in time, more than 80% of the people whom the models would infer to be Republican will be subject to this same inference again, despite cloaking the fine-grained features that drove the inference in the first place.

This is why we also investigate enhancing the longer-term efficacy of cloaking. Specifically, we examine grouping the fine-grained features into "metafeatures" (higher-level feature representations) and cloaking these metafeatures instead. The idea is that the metafeatures

[1] Following prior authors, we will capitalize "Like" when it refers to the action on Facebook.

will also include other similar behaviors that a user might take in the future, and which therefore may subsequently trigger the undesired inference. Our results show that this approach indeed increases the longer-term effectiveness of cloaking considerably and thus enhances privacy protection over time.

Importantly, the implications of cloaking digital footprints to reduce undesired inferences are not uniformly positive. As mentioned above, the same digital footprints may be used for different inferences as well. For example, a particular footprint might reveal not only sexual orientation but also the personality trait of Openness. While a user might be concerned about their data being used to infer their sexual orientation and subsequently discriminate against them, they might be appreciative of personalized services and ads that account for their level of openness to experience. That is, the same traces and mechanisms that may lead to discrimination, can also benefit users in the form of personalized content (e.g., individualized playlists, more relevant news, etc.). Desired personalization can not only lead to happier users but also to higher engagement [Fernández-Loría et al., 2017] and ultimately to higher platform revenue [Johnson et al., 2020].

In this paper we also examine the unintended consequences of cloaking. Specifically, suppressing certain digital footprints via cloaking mechanisms can have spillover effects: the data available for *desired* personalization tasks decreases, potentially reducing accuracy and effectiveness. To explore this privacy-personalization trade-off, we evaluate the impact of cloaking strategies on the accuracy of unrelated prediction tasks that are not the subject of cloaking. For example, we examine how cloaking for sexual orientation impacts the performance of a model predicting personality using the same large collection of digital footprints. Insights into the nature of this trade-off are crucial to empower users to make informed decisions about their online activity and about where on the privacy-personalization trade-off they want to be. To the best of our knowledge, our study is the first to evaluate empirically how a privacy intervention affects personalization levels, thereby offering new insights into the trade-off between privacy and personalization.

In summary, our study offers three main contributions:

- We assess the longer-term effectiveness of cloaking digital footprints, measuring the percentage of targeted individuals whose privacy remains protected over time. The results show that indeed the effectiveness of cloaking fine-grained features decreases steadily and markedly over time for most inference tasks.

- We introduce a new cloaking strategy based on metafeatures, and show that it enhances longer-term cloaking protection (as intended).

- We examine the privacy-personalization trade-off inherent in using cloaking to protect against unwanted inferences. Specifically, we show that cloaking for one task can affect the predictive performance of other personalization tasks. Moreover, the more-stable metafeature-based strategies have a stronger effect on other prediction tasks, highlighting the trade-off faced by users: better longer-term privacy protection indeed can reduce desired personalization performance more.

9.2 BACKGROUND

The trade-off between privacy and personalization is well recognized as a critical issue in our digital age [Garcia-Rivadulla, 2016, Chellappa and Sin, 2005, Habegger et al., 2014, Taylor et al., 2009]. On the one hand, personalization approaches might be appreciated by consumers who receive more relevant products and services as a result of targeted advertising and product design [Tran, 2017]. On the other hand, the ability to predict people's intimate traits and influence their behavior raises serious concerns for individuals and society at large. In countries where homosexuality is illegal, for example, the ability to infer sexual orientation from Facebook Likes could become a death sentence [Cabañas et al., 2018]. Similarly, health insurance companies could attempt to identify people with unhealthy habits or specific health problems, resulting in higher premiums or even rejection of coverage altogether [Cabañas et al., 2018]. The perhaps most well-known case of such an abuse is that of Cambridge Analytica, the UK-based PR firm which used psychological targeted advertisements on Facebook to interfere in the 2016 US presidential elections [Matz et al., 2020, Doward and Gibbs, 2017].

Given the seriousness of these potential transgressions and privacy violations, scientists, activists and policy makers have pushed for legislation that aims to prohibit the prediction of protected categories, such as race or religion. Facebook, for example, has faced years of criticism for offering advertisers 'interest' categories that have led to the exclusion of people of color from housing ads, fueled political polarization, and helped Big Pharma track users with specific illnesses [Waller, Angie and Lecher, Colin, 2022, Angwin and Parris Jr., 2016, Edelman, Gilad, 2019, Lecher, Colin, 2021]. In 2022, Facebook responded to the growing public pressure and changing regulatory landscape by removing the option to target users explicitly based on potentially sensitive traits such as health, race, sexual or political orientation, and religious beliefs [Waller, Angie and Lecher, Colin, 2022].[2]

However, non-profit news organization The Markup reported that Facebook's attempts at better protecting their users' privacy and preventing discrimination were only partially successful. For example, although *Hispanic Culture* was removed from the target categories available to advertisers, *Spanish Language* was not [Waller, Angie and Lecher, Colin, 2022]. Despite the fact that Facebook has since removed additional interests and categories related to protected traits, we argue that playing whack-a-mole across many millions of pages and categories is destined to fail. This is partially the case because few users for whom a protected trait is predicted will actually Like pages that explicitly reveal these traits. For example, less than 5% of users predicted to be homosexual were connected with explicitly homosexual pages such as *No H8 Campaign*, *Being Gay* or *I Love Being Gay* [Kosinski et al., 2013].

Consequently, the mere act of eliminating certain prediction categories from the platforms prediction or targeting engines is insufficient. This is particularly true when ads or content are targeted based on machine-learned models rather than the explicit choice of individual interests. Even if *Homosexuality* is removed as an explicit targeting category and *No H8 Campaign* is removed as a data item for prediction, algorithms likely will learn to use other digital footprints to target content or ads that would appeal to gay individuals.

[2] Some of the interest categories that will be no longer available include 'Gay Pride', 'Islamic Calendar', and 'Lung Cancer Awareness' [Silberling, Amanda, 2021, Waller, Angie and Lecher, Colin, 2022]. A comprehensive list of removed Facebook pages can be found here: https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

In order to substantially limit the predictability and use of sensitive traits across all users, platforms would have to ban an unreasonable number of pages from their inference algorithms, among them many seemingly neutral pages that will be hard to justify and would likely evoke concerns related to freedom of speech and expression. Moreover, implementing such paternalistic measures may undermine an individual's agency over what they choose to reveal about themselves. For example, should we force users to restrict their online identities if they feel perfectly safe and comfortable about their lifestyle and sexual orientation, and would be delighted to receive associated advertisements? In addition, generic one-size-fits-all approaches to restricting certain aspects of online behavior can have negative consequences for socially relevant causes that would benefit from personalization and civic engagement: climate activists and medical researchers, for example, have pointed out that the changes to Facebook's targeting platform have severely limited their ability to reach relevant audiences [Waller, Angie and Lecher, Colin, 2022].

In this paper, we examine an individualized approach that offers users more control and transparency over their online identities, and can be tailored to and by the individual: cloaking certain digital traces that are relevant for inferences about a particular individual. Chen et al. [2017] propose a "cloaking device" that reveals to users the digital footprints without which the prediction model would not have made the inference, and allows them to restrict inference procedures from using those footprints. Let us consider a digital footprint to be a specific aspect of online behavior that is stored about the individual on a technology platform, such as a particular song listened to on Spotify or a specific page Liked on Facebook. *Cloaking* a digital footprint means removing it from the set of data considered by an algorithm drawing inferences for that user. In the common case of a machine learning model in an AI inference system, where the digital footprints are the features used by the model, cloaking the digital footprint represented by feature x for user u would mean setting the value of x to whatever would be the value if the system had not saved that digital footprint for that user (for example, setting the feature value to zero as an indication that the user did not Like the page in question). This cloaking could be implemented by the platform, by providing users with the option to choose which inferences to avoid. Alternatively, such transparency could guide users to better decide which data they feel comfortable sharing in the first place; however, for many systems, the digital footprints are the result of simply using the system, so this latter alternative would involve restricting one's own behavior.

The reason why we (and previous authors) focus on cloaking the underlying features, and not the particular inferences, is because the latter cannot protect the user from closely related inferences in the future [Chen et al., 2017]. This is in line with the current advertising options on Facebook: as mentioned before, it is no longer possible to target people based on certain private traits (e.g., sexual orientation). Hence, targeting such advertisements has to rely on associated interests (e.g. Facebook Likes that are empirically related to a certain sexual orientation). This is natural since fine-grained behavioral data, such as Facebook Likes, are the primary features used by content-selection and ad targeting machine learning models on such platforms [Facebook, 2020, Andreou et al., 2019, Chaudhary et al., 2021, Lukka and James, 2014]. The previously proposed cloaking strategy has been shown to be effective in avoiding inferences at the time the cloaking takes place, with relatively little burden on the users [Chen et al., 2017]. However, prior work has not investigated how effective the cloaking would be over time, as users continue to leave new digital traces.

9.3 MATERIALS

We use data from the MyPersonality project, which contains the Liked Facebook pages of 220,489 volunteers in the United States, along with their scores on the Big 5 personality traits and personal characteristics such as gender, age, sexual orientation and political preferences [Kosinski et al., 2013]. A Facebook Like is a mechanism used by Facebook users to express their positive association with online content, and in this case we focus on the public pages they Liked, which can relate to products, public persons, music, sports, books, restaurants, or public statements they agree with. Using this data, it is possible to create a user-Like matrix X such that $x_{ij} = 1$ if user i Liked page j. Behavioral datasets, such as Facebook Likes, are usually very sparse as any user usually only takes a limited number of actions (in this case Like Facebook pages), while the total number of possible actions is very large [Junqué de Fortuny et al., 2013]. As described in more detail below, we assess the impact of cloaking the Likes that lead to the inferences of gender, political orientation and sexual orientation.[3] In this study, we use these as examples of the attributes individuals might wish to safeguard; the specific attributes deemed private of course will vary depending on the individual's preferences. The data is described in Table 9.1.

Table 9.1: Data description for the target variables that will be cloaked. We select only the instances that have a value for the corresponding trait.[4] The features are the Facebook pages that remain after pre-processing. *Active elements* shows the number of non-zero elements in the entire matrix; *Sparsity* is the percentage of zero elements over the total number of elements in the matrix. *Average Likes* is the average number of Likes a person associated with this trait has. *Balance* is the percentage of instances with a positive value for the target variable.

Target variable	Instances	Features	Active elements	Sparsity (in %)	Average Likes	Balance (in %)
Male	165,234	115,326	16,901,459	99.91	86.8	38.37
Female	165,234	115,326	16,901,459	99.91	112.0	61.63
Homosexual	22,477	115,326	2,197205	99.92	104.3	4.67
Lesbian	29,309	115,326	4,041,148	99.88	110.5	2.65
Democrat	36,534	115,326	4,190,576	99.90	134.0	17.27
Republican	36,534	115,326	4,190,576	99.90	124.2	10.24

PERSONALITY TRAITS Personality trait research suggests that personality consists of a range of consistent and relatively stable characteristics (traits) that determine how an individual will think, feel and behave [Matz et al., 2016]. The Big 5 (BF) Model of Personality is the most widely accepted model and proposes five independent traits to capture individual personality differences [Costa and McCrae, 1992, Matz et al., 2016]. The five traits are: 1)

[3] Only gender is still available as an explicit targeting option on Facebook, but machine learning models can still learn the other traits implicitly when optimizing a particular ad or content element.

[4] For the prediction task of homosexuality, only men whose data record has a value for sexual orientation will be considered, while for the prediction task of lesbian, only women with a value for sexual orientation will be taken into account.

Extraversion, the tendency to seek excitement and stimulation in the company of others; 2) *Openness*, the tendency to be intellectually curious, creative and unconventional; 3) *Neuroticism*, the tendency to experience negative emotions, and being anxious and nervous; 4) *Agreeableness*, the tendency to be trusting, compassionate and cooperative; and 5) *Conscientiousness*, the tendency to be organized and efficient [Matz et al., 2016, Ramon et al., 2021a]. The Big 5 personality traits were established using the International Personality Item Pool (IPIP) questionnaire with 20 items [Goldberg et al., 2006, Kosinski et al., 2013]. The traits are recorded on a 5-point Likert scale, and we only select the data instances that have a value for the corresponding trait. Research shows that the power of digital footprints to predict personality traits is in line with the typical strength of the relationship between personality and behavior, also known as the *personality coefficient* (a correlation between 0.30 and 0.40) [Meyer et al., 2001, Azucar et al., 2018].

9.4 methods

cloaking mechanism. As described above, *cloaking* refers to the mechanism of changing user data so that—from the perspective of the inference procedure—it was as if the user did not exhibit one or more specific behaviors. In this setting of Facebook Likes, cloaking can be defined as hiding specific Like pages from the prediction algorithm. This does not mean that the user actually has to *unLike* these pages (although this can be an alternative as well), but that the prediction algorithm no longer uses these "hidden" data attributes for the inference of that person. The cloaking mechanism introduced by Chen et al. [2017] relies on counterfactual explanations. These counterfactual explanations explain inferences made by machine learned models via the features that led to the inferences, defined specifically as a minimal subset of features the removal of which will change—in our case, inhibit—the system's inference [Martens and Provost, 2014, Wachter et al., 2017b, Verma et al., 2020, Fernández-Loría et al., 2022]. When using behavioral data, this corresponds to a minimal set of non-zero features of the instance, where changing (just) these feature values to zero would lead the model to draw a different inference [Ramon et al., 2020]. We apply counterfactual explanations instead of other explanation techniques as they give a direct way to alter the predicted outcome, in line with Chen et al. [2017]. Nevertheless, with suitable modifications other explanation techniques such as SHAP could also be adapted to support cloaking [Ramon et al., 2020, Lundberg and Lee, 2017].[5]

We use the procedure introduced in Martens and Provost [2014] to compute the counterfactual explanations.[6] This algorithm finds counterfactual explanations using a heuristic search that requires the decision to be based on a scoring function, such as a probability estimate from a predictive model [Fernández-Loría et al., 2022]. The search algorithm then uses this scoring function to first consider features that, when changed to their counterfactual values, reduce the score of the predicted class the most. When a set of features is found that would alter the outcome of the predicted class, these features are changed to their counterfactual value. This change is done by replacing the original feature value with the median value of that feature over the training data, which in the case of behavioral data will be 0 as this data is

[5] Fernández-Loría et al. [2022] detail why feature importance methods such as SHAP cannot be used directly for tasks like this that depend on inhibiting the inference.

[6] Python code available at https://github.com/ADMAntwerp/edc

extremely sparse. For example, in the Facebook data, there is no page that is Liked by the majority of users, so the median value of every feature will be 0. Counterfactual explanations will then point to the Facebook Likes a user has to cloak (or simply unLike). An example of such a counterfactual explanation could be: *If you would not have liked the pages 'The Tea Party Patriots' and 'Sarah Palin', you would not not have been predicted to be a Republican.* A user is considered to be successfully cloaked when his or her score falls below the predefined threshold for drawing the inference in question [Chen et al., 2017]. For example, this might be the model score threshold used for deciding to target content or an ad. The average size of a counterfactual explanation, i.e., the average number of Likes that have to be cloaked to avoid positive inference for each prediction task, can be found in Table 9.2.

Table 9.2: Model statistics. *Positive rate* indicates the percentage of (test set) instances that are predicted as positive by the machine learned model. *AUC* is the model's accuracy on the task, as measured by the area under the ROC curve. *Explanation size* is the average number of Likes that must be cloaked to avoid positive inferences.

Target variable	AUC (in %)	Positive rate (in %)	Explanation size (avg.)
Male	95.2	5.13	8
Female	95.2	4.75	6
Homosexual	89.4	5.72	4
Lesbian	77.8	7.88	2
Democrat	77.3	4.58	3
Republican	82.1	4.79	6

METAFEATURES Dimensionality reduction methods are techniques to reduce a high dimensional feature space into a lower-dimensional form. To group fine-grained features into higher-level metafeatures, we use Non-Negative Matrix Factorization (NMF) [Lee and Seung, 1999].[7] We chose this dimensionality reduction technique because the non-negativity constraint facilitates the interpretation of the extracted metafeatures, and it has been shown to provide interpretable results for fine-grained data applications [Contreras-Piña and Ríos, 2016, Ramon et al., 2021b].[8] We create 50 metafeatures for the Facebook Likes and assign each page to the metafeature or topic for which it has the highest weight to ensure mutual exclusivity (each feature only belongs to one metafeature) [Ramon et al., 2021b]. An example of two metafeatures is shown in Section 9.4.

Another option to create metafeatures is to use the categories that Facebook assigned to the Facebook Like pages itself. These categories are more broad such as 'Public Figure' or 'Musician/Band'. We will call these *domain-based metafeatures*, in line with Ramon et al. [2021b], to contrast with the data-driven metafeatures produced using NMF. The different datatypes used in the study are shown in Table 9.3. The results in Section 9.6 are generated

[7] Non-Negative Matrix Factorization (NMF) is a dimensionality reduction technique that decomposes a non-negative data matrix into two lower-dimensional, non-negative matrices. It is particularly useful for identifying latent features in data, when we would like to be able to interpret the latent features.

[8] Note that there exist many other techniques to generate the metafeatures, but we do not compare them in this work, as our goal is to study whether using metafeatures can give better performance, rather than to figure out what sort of metafeatures performs best.

using the data-driven MF. We report the results from using domain-based metafeatures in Appendix E.1.

Datatype	In this study
Fine-grained feature (FG)	Facebook Like
Metafeature (MF)	Data-driven MF (created by NMF) Domain-driven MF (assigned by Facebook)

Table 9.3: Features that are used in the study

9.5 experimental set-up

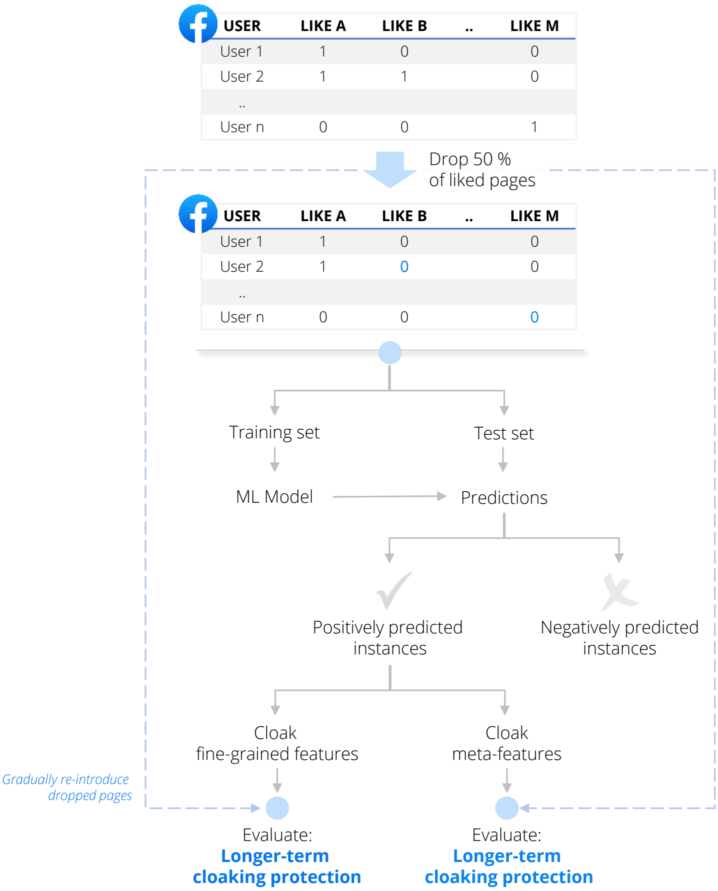
We focus on cloaking the inferences gender (*male* and *female*), sexual orientation (*homosexual* and *lesbian*) and political orientation (*democrat* and *republican*). We train Logistic Regression models with ℓ_2-regularization with the Scikit-learn library (Python). We chose Logistic Regression as the literature has shown that this is one of the best performing classification models for behavioral data [De Cnudde et al., 2020], and this type of model is commonly used to train models on behavioral data [Agarwal et al., 2014, Perlich et al., 2014, Clark and Provost, 2019, Ramon et al., 2021a]. We use 66% of the data for training, and the remaining 33% for testing. We also exclude users with fewer than 10 Likes and Facebook pages with fewer than 10 Likes. For fine-tuning the hyperparameters of the model, we perform a grid search on the training set by using three-fold cross-validation, where we tune the regularization parameter C of the ℓ_2-LR model. As is common in targeted advertising, we assume that a positive inference is drawn, which means that the user would be targeted, when the model assigns the user a score which places him or her in a specified top quantile of the score distribution produced by the prediction model [Chen et al., 2017, Perlich et al., 2014]. For online targeting, a typical value for this quantile is between 90 and 100, and we base our threshold for positive inference on the 95th percentile of the scores over the training set [Chen et al., 2017, Perlich et al., 2014]. The chosen threshold will of course depend on the budget of the advertising campaign, and can be adjusted based on campaign performance data and insights from initial targeting efforts.[9] The test set AUC and positive rate for each prediction task are reported in Table 9.2.

9.5.1 *Longer-term cloaking protection*

We study longer-term cloaking protection using the methodology depicted in Figure 9.1. We simulate a person's behavior over time by first holding back 50% of Likes for each user at random.[10] After dropping these pages, we train a regularized logistic regression model for every prediction task on the reduced training set. We use this model to make predictions on

[9] Note that a limitation of this approach is that it will be impossible to use the targeting thresholds that are used by mainstream social media platforms, given that this information is proprietary.

[10] This means that the simulation uses the assumption that people's behavior over time is stable in the short run, as their liking behavior does not change—since the data does not include time stamps. Verifying this on time-stamped data would be an avenue for follow-up research.



USER	LIKE A	LIKE B	..	LIKE M
User 1	1	0		0
User 2	1	1		0
..				
User n	0	0		1

USER	LIKE A	LIKE B	..	LIKE M
User 1	1	0		0
User 2	1	0		0
..				
User n	0	0		0

Drop 50 %
of liked pages
Training set
Test set
ML Model
Predictions
Positively predicted instances
Negatively predicted instances
Cloak fine-grained features
Cloak meta-features
Gradually re-introduce dropped pages
Evaluate:
Longer-term cloaking protection
Evaluate:
Longer-term cloaking protection

Figure 9.1: Experimental set-up to measure the longer-term cloaking protection.

the instances in the reduced test set and select the positively predicted instances. For these instances, we compare two cloaking strategies to inhibit positive inferences:

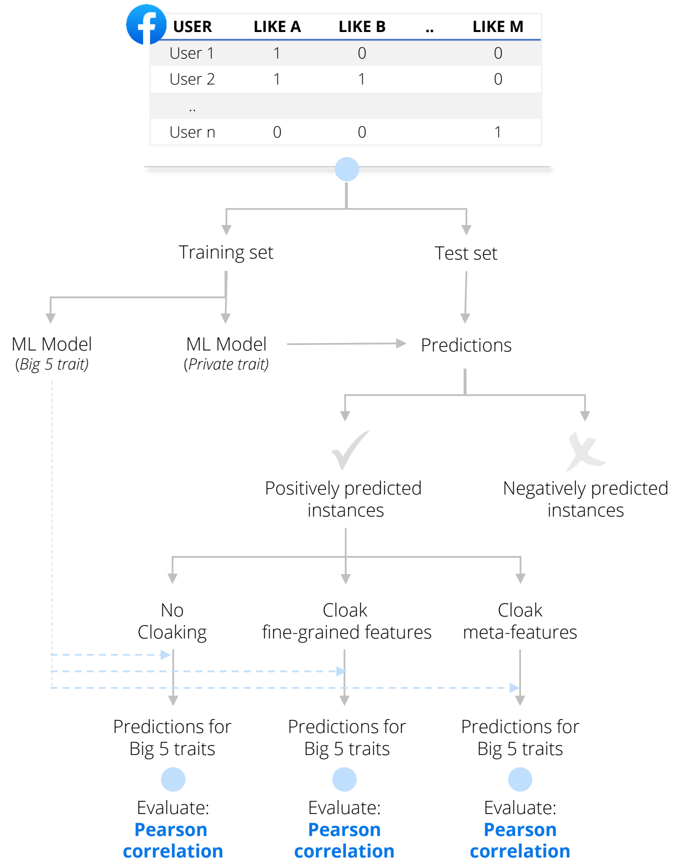
1. Cloaking the fine-grained features—i.e., the individual Likes. Specifically, we remove all the Liked pages in the corresponding counterfactual explanation of that instance, the same procedure used by Chen et al. [2017]. We call this strategy *FG*.

2. Cloaking the metafeatures, where we remove all the pages in the counterfactual explanation of that instance **and** the other Liked pages that belong to the same metafeatures as the pages in the counterfactual explanation. We call this strategy MF.

Using the second strategy leads to the number of Liked pages available for inference decreasing substantially more than when using the first strategy, which we will show in Figure 9.6.

We simulate an individual's behavior over time by gradually introducing the 50% of pages that initially were held back. We measure the **longer-term cloaking protection** of both strategies by computing the percentage of positively predicted instances for which cloaking this targeting task successfully inhibited future inferences for that same task and individual.

9.5.2 *Trade-off between privacy and personalization*



USER	LIKE A	LIKE B	..	LIKE M
User 1	1	0		0
User 2	1	1		0
..				
User n	0	0		1

Training set
Test set
ML Model (Big 5 trait)
ML Model (Private trait)
Predictions
Positively predicted instances
Negatively predicted instances
No Cloaking
Cloak fine-grained features
Cloak meta-features
Predictions for Big 5 traits
Predictions for Big 5 traits
Predictions for Big 5 traits
Evaluate: Pearson correlation
Evaluate: Pearson correlation
Evaluate: Pearson correlation

Figure 9.2: Experimental set-up to measure the impact of cloaking a private trait on other prediction tasks.

As discussed at the outset, although we may think myopically about a privacy-preserving action when taking it, such actions can have spillover effects. In particular, cloaking data in order to inhibit one inference can have effect on other inferences—possibly ones that we would not want to inhibit. Therefore, consumers and platforms should be interested in what effect hiding portions of someone's digital traces has on the performance of other inference tasks.

Define X as the initial complete data, and X_c as the cloaked data. To what extent does changing X to X_c affect the predictions of models predicting different target variables?

We show the set-up of our experiment in Figure 9.2. We examine the effect on a second set of prediction tasks when applying cloaking to the sensitive-trait-prediction tasks we described above. The new tasks involve predicting an individual's ratings for the Big 5 personality traits, the accuracy of which we measure using Pearson correlation, which is the most commonly used measure of prediction accuracy for predicting these personality traits [Kosinski et al., 2013, Azucar et al., 2018]. We choose the Big 5 traits as the set of tasks to assess spillover effects because they cover broad aspects of personality and are very well understood. We compare the effects of: not cloaking an individual's data, cloaking fine-grained footprints (FG), and cloaking metafeatures (MF).[11]

9.6 RESULTS

9.6.1 *Longer-term cloaking protection*

Let's consider first a single individual, whom we will call John, who has been using a particular technology platform and thereby leaving digital footprints. The platform's political-orientation model gives him a high enough score as a *Republican* in order for him to receive corresponding political ads. John no longer wants to receive advertisements related to his political orientation, maybe because he no longer identifies as such, or he wants to keep his political orientation private, or he simply finds these advertisements annoying. We want to see if, as he subsequently continues his usual behavior, and thereby continues to Like pages, John gets targeted as Republican again after using the cloaking device described by Chen et al. [2017].

We represent John in Figure 9.3. As described above, we simulate the point where the model uses half of his digital traces (88 Likes) as the current footprint data, and the point with all his digital footprints as his future data (175 Likes). Using the current data, John is targeted as a Republican.[12] John wants to inhibit this inference and receives the following advice (based on counterfactual explanations to bring him under the threshold): *If you would hide the Likes 'Conservative' and 'Chick-fil-A', you would no longer be targeted as Republican.* After cloaking these pages, John has 86 Likes remaining and is no longer targeted as Republican.[13]

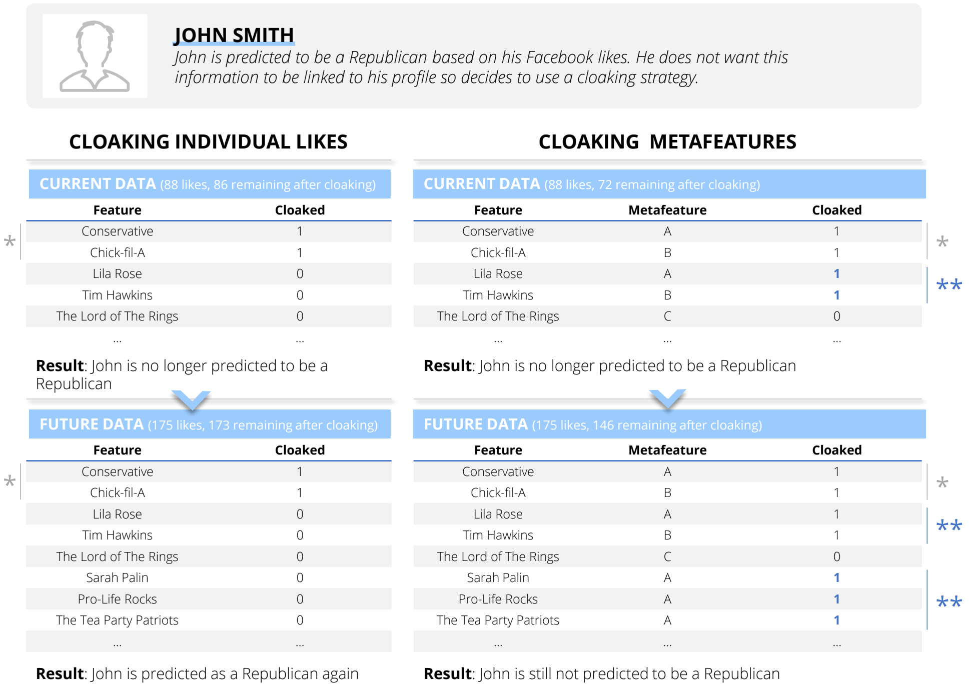
We move on to the future point. John, who has remained active, has Liked 87 new pages (essentially doubling his digital traces). Even though the two Liked pages from his initial counterfactual explanation are still cloaked, John gets re-predicted as Republican due to his new digital footprints.[14] This illustrates that cloaking is not necessarily robust in the longer

[11] The point here is not that inferences for the Big 5 traits are not privacy invasive; this of course will depend on the individual. Rather, the point simply is to examine the effects of cloaking some potentially sensitive inferences on other inferences that are broadly applicable.

[12] Prediction score = 0.161, which is above the targeting threshold of 0.148.

[13] Prediction score = 0.140, which is below the threshold of 0.148.

[14] His prediction score on the full data is 0.260; after cloaking the two Likes in his explanation, his prediction score is 0.225. This above the targeting threshold of 0.194. The threshold is different now because after



JOHN SMITH

John is predicted to be a Republican based on his Facebook likes. He does not want this information to be linked to his profile so decides to use a cloaking strategy.

CLOAKING INDIVIDUAL LIKES

CURRENT DATA (88 likes, 86 remaining after cloaking)

Feature	Cloaked
Conservative	1
Chick-fil-A	1
Lila Rose	0
Tim Hawkins	0
The Lord of The Rings	0
...	...

Result: John is no longer predicted to be a Republican

FUTURE DATA (175 likes, 173 remaining after cloaking)

Feature	Cloaked
Conservative	1
Chick-fil-A	1
Lila Rose	0
Tim Hawkins	0
The Lord of The Rings	0
Sarah Palin	0
Pro-Life Rocks	0
The Tea Party Patriots	0
...	...

Result: John is predicted as a Republican again

CLOAKING METAFEATURES

CURRENT DATA (88 likes, 72 remaining after cloaking)

Feature	Metafeature	Cloaked
Conservative	A	1
Chick-fil-A	B	1
Lila Rose	A	1
Tim Hawkins	B	1
The Lord of The Rings	C	0
...	...	...

Result: John is no longer predicted to be a Republican

FUTURE DATA (175 likes, 146 remaining after cloaking)

Feature	Metafeature	Cloaked
Conservative	A	1
Chick-fil-A	B	1
Lila Rose	A	1
Tim Hawkins	B	1
The Lord of The Rings	C	0
Sarah Palin	A	1
Pro-Life Rocks	A	1
The Tea Party Patriots	A	1
...	...	...

Result: John is still not predicted to be a Republican

Figure 9.3: Example of John. The column *Cloaked* signals the pages that are cloaked for each strategy and point in time.
*: Original features cloaked to ensure John is not predicted as republican.
**: Additional features cloaked because they are part of same metafeature as the original features.

term, as individuals continue to leave new digital footprints. (Note that Chen et al. pointed out as a limitation of the original cloaking design that if cloaking does not also cover closely associated features, one might end up being targeted again in the future [Chen et al., 2017].)

Cloaking based on metafeatures is intended to (partially) address this lack of robustness. Recall that cloaking metafeatures also cloaks other footprints that are (estimated to be) closely related to those suggested by the counterfactual explanation. So for our current example, the Facebook page 'Conservative' belongs to metafeature A, and the Facebook page 'Chick-fil-A' belongs to metafeature B. Typically, metafeatures such as these[15] are interpreted by looking at the top weighted fine-grained features for each metafeature [Wang and Zhang, 2012, O'callaghan et al., 2015, Contreras-Piña and Ríos, 2016]. These are shown in Table 9.4.

Metafeature A clearly is related to right-wing politics, and metafeature B to Christianity. Metafeature cloaking hides not only the Likes (fine-grained features) in the counterfactual

everyone in the dataset has acquired new digital traces, the scores for the top 5th percentile will be different.

[15] Specifically, those created by embedding the original data in a lower dimensional space.

Table 9.4: Interpretation of two metafeatures generated with NMF by showing the 10 features with the highest coefficients for each metafeature.

Metafeature A	Metafeature B
Being Conservative	*The Bible*
Sarah Palin	*Jesus Daily*
Conservative	*"I'm proud to be Christian" by Aaron Chavez*
Glenn Beck	*Casting Crowns*
Fox News	*Chris Tomlin*
Tea Party Patriots	*Third Day*
Mitt Romney	*TobyMac*
FreedomWorks	*Jeremy Camp*
Sean Hannity	*Switchfoot*
John McCain	*Skillet Music*

explanation, but also all the Likes that belong to the same metafeature as each of these Likes. When we also cloak all the Likes in the associated metafeatures, 14 additional pages are cloaked. These include 'Tim Hawkins' and 'Lila Rose'.[16] Subsequently, when John Likes pages in the future, the new pages associated with those same metafeatures will be hidden as well. For John, this leads to also hiding pages such as 'Sarah Palin', 'Tea Party Patriots' and 'Pro-Life Rocks'. In total, 29 new pages are cloaked in the future and the result is that John will not be predicted as Republican even after leaving his future footprints.[17]

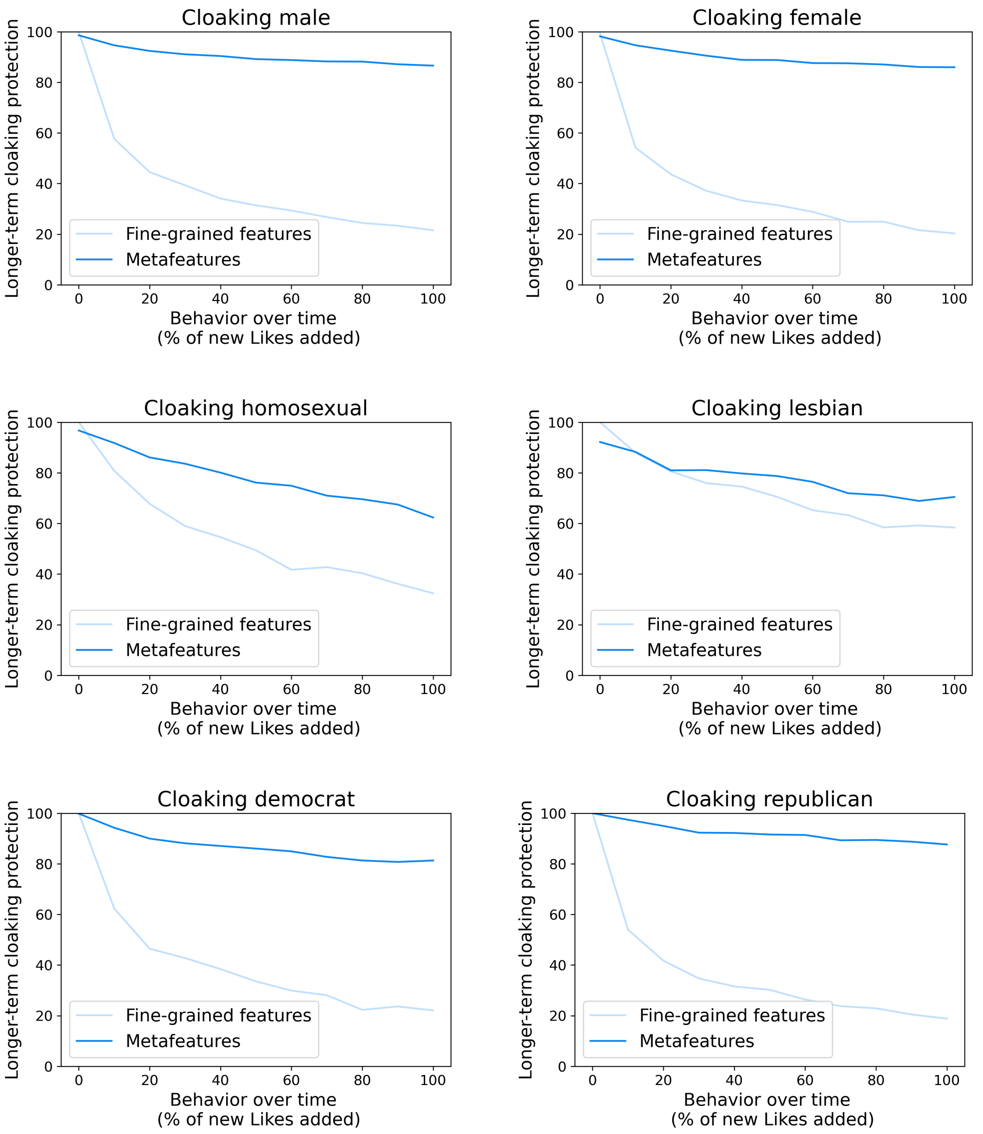
Moving beyond the specific example of John, we compare the longer-term cloaking protection of the two cloaking strategies in hiding gender, political orientation and sexual orientation. As shown in Figure 9.4, cloaking the fine-grained features offers less protection over time than cloaking the metafeatures. When using only the fine-grained features, people get targeted again relatively quickly. For example, when cloaking *male*, after adding 10% new Likes, only 57.6% of instances are still successfully cloaked. After adding all their new Likes (and thus doubling their digital traces), only 21.5% are still successfully cloaked. On the other hand, when we cloak the metafeatures instead, we see that 86.6% are still successfully cloaked when the digital traces are doubled.

We see the same patterns when cloaking *female* and political orientation (*Democrat* and *Republican*). Sexual orientation, especially *lesbian*, is more effectively cloaked over time than other tasks when using fine-grained features; cloaking the metafeatures is still a more effective longer-term cloaking strategy, but the difference between the strategies is smaller.[18] We conjecture that this could be related to the more severe class imbalance: there are fewer people whose true target label is *lesbian* in the targeted population (*True Positives*).

[16] This brings the prediction score further down to 0.120.

[17] The prediction score on the future data after cloaking the metafeatures is 0.126, which is well below the threshold of 0.194.

[18] We see that for the prediction task of *lesbian*, for a very small number of individuals, cloaking the metafeatures instead of the fine-grained can lower the number of successfully cloaked individuals, even without adding additional data.



Cloaking male

Cloaking female

Cloaking homosexual

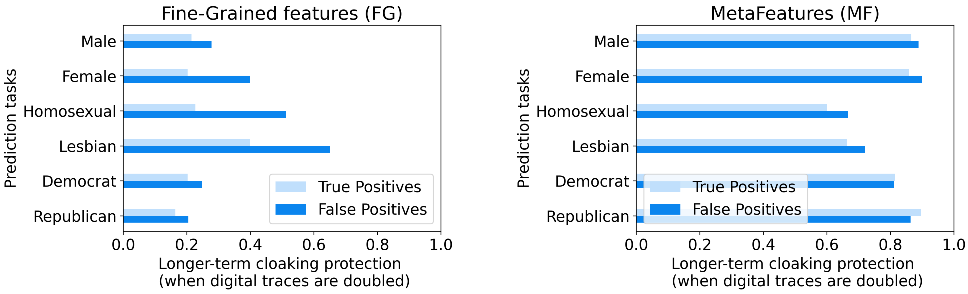
Cloaking lesbian

Cloaking democrat

Cloaking republican

Figure 9.4: Longer-term cloaking protection. We measure the longer-term cloaking protection as the percentage of positively predicted instances for which cloaking this targeting task successfully inhibits future inference. The population taken into account constitutes the intersection of individuals predicted as positive when using 1/2 of the data, and when using the full data. We measure the evolution over time on the x-axis by gradually re-adding the dropped pages.

We analyze whether there is a difference in longer-term cloaking protection between correctly predicted people (True Positives) and people who were incorrectly predicted as exhibiting the target trait (False Positives). Intuitively, we might think that people who were correctly inferred to have the predicted trait would be more likely to reveal themselves again over time. In Figure 9.5a, one can observe that the longer-term cloaking protection of True Positives is in fact lower when using the FG cloaking strategy. This aligns with intuition—that the True Positives have higher likelihood of repeating behaviors that could result in the same prediction. This difference almost disappears when we assess the longer-term cloaking protection with metafeatures (Figure 9.5b).



(a)

(b)

Figure 9.5: Is there a difference in longer-term cloaking protection between True Positives and False Positives? We measure this at the point where the digital traces have doubled.

We also present the results of two additional cloaking strategies in Appendix E.1. The first option involves using the categories assigned by Facebook to the Liked pages themselves (domain-based metafeatures). The advantage of using domain-based metafeatures is that they are readily available and by design comprehensible. However, as shown in Figure E.10, the data-driven metafeatures created by NMF are more effective in avoiding inferences over time, and in addition our analysis reveals that on average they hide fewer pages than the domain-based metafeatures. We conjecture that this is because they more accurately capture general patterns of behavior. For example, when examining the metafeatures in Table 9.4, we see that they are strongly associated with right-wing politics and Christianity, which are both highly predictive of being a Republican. On the other hand, domain-based metafeatures such as 'Public Figure' may be too general to capture these specific patterns.

A different strategy for increasing the robustness of cloaking involves adding a *tolerance level* to the initial counterfactual explanations. This means that instead of using the threshold of the decision-making system to generate the counterfactual explanations, for cloaking we employ a **lower** threshold. It is to be expected that when we bring someone just below the threshold (which is what counterfactual explanations do), the chances of them crossing the threshold again are relatively high. Therefore, we explore bringing individuals not only below the 95% threshold but also below the 90th percentile (while still using the 95th percentile as the threshold for prediction). This approach should provide an additional layer of protection from future targeting. As depicted in Figure E.11, it does indeed offer extended protection initially, but on average, the protection of the cloaking strategy still diminishes rapidly as more Likes are accumulated over time. This strategy has no impact on the pages that will be cloaked in the future, and this is clearly evident in the results. This highlights one of the major advantages of using metafeatures as the basis of a cloaking strategy.

9.6.2 *Trade-off between privacy and personalization*

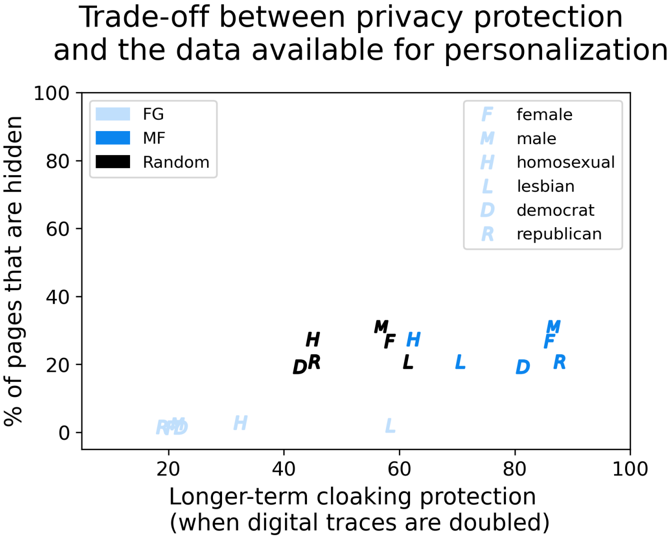
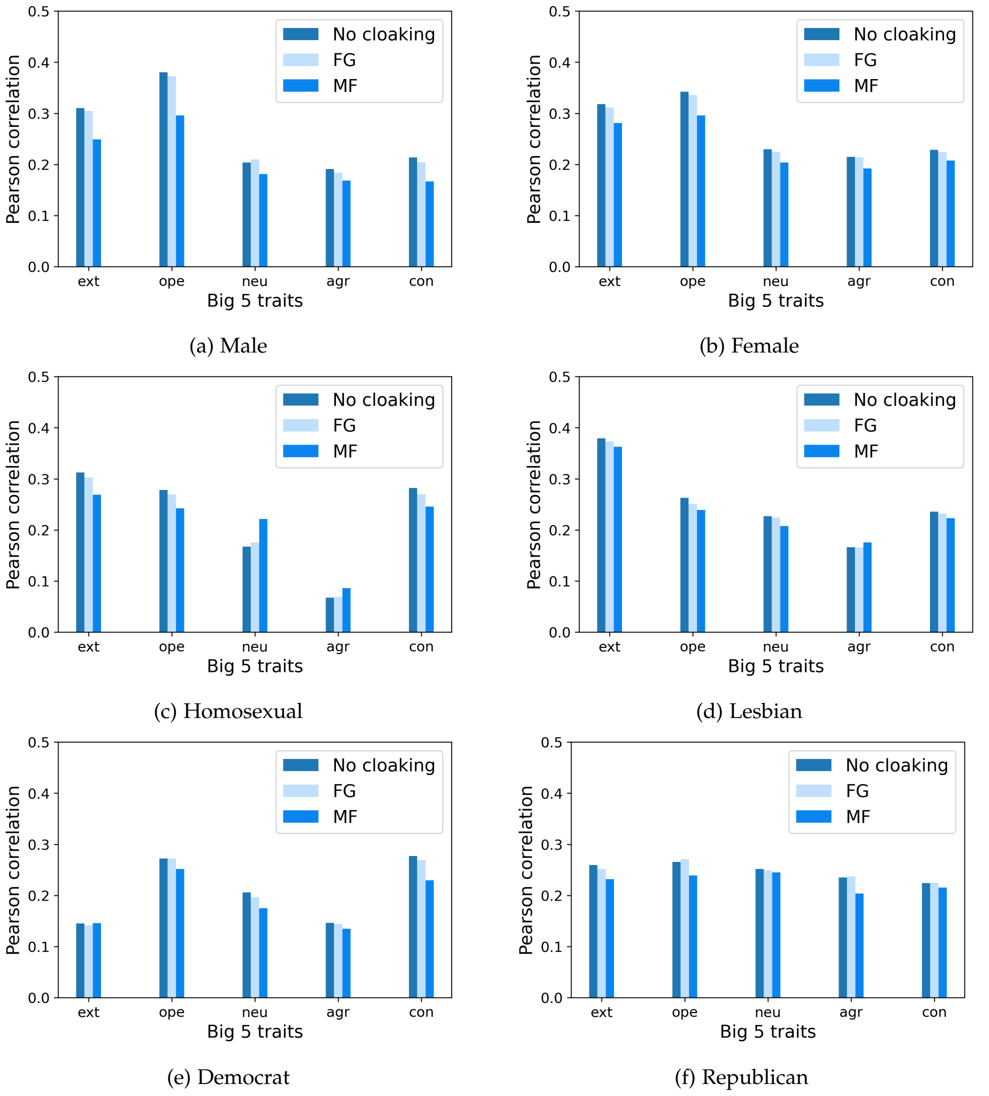


Figure 9.6: What percentage of people's digital footprints are hidden with each cloaking strategy? We measure loss in personalization by the average % of someone's Likes that have to be removed to cloak a trait, and privacy protection as the level of longer-term cloaking protection at the point when the individual's digital footprints are doubled.

Cloaking metafeatures hides larger portions of an individual's digital footprints, resulting in increased privacy protection but potentially losing the benefits of personalization. We assume users do not want to lose all personalization; otherwise, an individual could simply cloak all his Liked pages and no inferences would be made (this could still be a viable option for some users, although this will be a bad outcome from the perspective of the advertising platform). Figure 9.6 illustrates that cloaking metafeatures results in a substantial increase in privacy, but also in a substantial increase in the number of pages that are being hidden than when using fine-grained features. In the example of John, after cloaking the fine-grained features, he has 173 Likes left for personalization, while after cloaking the metafeatures, he only has 149 Likes left. Therefore it is important to assess the impact of cloaking an individual's sensitive traits on the ability to predict other things about the individual. To verify whether the additional protection from the MF strategy is not just due to the removal of more features, we also test out a *random* strategy. In this setting, we remove the features from the counterfactual explanation, and in addition remove additional features—chosen at random–to total the same number of features as removed by the metafeature cloaking strategy. Figure 9.6 shows that this random strategy (visualized in black) leads to some additional protection over the fine-grained strategy, but to significantly less protection than the metafeatures strategy for every prediction task, while the same number of Likes are removed. We note that the difference is especially large for the gender (male and female) and the political (democrat and republican) prediction tasks.



(a) Male

(b) Female

(c) Homosexual

(d) Lesbian

(e) Democrat

(f) Republican

Figure 9.7: Effect of cloaking on the predictive performance of the Big 5 traits.

We follow the set-up described in Section 9.5.2 to measure the impact of cloaking sensitive traits on the predictive performance of other prediction tasks (in this case the Big 5 traits). Figure 9.7 shows that the impact of cloaking metafeatures on the predictive performance of the Big 5 traits is larger on average than the impact of cloaking fine-grained features. For both strategies, the impact is largest when cloaking gender, followed by political orientation. The impact of both cloaking strategies on the predictive performance seems fairly small in most cases, but we cannot truly judge the losses in value (corresponding to the small losses in predictive power) in a study such as this.

9.7 discussion and conclusion

The digital traces we leave every day enable those who collect them to make intimate inferences about who we are [Kosinski et al., 2013]. While such inferences might lead to desired personalization outcomes, they also pose a considerable threat to individuals' privacy and self-determination. In this paper, we examined the effectiveness and impact of two related privacy-enhancing cloaking strategies that conceal a portion of users' digital footprints from inference procedures, in order to limit the ability of platforms to make predictions about underlying sensitive and psychological traits. Although previous work has shown that such cloaking mechanisms (Chen et al.'s specifically) can be effective in the short-term [Chen et al., 2017], our findings suggest that the corresponding cloaking effectiveness can decline rapidly over time. That is, as people continue to generate traces after the cloaking has been implemented, the system often can draw those same inferences from the new data. We introduce a new cloaking strategy—one that is based on cloaking metafeatures rather than individual footprints—and show how this strategy offers better longer-term privacy protection.

Our findings also highlight the potential trade-off between privacy protection and personalized services. That is, while individuals might be interested in cloaking certain aspects of their identity (e.g., their sexual orientation), they might appreciate the benefits they receive from sharing other aspects (e.g., their openness). We show that cloaking a particular trait likely has spillover effects on other traits that were not intentionally targeted. Although the trade-off between personalization and privacy is not a new idea [Garcia-Rivadulla, 2016], we are not aware of empirical analyses of the actual trade-offs introduced by privacy-enhancing techniques like cloaking.[19]

The extent to which trading off personalization for enhanced privacy protection is desirable will depend on the specific context and preferences of the user [Westin, 2003]. While some users might be willing to forsake targeted advertising for higher levels of privacy, others might favor convenience and service over the ability to conceal potentially unwanted aspects of their identity. The same is true for companies who might trade-off the ability to get highly granular consumer insights on all levels for a higher likelihood that consumers will stay on the platform and refrain from opting out of tracking and personalization altogether. We argue that different forms of cloaking can provide solutions that operate between the two extremes. On the one hand, they allow companies to keep collecting large amounts of data and monetize it within the boundaries set by users. On the other hand, users gain control over the level of personalization they feel comfortable with, while having the ability to inhibit unwanted inferences.

[19] Prior work has shown a trade-off between privacy protection and advertising effectiveness [Goldfarb and Tucker, 2011], and Cloarec et al. [2024] investigate this trade-off on an eHealth platform, but to our knowledge, there has not been research that evaluates this at the level of a specific prediction task.

9.7.1 *Practical Implications*

One of the main practical implications of this research is to extend the options that could be available for individuals to have control when it comes to protecting their privacy. While data protection regulations such as the General Data Protection Regulation (GDPR) in Europe or the California Consumer Privacy Act (CCPA) have pushed for increased consumer control [Harris, 2020, Van Ooijen and Vrabec, 2019], there is a growing body of research suggesting that—without any support—individuals struggle to act as responsible stewards of their personal data [Garcia-Rivadulla, 2016]. Research on the *privacy paradox*, for example, reveals a stark discrepancy between a user's expressed concerns regarding online privacy and their actual behavior when sharing personal information [Barth and De Jong, 2017, Kokolakis, 2017]. Despite expressing concerns about their privacy, individuals are often willing to share personal information online in exchange for personalized recommendations [Barth and De Jong, 2017]. For instance, even though 93% of USA citizens consider it important to maintain control over who can access their data, only a small fraction of people actually read the privacy policies of the services collecting their data [Matz et al., 2020, Madden and Rainie, 2015, Solove, 2012]. One (obvious) reason why consumers do not succeed in achieving desired levels of privacy is their lack of knowledge about how their data is actually being collected and used [Acquisti et al., 2020]. Related is the *acceptability gap*, which shows that users are more accepting of personalized services than of the collection of personal data required for these services [Kozyreva et al., 2021]. They overlook the relationship between them, and as a result, fail to engage in an adequate comparison of the value received from personalization to the value of keeping data private Kozyreva et al. [2021]. As a consequence, most people tend to overvalue the short-term benefits of actions, such as using an app, over the long-term privacy risks, which are delayed and intangible [Acquisti, 2004].

Complicating matters further, research has suggested that people's apparent inaction regarding their privacy is also the result of them feeling that they have no control over the situation, and as a consequence simply give up (a phenomenon researchers have called *digital resignation*) [Acquisti et al., 2020, Draper and Turow, 2019]. Finally, it may simply be that the perceived cost of protecting privacy is simply too high: either not using a service or possibly navigating an ultra-complicated web of documents and settings. In all of these cases, providing transparency into how data is used and control over its use seems vital for consumer welfare and, in particular, for users to make informed privacy decisions [Matz et al., 2020]. Both privacy and transparency are essential prerequisites for establishing a trustworthy AI system [Liu et al., 2022].

In this paper, we analyze and extend a tool that could help guide individuals in making choices on their privacy settings online. Since the implications of sharing personal data are often difficult to anticipate, let alone to trade off for immediate convenience rewards, we need easy ways for people to move the dial between oversharing and undersharing. Cloaking offers such a lever and might encourage platforms to offer more mechanisms for users to control data-driven inferences and personalization (including targeted advertising) either through editing of the data items—the digital footprints—that are stored about them or through an explicit cloaking mechanism that hides footprints from the AI inference systems specifically.

Notably, the cloaking methodology depends on the cooperation of the platforms that collect this kind of data (like Facebook, Google, Spotify). While platforms might try to resist the introduction of technology that limits their ability to commercialize consumer insights, we argue that introducing increased consumer control in the form of cloaking could eventually benefit them in the long-run. As stricter data protection regulations are introduced around the world—often empowering consumers to revoke access to their personal data—platforms might be forced to provide sufficient transparency and control in order to retain users and prevent them from opting out of data collection entirely. Moreover, a gradual shift to higher levels of platform-supported user control might prevent legislators from introducing more paternalistic regulatory actions.

9.7.2 *Limitations*

Of course there are limitations to this study. First of all, note that we use an archival dataset of Facebook Likes, which was collected until 2012. This raises the question how valid these findings remain today. Note that the sociodemographics of Facebook users is different from that of users of Instagram or Tiktok, and from the general population [Cavalcante, Kaylan, 2023, Ribeiro et al., 2020]. Additionally, we only look at Facebook Likes, while Kim and Yang [2017] shows that other patterns of behavior on Facebook such as *following*, *commenting* or *sharing* are distinctly different from Facebook Likes and are driven by different things. It would be an interesting avenue of future research to repeat this study on a more recent dataset of Facebook traces (not necessarily Likes) or on data from other platforms such as Tiktok or Instagram.

Another limitation that we already shortly addressed in Section 9.5.1, is the non-temporality of the data. Our simulation assumes that people's behavior is stable over time, but it would be interesting to conduct this analysis and verify the results on actual time-stamped data. The metafeatures would have to be updated continuously, when new Like pages are created, and users Like more things. However, from the viewpoint of the platform, this would still be a lot more efficient than recloaking every undesired inference once new pages and Likes become available.

A last potential limitation in our methodology is the use of static thresholds on the output score from the machine learned models. To our understanding, the most common practice is to use fixed thresholds. However, thresholds are changed for various reasons, and more sophisticated advertising systems may dynamically adjust targeting thresholds—for example, to allocate a budget intelligently over the budget period, to deal with time-of-day differences, etc. In addition, we cannot know the targeting thresholds that are actually in place, as this information is proprietary. For our experiments, we base our threshold for positive inference on the 95th percentile of the scores over the training set, which is in line with the literature [Chen et al., 2017, Perlich et al., 2014]. In the case of dynamic thresholds, the threshold used for cloaking would have to be considered carefully. One would want to take into account the potential threshold range and cloak based on some notion of the expected minimum threshold that might be used.

Part V

CONCLUSION

10

Conclusion

This final Chapter summarizes the main empirical and methodological contributions from this thesis, discusses some general limitations, and presents directions for future research.

10.1 Main findings

"Technology is neither good nor bad, nor is it neutral."

Kranzberg's First Law of Technology

In this thesis, I explore the potential impact that Explainable AI techniques can have on society, highlighting the dual nature of their effects. I want to emphasize that these techniques can both bring positive advancements and harbor potential risks, depending on the application. My research discusses both these effects

First, in Chapter 4, I investigate the necessity of Explainable AI by comparing the performance of black-box and white-box models. My findings indicate that, on average, black-box models exhibit superior performance; however, the performance difference is often marginal. Hence, I advocate for the use of white-box models in contexts where transparency is considered critical, while reserving black-box models and explainable AI techniques for other settings.

My research also highlights potential dangers, as discussed in Chapters 5 and 8. In Chapter 5, I demonstrate the potential of manipulation arising from the abundance of possible explanations for each situation. This leads to considerable influence for the explanation provider, and I explore some scenarios in which this might lead the provider to behave unethically. Additionally, Chapter 8 focuses on the privacy implications that counterfactual explanations may entail, potentially revealing sensitive information about data subjects.

However, amidst these challenges, I also identify some benefits of employing Explainable AI, and counterfactual explanations in specific in Chapters 6 and 9. In Chapter 6, I illustrate how counterfactual explanations can serve as a tool for assessing bias in machine learning models, thereby contributing to avoiding discriminatory outcomes. In Chapter 9, I analyze how digital users can use XAI-tools to avoid undesired inferences based on the behavioral traces they leave behind. More in general, in Chapter 7, I note that the impact of bias mitigation methods is very opaque, and that more transparency is needed in their operational dynamics and who they impact.

In conclusion, my thesis highlights the significance of Explainable AI and transparency in Machine Learning as powerful tools with the potential to both advance and harm society. Through my evaluation of the challenges, and opportunities, I advocate for responsible and context-aware deployment of ML and XAI techniques.

10.2 LIMITATIONS

While I address the limitations specific to each study within each chapter, I will also discuss some general limitations of my studies and the overarching research domain here.

First off, the majority of my work is positioned within the realm of XAI, and it is important to recognize that many of the techniques described are far from perfect and that using them can create additional challenges. A significant concern within XAI is the mentioned 'disagreement problem,' where the multiplicity of potential explanations for a single phenomenon undermines the reliability of any given explanation. This issue, detailed in Chapter 5, suggests that the selection of explanations could be biased, leading to a preference for justifications that align with the selector's views or intentions [Bordt et al., 2022]. Additionally, the phenomenon of the 'anchoring effect,' as explored by Chu et al. [2020] presents another challenge. Their research demonstrates that individuals exposed to arbitrary explanations from flawed machine learning models tend to develop unwarranted trust in these models, irrespective of their validity. These concepts highlight that we need to remain skeptical when using XAI.

Secondly, the aspiration to mitigate complex sociological dilemmas through AI introduces the risk of 'techno-solutionism,' where technology is perceived as a universal remedy [Morozov, 2013]. This perspective fails to recognize that issues such as transparency, fairness, and privacy are deeply entrenched in societal and systemic structures, and thus, cannot be fully resolved through technological means alone. Although the methods I describe in my thesis are based in the field of computer science, for their application collaboration with legal experts, social scientists, policymakers, and other relevant fields is needed. Many of these decisions should not be up to data scientists alone, and I emphasize the need for ongoing societal and policy engagement. It is also necessary to question whether it is even appropriate to use an algorithm in the first place; in some settings, we might not want to relinquish the control to a black-box model. However, it is equally important to note that idealizing human judgment as a superior alternative ignores the inherent biases and complexities characterizing human decision-making processes. One should carefully consider whether to opt for human or algorithmic decision-making, as both can be flawed. The best approach may vary depending on the specific application.

Lastly, specific to the field of fairness, the frameworks in both Chapter 6 and Chapter 7 assume access to a static, sensitive attribute. In reality, access to these attributes might be hard to obtain, as legal and ethical constraints may impose constraints on obtaining or utilizing certain sensitive information [Haeri and Zweig, 2020, Veale and Binns, 2017, Johnson, 2021, Holstein et al., 2019]. Additionally, as society is evolving, some of these sensitive attributes are no longer static. Take for example gender: individuals may change gender or may not identify as one of the binary gender categories. This is an obvious limitations of both my studies in the fairness domain but also opens up opportunities for future research, which I will discuss in Section 10.3.

10.3 future research directions

Given the relevance of ethical AI and the continuous emergence of new techniques, the potential for future research in this domain is extensive. I foresee several directions for future research (again besides specific questions that are mentioned in each study separately).

First, building on the contributions presented in Chapter 5, there are numerous possibilities for further investigation into the disagreement problem. One potential avenue is to explore the extent to which the non-deterministic nature of the explanation algorithms influences the obtained results. It would also be interesting to investigate the factors that contribute to disagreement, such as specific data instances or machine learning models. For instance, analyzing whether certain models yield greater diversity in explanations or if data instances near decision boundaries result in increased disagreement would be intriguing research inquiries. Furthermore, besides investigating the issue further, it is also worthwhile to think about potential solutions. It would be useful for regulators if there would be a technique that could assess how likely it is that an explanation is manipulated, and not generated in a standard manner. Developing a framework or methodology to assess this would be a valuable research contribution.

Second, most of my research focuses on applications of tabular data. Only in Chapter 9, I use a behavioral dataset (Facebook likes). Working with tabular or behavioral datasets leads to very different challenges, as tabular datasets are structured datasets with a limited number of attributes that are typically consciously collected, while behavioral datasets portray a picture of people's behavior over time and are generated automatically though tracking systems [De Cnudde et al., 2019]. In the future, I would like to extend some of my findings to this type of datasets as well. Addressing fairness in this domain present unique challenges and has remained largely unexplored. Given the complexity of contextual biases in behavioral data, transparency mechanisms will be of vital importance to evaluate them.

A third challenge I aim to address in the fairness area is linked to one of the limitations I just mentioned, namely scenarios where there is no access to the sensitive attribute . Most existing fairness metrics and mitigation strategies presuppose the availability of this attribute, but in practical situations this assumption may not hold. Moreover, as society evolves, new forms of discrimination can emerge beyond the traditionally protected attributes [Wachter, 2022]. To tackle this issue, I aim to develop solutions that do not rely on explicit access to a sensitive attribute. My approach would include various techniques, including clustering, uncertainty estimation, and explainable artificial intelligence (XAI) methods, to identify and address discriminated subgroups, even when the sensitive attribute is unobservable or rapidly changing. The goal of this research is to enhance the fairness and equity of algorithms in situations where traditional fairness metrics fall short due to a lack of access to the sensitive attribute.

A last research area that I want to move forward in, is the direction of 'Fair AI In Practice'. As outlined in Chapter 7, many of the bias mitigation methods result in scenarios that do not align with real-world applications. I want to contribute to frameworks that can be readily applied by practitioners. One of the research goals I have here is studying the cost of *resource-constrained* fairness. In most applications of fair machine learning, the resources are approximately fixed (think about student admission, healthcare screening or hiring), and current estimates of the cost do not take this into account. Being fair in a resource-constrained context entails redistributing resources from privileged groups to protected groups, which will be associated with costs and benefits for both groups. Additionally, I would like to conduct various case studies where I examine my findings in real-world applications, starting from the data collection to the implementation of the machine learning model.

Bibliography

Broad agency announcement, explainable artifcial intelligence (xai). https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf, 2016. Accessed: 2020-11-12.

Waller, Angie and Lecher, Colin. Facebook promised to remove "sensitive" ads. here's what it left behind. https://themarkup.org/citizen-browser/2022/05/12/facebook-promised-to-remove-sensitive-ads-heres-what-it-left-behind, 2022. Accessed: 2023-03-13.

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Alessandro Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 21–29, 2004.

Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology*, 30(4):736–758, 2020.

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *The Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

Aditi Agrawal. New york regulator orders probe into goldman sachs' credit card practices over apple card and sexism. https://www.medianama.com/2019/11/223-apple-card-sexism-goldman-sachs/, November 12, 2019. Medianama, Online, accessed February 1, 2022.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. Characterizing the risk of fairwashing. *Advances in Neural Information Processing Systems*, 34:14822–14834, 2021.

Subhi J Al'Aref, Gurpreet Singh, Alexander R van Rosendael, Kranthi K Kolli, Xiaoyue Ma, Gabriel Maliakal, Mohit Pandey, Bejamin C Lee, Jing Wang, Zhuoran Xu, et al. Determinants of in-hospital mortality after percutaneous coronary intervention: a machine learning approach. *Journal of the American Heart Association*, 8(5):e011160, 2019.

Edesio Alcobaça, Felipe Siqueira, Adriano Rivolli, Luís Paulo F Garcia, Jefferson Tales Oliva, André CPLF de Carvalho, et al. Mfe: Towards reproducible meta-feature extraction. *J. Mach. Learn. Res.*, 21:111–1, 2020.

Hiva Allahyari and Niklas Lavesson. User-oriented assessment of classification model understandability. In *11th scandinavian conference on Artificial intelligence*. IOS Press, 2011.

Xavier Amatriain and Justin Basilico. Netflix recommendations: Beyond the 5 stars., 2020. URL https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429.

Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the facebook advertising ecosystem. In *NDSS 2019-Proceedings of the Network and Distributed System Security Symposium*, pages 1–15, 2019.

Julia Angwin and Terry Parris Jr. Facebook lets advertisers exclude users by rac. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race, 2016. Accessed: 2023-03-13.

Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6679–6687, 2021.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

André Artelt, Valerie Vaquet, Riza Velioglu, Fabian Hinder, Johannes Brinkrolf, Malte Schilling, and Barbara Hammer. Evaluating robustness of counterfactual explanations. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 01–09. IEEE, 2021.

Irit Askira-Gelman. Knowledge discovery: comprehensibility of the results. In *Proceedings of the thirty-first Hawaii international conference on system sciences*, volume 5, pages 247–255. IEEE, 1998.

Arthur Asuncion and David Newman. UCI Machine Learning Repository, 2007.

Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, Liam Murphy, et al. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Transactions on data privacy*, 7(3):337–370, 2014.

Danny Azucar, Davide Marengo, and Michele Settanni. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences*, 124:150–159, 2018.

Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

Hubert Baniecki, Wojciech Kretowicz, and Przemyslaw Biecek. Fooling partial dependence via data poisoning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part III*, pages 121–136. Springer, 2023.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.

Alistair Barr. Google mistakenly tags black people as gorillas, showing limits of algorithms., 2015.

Susanne Barth and Menno DT De Jong. The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and informatics*, 34(7):1038–1058, 2017.

Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. Bias on demand: A modelling framework that generates synthetic data with bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1002–1013, 2023.

Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228. IEEE, 2005.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

Richard Berk. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.

Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *ESANN*, 2016.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020.

Emily Black, Samuel Yeom, and Matt Fredrikson. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1): 1–21, 2017.

Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 891–905, 2022.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, and Marcin Detyniecki. Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 188–203, 2023.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001a.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001b.

Dieter Brughmans, Pieter Leyman, and David Martens. Nice: an algorithm for nearest instance counterfactual explanations. *Data Mining and Knowledge Discovery*, pages 1–39, 2023a.

Dieter Brughmans, Lissa Melis, and David Martens. Disagreement amongst counterfactual explanations: How transparency can be deceptive. *arXiv preprint arXiv:2304.12667*, 2023b.

Tobias Budig, Selina Herrmann, and Alexander Dietz. Trade-offs between privacy-preserving and explainable machine learning in healthcare. *cii Student Papers-2021*, page 59, 2021.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018.

José González Cabañas, Ángel Cuevas, and Rubén Cuevas. Facebook use of sensitive data for advertising in europe. *arXiv preprint arXiv:1802.05030*, 2018.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18. IEEE, 2009.

Alison Callahan and Nigam H Shah. Machine learning in healthcare. In *Key Advances in Clinical Informatics*, pages 279–291. Elsevier, 2017.

Rachele Carli, Amro Najjar, and Davide Calvaresi. Risk and exposure of xai in persuasion and argumentation: The case of manipulation. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 204–220. Springer, 2022.

E. Carrizosa, J. Ramírez Ayerbe, and D. Romero Morales. Mathematical optimization modelling for group counterfactual explanations. Forthcoming in *European Journal of Operational Research* https://www.researchgate.net/publication/368958766_Mathematical_Optimization_Modelling_for_Group_Counterfactual_Explanations, 2024a.

Emilio Carrizosa, Amaya Nogales-Gómez, and Dolores Romero Morales. Clustering categories in support vector machines. *Omega*, 66:28–37, 2017.

Emilio Carrizosa, Jasone Ramírez-Ayerbe, and Dolores Romero Morales. Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Systems with Applications*, 238:121954, 2024b.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 2020.

Cavalcante, Kaylan. Facebook influencer marketing: Complete guide for brands. https://insense.pro/blog/facebook-influencer-marketing, 2023. Accessed: 2024-06-13.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. pages 292–303. IEEE, 2021.

Kiran Chaudhary, Mansaf Alam, Mabrook S Al-Rakhami, and Abdu Gumaei. Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics. *Journal of Big Data*, 8(1):1–20, 2021.

Rahul Chavan. Understanding train, test, and validation dataset split in simple, quick terms, 2023. URL https://medium.com/@rahulchavan4894/understanding-train-test-and-validation-dataset-split-in-simple-quick-terms-5a8630fe58c8.

Ramnath K Chellappa and Raymond G Sin. Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information technology and management*, 6: 181–202, 2005.

Daizhuo Chen, Samuel P Fraiberger, Robert Moakler, and Foster Provost. Enhancing transparency and control when drawing data-driven inferences about individuals. *Big data*, 5(3): 197–212, 2017.

Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology*, 32(4):1–30, 2023.

Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

Karine Chubarian and György Turán. Interpretability of bayesian network classifiers: Obdd approximation and polynomial threshold functions. In *ISAIM*, 2020.

Jessica Clark and Foster Provost. Unsupervised dimensionality reduction versus supervised regularization for classification from sparse data. *Data Mining and Knowledge Discovery*, 33: 871–916, 2019.

Julien Cloarec, Charlotte Cadieu, and Nour Alrabie. Tracking technologies in eHealth: Revisiting the personalization-privacy paradox through the transparency-control framework. *Technological Forecasting and Social Change*, 200:123101, 2024.

William W Cohen. Fast effective rule induction. In *Machine learning proceedings 1995*, pages 115–123. Elsevier, 1995.

Roberto Confalonieri, Tillman Weyde, Tarek R Besold, and Fermín Moscoso del Prado Martín. Trepan reloaded: A knowledge-driven approach to explaining artificial neural networks. *arXiv preprint arXiv:1906.08362*, 2019.

Constanza Contreras-Piña and Sebastián A Ríos. An empirical comparison of latent sematic models for applications in industry. *Neurocomputing*, 179:176–185, 2016.

A Feder Cooper, Solon Barocas, Christopher De Sa, and Siddhartha Sen. Variance, self-consistency, and arbitrariness in fair classification. *arXiv preprint arXiv:2301.11562*, pages 1–84, 2023.

Madison Coots, Soroush Saghafian, David Kent, and Sharad Goel. Reevaluating the role of race and ethnicity in diabetes screening. *arXiv preprint arXiv:2306.10220*, 2023.

Sam Corbett-Davies, Johann D Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1): 14730–14846, 2023.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. *EUROSIS-ETI*, pages 5–12, 2008.

Paul T Costa and Robert R McCrae. Normal personality assessment in clinical practice: The neo personality inventory. *Psychological assessment*, 4(1):5, 1992.

Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8:24–30, 1995.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *Parallel Problem Solving from Nature–PPSN XVI: 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I*, pages 448–469. Springer, 2020.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

Sofie De Cnudde, Yanou Ramon, David Martens, and Foster Provost. Deep learning on big, sparse, behavioral data. *Big data*, 7(4):286–307, 2019.

Sofie De Cnudde, David Martens, Theodoros Evgeniou, and Foster Provost. A benchmarking study of classification techniques for behavioral data. *International journal of data science and analytics*, 9(2):131–173, 2020.

Enric Junqué de Fortuny and David Martens. Active learning-based pedagogical rule extraction. *IEEE transactions on neural networks and learning systems*, 26(11):2664–2677, 2015.

Raphael Mazzine Barbosa de Oliveira and David Martens. A framework and benchmarking study for counterfactual generating methods on tabular data. *Applied Sciences*, 11(16):7274, 2021.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

Janez Demšar, Tomaž Curk, Aleš Erjavec, Črt Gorup, Tomaž Hočevar, Mitar Milutinovič, Martin Možina, Matija Polajnar, Marko Toplak, Anže Starič, et al. Orange: data mining toolbox in python. *the Journal of machine Learning research*, 14(1):2349–2353, 2013.

Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. 2020.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

Jamie Doward and Alice Gibbs. Did Cambridge Analytica influence the Brexit vote and the US election. *The Guardian*, 4(3), 2017.

Oran Doyle. Direct discrimination, indirect discrimination and autonomy. *Oxford Journal of Legal Studies*, 27(3):537–553, 2007.

Nora A Draper and Joseph Turow. The corporate cultivation of digital resignation. *New media & society*, 21(8):1824–1839, 2019.

Kirill Dubovikov. *Managing Data Science*. Packt Publishing, November 2019. ISBN 9781838826321.

Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer, 2006.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Edelman, Gilad. How facebook's political ad system is designed to polarize. https://www.wired.com/story/facebook-political-ad-system-designed-polarize/, 2019. Accessed: 2023-03-13.

Khaled El Emam and Fida Kamal Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637, 2008.

Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, et al. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

Mark Elliot and Angela Dale. Scenarios of attack: the data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics*, 14(Spring):6–10, 1999.

European Commission. 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. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence, 2021. European Commission, Online, accessed February 24, 2022.

Facebook. Good questions, real answers: How does facebook use machine learning to deliver ads? https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads, 2020. Accessed: 2024-01-31.

Marco Favier, Toon Calders, Sam Pinxteren, and Jonathan Meyer. How to be fair? a study of label and selection bias. *Machine Learning*, 112(12):5081–5104, 2023.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Thomas A Feo and Mauricio GC Resende. Greedy randomized adaptive search procedures. *Journal of global optimization*, 6(2):109–133, 1995.

Carlos Fernandez, Foster Provost, and Xintian Han. Counterfactual explanations for data-driven decisions. In *40th International Conference on Information Systems, ICIS 2019*. Association for Information Systems, 2020.

Carlos Fernández-Loría, Foster Provost, and Xintian Han. Explaining data-driven decisions made by ai systems: The counterfactual approach. *MIS Quarterly*, 46(3):1635–1660, 2022.

Carlos Fernández-Loría, Foster Provost, Jesse Anderton, Benjamin Carterette, and Praveen Chandar. A comparison of methods for treatment assignment with an application to playlist generation. *Information Systems Research*, 34(2):786–803, 2017.

Will Fleisher. What's fair about individual fairness? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 480–490, 2021.

Hortense Fong, Vineet Kumar, Anay Mehrotra, and Nisheeth K Vishnoi. Fairness for auc via feature augmentation. *arXiv preprint arXiv:2111.12823*, 2021.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.

Alex A Freitas. Automated machine learning for studying the trade-off between predictive accuracy and interpretability. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 48–66. Springer, 2019.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

Roland G Fryer Jr, Glenn C Loury, and Tolga Yuret. An economic analysis of color-blind affirmative action. *The Journal of Law, Economics, & Organization*, 24(2):319–355, 2008.

Archon Fung and Erik Olin Wright. Deepening democracy: Innovations in empowered participatory governance. *Politics & society*, 29(1):5–41, 2001.

Glenn Fung, Sathyakama Sandilya, and R Bharat Rao. Rule extraction from linear support vector machines. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 32–40, 2005.

Isel del Carmen Grau García. *Self-labeling Grey-box Model: An Interpretable Semi-supervised Classifier*. Ph.d. thesis, Queens University Belfast, United Kingdom, 2020.

Sandra Garcia-Rivadulla. Personalization vs. privacy: An inevitable trade-off? *IFLA journal*, 42(3):227–238, 2016.

Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1296. PMLR, 2020.

R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 325–336, 2020.

Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769, 2007.

Gabriel Ghinita, Panagiotis Karras, Panos Kalnis, and Nikos Mamoulis. A framework for efficient data anonymization under privacy and accuracy constraints. *ACM Transactions on Database Systems (TODS)*, 34(2):1–47, 2009.

Leilani H Gilpin, Andrew R Paley, Mohammed A Alam, Sarah Spurlock, and Kristian J Hammond. " explanation" is not a technical term: The problem of ambiguity in xai. *arXiv preprint arXiv:2207.00007*, 2022.

Aristides Gionis and Tamir Tassa. k-anonymization with minimal loss of information. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):206–219, 2008.

Sofie Goethals. How counterfactual explanations can be used to detect bias in a machine learning model. EURO: 32th European Conference on Operational Research, Espoo, Finland, 2022.

Sofie Goethals. Explainability methods to measure discrimination in machine learning models. BIAS workshop at ECML: Third Workshop on Bias and Fairness in AI, Turin, Italy, 2023a.

Sofie Goethals. The trade-offs of obscuring your digital footprints. ECDA: European Conference on Data Analysis, Antwerp, Belgium, 2023b.

Sofie Goethals. Explainability methods to measure discrimination in machine learning models. EWAF: European Workshop on Algorithmic Fairness, Winterthur, Switzerland, 2023c.

Sofie Goethals. The trade-offs of obscuring your digital footprints. ORBEL: 36th Annual Conference of the Belgian Operational Research Society, Liege, Belgium, 2023d.

Sofie Goethals. Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem. XKDD workshop at ECML: International Workshop on Explainble Knowledge Discovery in Data Mining, Turin, Italy, 2023e.

Sofie Goethals and Toon Calders. Reranking individuals: The effect of fair classification within-groups. Poster in the European Workshop on Algorithmic Fairness, Mainz, Germany, 2024.

Sofie Goethals, David Martens, and Theodoros Evgeniou. The Non-linear Nature of the Cost of Comprehensibility. *Journal of Big Data*, 9(1):1–23, 2022.

Sofie Goethals, Travis Greene, David Martens, and Galit Shmueli. Algorithmic Explanations as Ad Opportunities. Poster in the Workshop on Decision Intelligence and Analytics for Online Marketplaces at KDD, Long Beach, California, 2023a.

Sofie Goethals, David Martens, and Toon Calders. PreCoF: Counterfactual Explanations for Fairness. *Machine Learning*, pages 1–32, 2023b.

Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–24, 2023c.

Sofie Goethals, Toon Calders, and David Martens. Beyond Accuracy-Fairness: Stop evaluating bias mitigation methods solely on between-group metrics. *Under review*, 2024a.

Sofie Goethals, Eoin Delaney, Brent Mittelstadt, and Chris Russell. Resource-constrained fairness. *Under review*, 2024b.

Sofie Goethals, David Martens, and Theodoros Evgeniou. Manipulation Risks in Explainable AI: The Implications of the Disagreement Problem. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (forthcoming)*. Springer, 2024c.

Sofie Goethals, Sandra Matz, Foster Provost, Yanou Ramon, and David Martens. The Impact of Cloaking Digital Footprints on User Privacy and Personalization. *Under review*, 2024d.

Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable AI: current status and future directions. *arXiv preprint arXiv:2107.07045*, 2021.

Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in personality*, 40(1): 84–96, 2006.

Avi Goldfarb and Catherine E Tucker. Privacy regulation and online advertising. *Management science*, 57(1):57–71, 2011.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Travis Greene, Sofie Goethals, David Martens, and Galit Shmueli. Monetizing Explainable AI: A double-edged sword. *Under review*, 2023.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Benjamin Habegger, Omar Hasan, Lionel Brunie, Nadia Bennani, Harald Kosch, and Ernesto Damiani. Personalization vs. privacy in big data analysis. *International Journal of Big Data*, pages 25–35, 2014.

Maryam Amir Haeri and Katharina Anna Zweig. The crucial role of sensitive attributes in fair classification. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2993–3002. IEEE, 2020.

Alaa Hamoud. Selection of best decision tree algorithm for prediction and classification of students' action. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, 16(1):26–32, 2016.

David J Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123, 2009.

Moritz Hardt and Solon Barocas. Fairness in machine learning. In *Neural Information Processing Symposium, Tutorials Track*, 2017.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

Rebecca Harris. Forging a path towards meaningful digital privacy: Data monetization and the CCPA. *Loy. LAL Rev.*, 54:197, 2020.

Joseph F. Healey, Andi Stepnick, and Eileen O'Brien. *Race, ethnicity, gender, & class : the sociology of group conflict and change*. SAGE, eigth edition, 2019.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32, 2019.

LaTasha Hill. Less talk, more action: How law schools can counteract racial bias of LSAT scores in the admissions process. *U. Md. LJ Race, Religion, Gender & Class*, 19:313, 2019.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 13–38. Springer, 2022.

Max Hort, Jie M Zhang, Federica Sarro, and Mark Harman. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 994–1006, 2021.

Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 2023.

Knut T Hufthammer, Tor H Aasheim, Sølve Ånneland, Håvard Brynjulfsen, and Marija Slavkovik. Bias mitigation with aif360: A comparative study. In *Norsk IKT-konferanse for forskning og utdanning*, number 1, 2020.

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288, 2002.

Shomik Jain, Kathleen Creel, and Ashia Wilson. Scarce resource allocations that rely on machine learning should be randomized. *arXiv preprint arXiv:2404.08592*, 2024.

Patrick Janssen and Bert M Sadowski. Bias in algorithms: On the trade-off between accuracy and fairness. 2021.

Ulf Johansson, Cecilia Sönströd, Tuve Löfström, and Henrik Boström. Obtaining accurate and comprehensible classifiers using oracle coaching. *Intelligent Data Analysis*, 16(2):247–263, 2012.

Ulf Johansson, Cecilia Sönströd, and Rikard König. Accurate and interpretable regression trees using oracle coaching. In *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 194–201. IEEE, 2014.

Gabbrielle M Johnson. Algorithmic bias: on the implicit biases of social technology. *Synthese*, 198(10):9941–9961, 2021.

Garrett A Johnson, Scott K Shriver, and Shaoyin Du. Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, 39(1):33–51, 2020.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

Enric Junqué de Fortuny, David Martens, and Foster Provost. Predictive modeling with big data: is bigger really better? *Big data*, 1(4):215–226, 2013.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Faisal Kamiran and Indrė Žliobaitė. Explainable and non-explainable discrimination in classification. In *Discrimination and Privacy in the Information Society*, pages 155–170. Springer, 2013.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*, pages 924–929. IEEE, 2012.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, 2013.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys (CSUR)*, 2021.

Ujwal Kayande, Arnaud De Bruyn, Gary L Lilien, Arvind Rangaswamy, and Gerrit H Van Bruggen. How incorporating feedback mechanisms in a dss affects dss evaluations. *Information Systems Research*, 20(4):527–546, 2009.

Mark T Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai). In *International Conference on Case-Based Reasoning*, pages 163–178. Springer, 2020.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Maurice George Kendall. Rank correlation methods. 1948.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning.

Advances in Neural Information Processing Systems, 30, 2017.

Cheonsoo Kim and Sung-Un Yang. Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review*, 43(2):441–449, 2017.

Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems*, 31, 2018.

Pauline T Kim. Auditing algorithms for discrimination. *U. Pa. L. Rev. Online*, 166:189, 2017.

Slava Kisilevich, Yuval Elovici, Bracha Shapira, and Lior Rokach. kactus 2: Privacy preserving in classification tasks using k-anonymity. In *Annual Workshop on Information Privacy and National Security*, pages 63–81. Springer, 2008.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017.

Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *International conference on machine learning*, pages 2737–2746. PMLR, 2018.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15):5802–5805, 2013.

Anastasia Kozyreva, Philipp Lorenz-Spreen, Ralph Hertwig, Stephan Lewandowsky, and Stefan M Herzog. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from germany, great britain, and the united states. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.

Natasa Krco, Thibault Laugel, Jean-Michel Loubes, and Marcin Detyniecki. When mitigating bias is unfair: A comprehensive study on the impact of bias mitigation algorithms. *arXiv preprint arXiv:2302.07185*, 2023.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

Michał Kuźba and Przemysław Biecek. What would you ask the machine learning model? identification of user needs for model explanations based on human-model conversations. In *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 447–459. Springer, 2020.

Kweku Kwegyir-Aggrey, Jessica Dai, A Feder Cooper, John Dickerson, Keegan Hines, and Suresh Venkatasubramanian. Repairing regressors for fair binary classification at any decision threshold. In *NeurIPS 2023 Workshop Optimal Transport and Machine Learning*, 2023.

Carmen Lacave and Francisco J Díez. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127, 2002.

Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294*, 2019.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1452, 2022.

Lecher, Colin. How big pharma finds sick users on facebook. https://themarkup.org/citizen-browser/2021/05/06/how-big-pharma-finds-sick-users-on-facebook, 2021. Accessed: 2023-03-13.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Eunjin Lee, David Braines, Mitchell Stiffler, Adam Hudler, and Daniel Harborne. Developing the sensitivity of lime for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pages 349–356. SPIE, 2019.

Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60, 2005.

Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*, pages 25–25. IEEE, 2006.

Daphne Lenders and Toon Calders. Real-life performance of fairness interventions-introducing a new benchmarking dataset for fair ml. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 350–357, 2023.

Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pages 106–115. IEEE, 2006.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2021.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

Haochen Liu, Yiqi Wang, Wenqi Fan, Xiaorui Liu, Yaxin Li, Shaili Jain, Yunhao Liu, Anil Jain, and Jiliang Tang. Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, 14(1):1–59, 2022.

Ana C Lorena, Luís PF Garcia, Jens Lehmann, Marcilio CP Souto, and Tin Kam Ho. How complex is your classification problem? a survey on measuring classification complexity. *ACM Computing Surveys (CSUR)*, 52(5):1–34, 2019.

Grigorios Loukides and Aris Gkoulalas-Divanis. Utility-preserving transaction data anonymization with low information loss. *Expert systems with applications*, 39(10):9764–9777, 2012.

Stella Lowry and Gordon Macpherson. A blot on the profession. *British medical journal (Clinical research ed.)*, 296(6623):657, 1988.

Julián Luengo and Francisco Herrera. An automatic extraction method of the domains of competence for learning classifiers using data complexity measures. *Knowledge and Information Systems*, 42(1):147–180, 2015.

Ville Lukka and Paul TJ James. Attitudes toward facebook advertising. *Journal of management and Marketing Research*, 14:1, 2014.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1): 2522–5839, 2020.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

Mary Madden and Lee Rainie. Americans' attitudes about privacy, security and surveillance. 2015.

Spyros Makridakis and Michele Hibon. The m3-competition: results, conclusions and implications. *International journal of forecasting*, 16(4):451–476, 2000.

Catherine Marsh, Chris Skinner, Sara Arber, Bruce Penhale, Stan Openshaw, John Hobcraft, Denise Lievesley, and Nigel Walford. The case for samples of anonymized records from the 1991 census. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154(2): 305–340, 1991.

David Martens. Fat flow: A data science ethics framework. 2020.

David Martens. *Data Science Ethics: Concepts, Techniques, and Cautionary Tales*. Oxford University Press, 2022.

David Martens and Foster Provost. Explaining data-driven document classifications. *MIS quarterly*, 38(1):73–100, 2014.

David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research*, 183(3):1466–1476, 2007.

David Martens, BB Baesens, and Tony Van Gestel. Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):178–191, 2008a.

David Martens, Johan Huysmans, Rudy Setiono, Jan Vanthienen, and Bart Baesens. Rule extraction from support vector machines: an overview of issues and application in credit scoring. *Rule extraction from support vector machines*, pages 33–63, 2008b.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. A survey on computational propaganda detection. *arXiv preprint arXiv:2007.08024*, 2020.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

Sandra Matz, Yin Wah Fiona Chan, and Michal Kosinski. Models of personality. *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*, pages 35–54, 2016.

Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48):12714–12719, 2017.

Sandra C Matz, Ruth E Appel, and Michal Kosinski. Privacy in the age of psychological targeting. *Current opinion in psychology*, 31:116–121, 2020.

Raphael Mazzine and Sofie Goethals. Counterfactual explanations for employment services. FEAST workshop at ECML: International workshop on Fair, Effective And Sustainable Talent management using data science, 2021.

Raphael Mazzine, Sofie Goethals, Dieter Brughmans, and David Martens. Counterfactual explanations for employment services. In *International workshop on Fair, Effective And*

Sustainable Talent management using data science, pages 1–7, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

Gregory J Meyer, Stephen E Finn, Lorraine D Eyde, Gary G Kay, Kevin L Moreland, Robert R Dies, Elena J Eisman, Tom W Kubiszyn, and Geoffrey M Reed. Psychological testing and psychological assessment: A review of evidence and issues. *American psychologist*, 56(2):128, 2001.

Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228, 2004.

Donald Michie, David J Spiegelhalter, and Charles C Taylor. *Machine learning, neural and statistical classification*. Citeseer, 1994.

Silvia Milano, Brent Mittelstadt, Sandra Wachter, and Christopher Russell. Epistemic fragmentation poses a threat to the governance of online targeting. *Nature Machine Intelligence*, 3(6):466–472, 2021.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Marius Miron, Songül Tolan, Emilia Gómez, and Carlos Castillo. Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Intelligence and Law*, 29(2):111–147, 2021.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. The unfairness of fair machine learning. *Available at SSRN 4331652*, 2023.

Ioannis Mollas, Nick Bassiliades, and Grigorios Tsoumakas. Truthful meta-explanations for local interpretability of machine learning models. *Applied Intelligence*, 53(22):26927–26948, 2023.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Evgeny Morozov. *To save everything, click here: The folly of technological solutionism*. PublicAffairs, 2013.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

Sandrine R Müller, Heinrich Peters, Sandra C Matz, Weichen Wang, and Gabriella M Harari. Investigating the relationships between mobility behaviours and indicators of subjective well–being using smartphone–based experience sampling and gps tracking. *European Journal of Personality*, 34(5):714–732, 2020.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

Francesca Naretto, Anna Monreale, and Fosca Giannotti. Evaluating the privacy exposure of interpretable global explainers. In *2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 13–19. IEEE, 2022.

Michael Neely, Stefan F Schouten, Maurits JR Bleeker, and Ana Lucic. Order in the court: Explainable ai methods prone to disagreement. *arXiv preprint arXiv:2105.03287*, 2021.

Peter Bjorn Nemenyi. *Distribution-free multiple comparisons.* Princeton University, 1963.

Helen Nissenbaum. Accountability in a computerized society. *Science and engineering ethics*, 2: 25–42, 1996.

Natalia Norori, Qiyang Hu, Florence Marcelle Aellen, Francesca Dalia Faraci, and Athina Tzovara. Addressing bias in big data and ai for health care: A call for open science. *Patterns*, 2(10), 2021.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):1–13, 2017.

Derek O'callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13): 5645–5657, 2015.

Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1895–1904, 2022.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. On counterfactual explanations under predictive multiplicity. In *Conference on Uncertainty in Artificial Intelligence*, pages 809–818. PMLR, 2020.

Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. On the privacy risks of algorithmic recourse. In *International Conference on Artificial Intelligence and Statistics*, pages 9680–9696. PMLR, 2023.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19, 2000.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 560–568, 2008.

Claudia Perlich, Foster Provost, and Jeffrey Simonoff. Tree induction vs. logistic regression: A learning-curve analysis. 2003.

Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95 (1):103–127, 2014.

Emmanuel Pintelas, Ioannis E Livieris, and Panagiotis Pintelas. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. *Algorithms*, 13(1):17, 2020.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Sergei Popov, Stanislav Morozov, and Artem Babenko. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312*, 2019.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Md Ileas Pramanik, Raymond YK Lau, and Wenping Zhang. K-anonymity through the enhanced clustering method. In *2016 IEEE 13th International Conference on e-Business Engineering (ICEBE)*, pages 85–91. IEEE, 2016.

Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105:1257, 2019.

Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.", 2013.

Foster J Provost, Tom Fawcett, Ron Kohavi, et al. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

Sujin Pyo, Jaewook Lee, Mincheol Cha, and Huisu Jang. Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the korean stock markets. *PloS one*, 12(11):e0188107, 2017.

Pengrui Quan, Supriyo Chakraborty, Jeya Vikranth Jeyakumar, and Mani Srivastava. On the amplification of security and privacy risks by post-hoc explanations in machine learning models. *arXiv preprint arXiv:2206.14004*, 2022.

J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

Yanou Ramon. *Rule-based explanation methods to gain insight into classification models using behavioral data*. PhD thesis, University of Antwerp, 2022.

Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Advances in Data Analysis and Classification*, 14(4):801–819, 2020.

Yanou Ramon, RA Farrokhnia, Sandra C Matz, and David Martens. Explainable ai for psychological profiling from behavioral data: an application to big five personality predictions from financial transaction records. *Information*, 12(12):518, 2021a.

Yanou Ramon, David Martens, Theodoros Evgeniou, and Stiene Praet. Can metafeatures help improve explanations of prediction models when using behavioral and textual data? *Machine Learning*, pages 1–40, 2021b.

Charan Reddy. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. 2022.

Lauren Rhue, Sofie Goethals, and Arun Sundararajan. Evaluating LLMs for gender disparities in notable persons. *arXiv preprint arXiv:2403.09148*, 2024.

Filipe N Ribeiro, Fabrício Benevenuto, and Emilio Zagheni. How biased is the population of Facebook users? Comparing the demographics of Facebook users with census data to generate correction factors. In *Proceedings of the 12th ACM Conference on Web Science*, pages 325–334, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*, 2017.

Adriano Rivolli, Luís PF Garcia, Carlos Soares, Joaquin Vanschoren, and André CPLF de Carvalho. Characterizing classification datasets: a study of meta-features for meta-learning. *arXiv preprint arXiv:1808.10406*, 2018.

Saumendu Roy, Gabriel Laberge, Banani Roy, Foutse Khomh, Amin Nikanjam, and Saikat Mondal. Why don't xai techniques agree? characterizing the disagreements between post-hoc explanations of defect predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 444–448. IEEE, 2022.

Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2), 2019.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1):1, 2020.

Stefan Rüping et al. Learning interpretable models. *Universität Dortmund*, 2006.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Robert J Sampson, John H Laub, and Christopher Wimer. Does marriage reduce crime? a counterfactual approach to within-individual causal effects. *Criminology*, 44(3):465–508, 2006.

Teresa Scantamburlo, Joachim Baumann, and Christoph Heitz. On prediction-modelers and decision-makers: why fairness requires more than a fair prediction model. *AI & SOCIETY*, pages 1–17, 2024.

Carel Schwartzenberg, Tom van Engers, and Yuan Li. The fidelity of global surrogates in interpretable machine learning. *BNAIC/BeneLearn 2020*, page 269, 2020.

Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aïvodji, Aisha Alaagib, Sébastien Gambs, and Nicolas Papernot. Washing the unwashable: On the (im) possibility of fairwashing detection. *Advances in Neural Information Processing Systems*, 35:14170–14182, 2022.

Lloyd S Shapley et al. A value for n-person games. 1953.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172, 2020.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Reza Shokri, Martin Strobel, and Yair Zick. Privacy risks of explaining machine learning models. *CoRR*, abs/1907.00164, 2019. URL http://arxiv.org/abs/1907.00164.

Reza Shokri, Martin Strobel, and Yair Zick. Exploiting transparency measures for membership inference: a cautionary tale. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI). AAAI*, volume 13, 2020.

Silberling, Amanda. Facebook will no longer allow advertisers to target political beliefs, religion, sexual orientation. https://techcrunch.com/2021/11/09/facebook-will-no-longer-allow-advertisers-to-target-political-beliefs-religion-sexual-orientation/, 2021. Accessed: 2023-03-13.

MS Simi, K Sankara Nayaki, and M Sudheep Elayidom. An extensive study on data anonymization algorithms based on k-anonymity. In *IOP Conference Series: Materials Science and Engineering*, volume 225, page 012279. IOP Publishing, 2017.

Amanpreet Singh, Narina Thakur, and Aakanksha Sharma. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1310–1315. Ieee, 2016.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. *Advances in neural information processing systems*, 34:62–75, 2021.

Djordje Slijepčević, Maximilian Henzl, Lukas Daniel Klausner, Tobias Dam, Peter Kieseberg, and Matthias Zeppelzauer. k-anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security*, 111:102488, 2021.

Kacper Sokol and Peter Flach. Explainability is in the mind of the beholder: Establishing the foundations of explainable artificial intelligence. *arXiv preprint arXiv:2112.14466*, 2021.

Kacper Sokol and Peter A Flach. Counterfactual explanations of machine learning predictions: opportunities and challenges for ai safety. In *SafeAI@ AAAI*, 2019.

Kacper Sokol, Raul Santos-Rodriguez, and Peter Flach. FAT Forensics: A Python toolbox for algorithmic fairness, accountability and transparency. *arXiv preprint arXiv:1909.05167*, 2019.

Daniel J Solove. Introduction: Privacy self-management and the consent dilemma. *Harv. L. Rev.*, 126:1880, 2012.

Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2239–2248, 2018.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

M Sumana and KS Hareesha. Anonymity: an assessment and perspective in privacy preserving data mining. *International Journal of Computer Applications*, 6(10), 2010.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.

Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05): 571–588, 2002a.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002b.

David G Taylor, Donna F Davis, and Ravi Jillapalli. Privacy concern and online personalization: The moderating effects of information control and compensation. *Electronic commerce research*, 9:203–223, 2009.

Trang P Tran. Personalized ads on facebook: An effective marketing tool for online marketers. *Journal of Retailing and Consumer Services*, 39:230–242, 2017.

Bogdan Trawiński, Magdalena Smketek, Zbigniew Telec, and Tadeusz Lasota. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *International Journal of Applied Mathematics and Computer Science*, 22:867–881, 2012.

Paul Van der Laan. The 2001 census in the netherlands. In *Conference the Census of Population*, 2000.

José Van Dijck and Thomas Poell. Understanding social media logic. *Media and communication*, 1(1):2–14, 2013.

Iris Van Ooijen and Helena U Vrabec. Does the GDPR enhance consumers' control over personal data? An analysis from a behavioural perspective. *Journal of Consumer Policy*, 42: 91–107, 2019.

Henk CA Van Tilborg and Sushil Jajodia. *Encyclopedia of cryptography and security*. Springer Science & Business Media, 2014.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (fairware)*, pages 1–7. IEEE, 2018.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: Challenges revisited. *arXiv preprint arXiv:2106.07756*, 2021.

Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira, and David Martens. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, pages 1–21, 2022a.

Tom Vermeire, Thibault Laugel, Xavier Renard, David Martens, and Marcin Detyniecki. How to choose an explainability method? towards a methodical implementation of xai in practice. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I*, pages 521–533. Springer, 2022b.

Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller, and Bernhard Schölkopf. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9584–9594, 2022.

Sandra Wachter. The theory of artificial immutability: Protecting algorithmic groups under anti-discrimination law. *Tulane Law Review*, 97:149, 2022.

Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017a.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017b.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. *W. Va. L. Rev.*, 123:735, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and AI. *Computer Law & Security Review*, 41:105567, 2021.

Yu-Xiong Wang and Yu-Jin Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on knowledge and data engineering*, 25(6):1336–1353, 2012.

Alan F Westin. Social and political dimensions of privacy. *Journal of Social Issues*, 59(2):431–453, 2003.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

David M White. The requirement of race-conscious evaluation of LSAT scores for equitable law school admissions. *Berkeley La Raza LJ*, 12:399, 2000.

Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Linda F Wightman. LSAC National Longitudinal Bar Passage Study. LSAC research report series. 1998.

Steve Wozniak. Tweet. https://twitter.com/stevewoz/status/1193330241478901760, November 10, 2019. Twitter, Online, accessed February 1, 2022.

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, pages 563–574. Springer, 2019.

Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 785–790, 2006a.

Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization for privacy preservation with less information loss. *ACM SIGKDD Explorations Newsletter*, 8(2):21–30, 2006b.

Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. Algorithmic decision making with conditional fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2125–2135, 2020.

Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6, 2017.

Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. *arXiv preprint arXiv:1906.01747*, 2019.

Samuel Yeom and Michael Carl Tschantz. Avoiding disparity amplification under different worldviews. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 273–283, 2021.

Xun Yi, Russell Paulet, Elisa Bertino, Xun Yi, Russell Paulet, and Elisa Bertino. *Homomorphic encryption*. Springer, 2014.

Razieh Nokhbeh Zaeem and K Suzanne Barber. The effect of the gdpr on privacy policies: Recent progress and future promise. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):1–20, 2020.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 180(3): 689–722, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019.

Zhi-Hua Zhou. Rule extraction: Using neural networks or for neural networks? *Journal of Computer Science and Technology*, 19(2):249–253, 2004.

Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.

Part VI

SUPPLEMENTARY MATERIAL

11 Appendices

A The cost of comprehensibility

A.1 *Materials*

A.1.1 *Datasets*

The used datasets can be found in Table A.1.

Table A.1: Used datasets

Dataset	# observations	# features	Imbalance
adult	48842	14	0.27
agaricus_lepiota	8145	22	0
analcatdata_aids	50	4	0
analcatdata_asbestos	83	3	0.01
analcatdata_bankruptcy	50	6	0
analcatdata_boxing1	120	3	0.09
analcatdata_boxing2	132	3	0.01
analcatdata_creditscore	100	6	0.21
analcatdata_cyyoung8092	97	10	0.26
analcatdata_cyyoung9302	92	10	0.34
analcatdata_fraud	42	11	0.15
analcatdata_japansolvent	52	9	0
analcatdata_lawsuit	264	4	0.73
appendicitis	106	7	0.36
australian	690	14	0.01
backache	180	32	0.52
biomed	209	8	0.08
breast_cancer_wisconsin	569	30	0.06

Continued on next page

Table A.1 continued from previous page

Dataset	# observations	# features	Imbalance
breast_cancer	286	9	0.16
breast_w	699	9	0.1
breast	699	10	0.1
buggyCrx	690	15	0.01
bupa	345	5	0
chess	3196	36	0
churn	5000	20	0.51
clean1	476	168	0.02
clean2	6598	168	0.48
cleve	303	13	0.01
coil2000	9822	85	0.78
colic	368	22	0.07
corral	160	6	0.02
credit_a	690	15	0.01
credit_g	1000	20	0.16
crx	690	15	0.01
diabetes	768	8	0.09
dis	3772	29	0.94
flare	1066	10	0.43
GAMETES_Epistasis_2_Way_1000atts_0.4H_EDM_1_EDM_1_1	1600	1000	0
GAMETES_Epistasis_2_Way_20atts_0.1H_EDM_1_1	1600	20	0
GAMETES_Epistasis_2_Way_20atts_0.4H_EDM_1_1	1600	20	0
GAMETES_Epistasis_3_Way_20atts_0.2H_EDM_1_1	1600	20	0
GAMETES_Heterogeneity_20atts_1600_Het_0.4_0.2_50_EDM_2_001	1600	20	0
GAMETES_Heterogeneity_20atts_1600_Het_0.4_0.2_75_EDM_2_001	1600	20	0
german	1000	20	0.16
glass2	163	9	0
haberman	306	3	0.22
heart_c	303	13	0.01
heart_h	294	13	0.08
heart_statlog	270	13	0.01
hepatitis	155	19	0.34
Hill_Valley_with_noise	1212	100	0
Hill_Valley_without_noise	1212	100	0
horse_colic	368	22	0.07
house_votes_84	435	16	0.05
hungarian	294	13	0.08
hypothyroid	3163	25	0.82
ionosphere	351	34	0.08
irish	500	5	0.01
kr_vs_kp	3196	36	0

Continued on next page

Table A.1 continued from previous page

Dataset	# observations	# features	Imbalance
labor	57	16	0.09
lupus	87	3	0.04
magic	19020	10	0.09
mofn_3_7_10	1324	10	0.31
molecular_biology_promoters	106	57	0
monk1	556	6	0
monk2	601	6	0.1
monk3	554	6	0
mushroom	8124	22	0
mux6	128	6	0
parity5	32	5	0
parity5+5	1124	10	0
phoneme	5404	5	0.17
pima	768	8	0.09
postoperative_patient_data	88	8	0.21
prnn_crabs	200	7	0
prnn_synth	250	2	0
profb	672	9	0.11
ring	7400	20	0
saheart	462	9	0.09
sonar	208	60	0
spambase	4601	57	0.04
spect	267	22	0.35
spectf	349	44	0.21
threeOf9	512	9	0
tic_tac_toe	958	9	0.09
tokyo1	959	44	0.08
twonorm	7400	20	0
vote	435	16	0.05
wdbc	569	30	0.06
xd6	973	9	0.11

A.1.2 *Dataset properties*

For the analysis of the dataset properties, we use the metafeature toolbox of Alcobaba [Alcobaça et al., 2020], that automatically extracts metafeatures out of the dataset. The metafeatures of this toolbox are based on those described in [Rivolli et al., 2018]. We select the metafeatures out of the groups: general, statistical, info-theory and complexity. The general metafeatures represent the basic information about the dataset. They capture metrics such as the number of instances, attributes, or other information about the predictive attribute [Rivolli et al., 2018]. The statistical measures represent information about the data distribution like the number of outliers, the variance, the skewness or the correlation in the data , and others [Rivolli et al., 2018]. The information-theoretic measures capture the amount of information present in the data such as the joint entropy, class entropy, class concentration, and others

[Rivolli et al., 2018]. The last group of measures we include is the group of information-complexity based on [Lorena et al., 2019]. We do not include the clustering, landmarking or model-based metafeatures because they already fit a model to the dataset and extract information from this model. The used dataset properties can be seen in Table A.2.[1]

Metafeature name	*Description*
AttrConc (mean)	*Concentration coef. of each pair of distinct attributes.*
AttrEnt (mean)	*Shannon's entropy for each predictive attribute.*
AttrToInst	*The ratio between the number of attributes.*
C1	*The entropy of class proportions.*
C2	*The imbalance ratio.*
CanCor (mean)	*Canonical correlations of data.*
CatToNum	*The ratio between the number of categoric and numeric features.*
ClassConc (mean)	*Concentration coefficient between each attribute and class.*
ClassEnt	*Target attribute Shannon's entropy.*
ClsCoef	*Clustering coefficient.*
Cor (mean)	*The absolute value of the correlation of distinct dataset column pairs.*
Cov (mean)	*The absolute value of the covariance of distinct dataset attribute pairs.*
Density	*Average density of the network.*
Eigenvalues (mean)	*Eigenvalues of covariance matrix from dataset.*
EqNumAttr	*Number of attributes equivalent for a predictive task.*
F1 (mean)	*Maximum Fisher's discriminant ratio.*
F1v (mean)	*Directional-vector maximum Fisher's discriminant ratio.*
F2 (mean)	*Volume of the overlapping region.*
F3 (mean)	*Feature maximum individual efficiency.*
F4 (mean)	*Collective feature efficiency.*
FreqClass (mean)	*Relative frequency of each distinct class.*
Gmean (mean)	*Geometric mean of each attribute.*
Gravity	*Distance between minority and majority classes center of mass.*
Hmean (mean)	*Harmonic mean of each attribute.*
Hubs (mean)	*Hub score*
InstToAttr	*Ratio between the number of instances and attributes.*
IqRange (mean)	*Interquartile range (IQR) of each attribute.*
JointEnt (mean)	*Joint entropy between each attribute and class.*
Kurtosis (mean)	*Kurtosis of each attribute.*
L1 (mean)	*Sum of error distance by linear programming.*
L2 (mean)	*OVO subsets error rate of linear classifier.*
L3 (mean)	*Non-Linearity of a linear classifier.*
LhTrace	*Lawley-Hotelling trace.*
Lsc	*Local set average cardinality.*
Mad (mean)	*Median Absolute Deviation (MAD) adjusted by a factor.*
Max (mean)	*Maximum value from each attribute.*
Mean (mean)	*Mean value of each attribute.*
Median (mean)	*Median value from each attribute.*
Min (mean)	*Minimum value from each attribute.*
MutInf (mean)	*Mutual information between each attribute and target.*
N1	*Fraction of borderline points.*
N2 (mean)	*Ratio of intra and extra class nearest neighbor distance.*
N3 (mean)	*Error rate of the nearest neighbor classifier.*
N4 (mean)	*Non-linearity of the k-NN Classifier.*
NrAttr	*Total number of attributes.*

[1] Based on the list: https://pymfe.readthedocs.io/en/latest/auto_pages/meta_features_description.html

NrBin	*Number of binary attributes.*
NrCat	*Number of categorical attributes.*
NrClass	*Number of distinct classes.*
NrCorAttr	*Number of distinct highly correlated pair of attributes.*
NrDisc	*Number of canonical correlation between each attribute and class.*
NrInst	*Number of instances (rows) in the dataset.*
NrNorm	*Number of attributes normally distributed based in a given method.*
NrNum	*Number of numeric features.*
NrOutliers	*Number of attributes with at least one outlier value.*
NsRatio	*Noisiness of attributes.*
NumToCat	*Number of numerical and categorical features.*
Ptrace	*Pillai's trace.*
Range (mean)	*Range (max - min) of each attribute.*
RoyRoot	*Roy's largest root.*
Sd (mean)	*Standard deviation of each attribute.*
SdRatio	*Statistical test for homogeneity of covariances.*
Skewness (mean)	*Skewness for each attribute.*
Sparsity (mean)	*(Possibly normalized) sparsity metric for each attribute.*
T1 (mean)	*Fraction of hyperspheres covering data.*
T2	*Average number of features per dimension.*
T3	*Average number of PCA dimensions per points.*
T4	*Ratio of the PCA dimension to the original dimension.*
TMean (mean)	*Trimmed mean of each attribute.*
Var (mean)	*Variance of each attribute.*
WLambda	*Wilks' Lambda value.*

Table A.2: Dataset properties used in the analysis

A.2 *Extra results on accuracy*

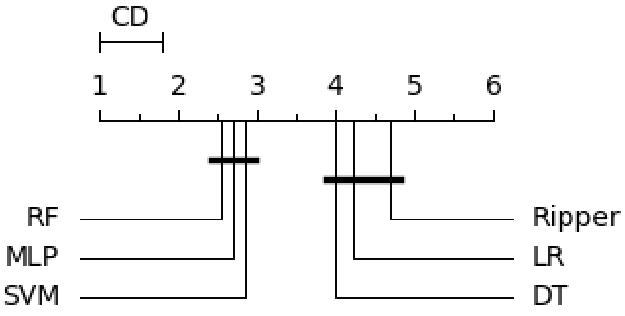
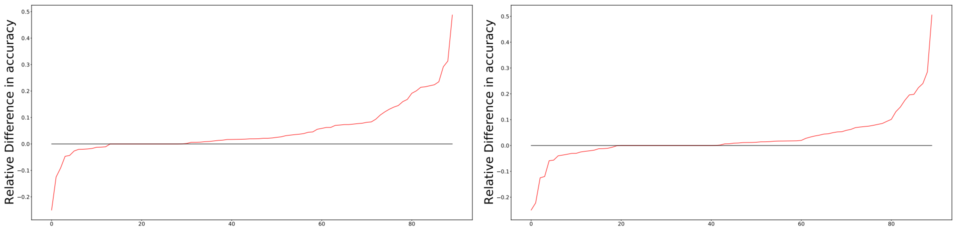


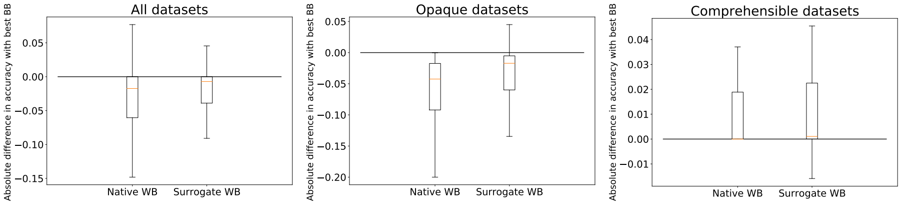
Figure A.1: Critical difference diagram of the comparison of classifiers for accuracy. Models that are not connected by the bold line have a significant difference in performance (at a 5% level with the Nemenyi test).



(a) Non-linearity of the cost of comprehensibility

(b) Non-linearity of the cost of explainability

Figure A.2: Comparing black box and white box models. For both plots, the datasets are ordered according to the gap in accuracy between the best black box and the best native (left figure) or surrogate (right) white box model (right).The y-axis measures the relative difference in the accuracy, defined as the ratio of the difference between the black and white box accuracy divided by that of the best model.



(a) Comparison of performance of the native and surrogate white box models across all the datasets.

(b) Comparison of performance of the native and surrogate white box models across all opaque datasets.

(c) Comparison of performance of the native and surrogate white box models across all comprehensible datasets.

Figure A.3: Comparison across datasets of best black box model for each dataset, surrogate white box model mimicking this best black box, and best native white box model. BB stands for black box and WB for white box. The line at 0 indicates the performance of the best black box model. The y-axis indicates the absolute difference in accuracy from the best black box model.

We report the empirical results as in the main article, this time using the accuracy of the models as our metric instead of the f1-score. All results are in line with the results for the f1-score. The hypothesis of the Friedman test is rejected with a value of $2.09 \cdot e^{-23}$. In Figure A.1, we show that the black box models are significantly better than the white box models but not significantly different from each other. The same can be said for the white box models. We see a non-linear nature of the cost of comprehensibility and explainability in Figures A.2a and A.2b. Finally, from the boxplots in Figure A.3 we see again that for the opaque datasets the surrogate white box models are better on average than the native ones. We also reject the hypothesis that the native and surrogate white boxes perform equally well (p-value $9.63 \cdot e^{-6}$) on average across all datasets. When we perform the same analysis for the two different types of datasets, we see again that the surrogate white box models outperform the native white box ones for the opaque datasets (Wilcoxon test p-value of $2.71 \cdot e^{-6}$), while the two are not significantly different for the comprehensible

Table A.3: The dataset properties that are significant when explaining the cost of comprehensibility using a number of standard dataset properties as independent variables in a regression model where the cost is the dependent variable.

Variable	MSE	Pr(>F)	Coef
F1v	0.089	0.0002	0.116
L3	0.059	0.003	0.096
T4	0.045	0.012	0.063
L2	0.044	0.012	0.089
L1	0.042	0.014	0.082
N4	0.038	0.020	0.094
CanCor	0.032	0.034	-0.075
F3	0.031	0.036	0.087
F3	0.031	0.036	0.087
EqNumAttr	0.029	0.044	-0.171
NsRatio	0.029	0.044	-0.171

datasets (Wilcoxon test p-value of 0.53). All these results are comparable with the results obtained when using f1-score as a metric.

Finally, we also compare the dataset properties that predict whether a dataset is *opaque* or *comprehensible* and see if they are the same for both metrics. We see in Table A.3 that the same dataset properties are important in predicting the gap in *accuracy* as in predicting the gap in *f1-score*, but that now some more attributes are significant. *F1v*, *L1*, *EqNumAttr* and *NsRatio* were already significant in predicting the gap in *f1-score*. The linearity measures *L2* and *L3* are now also significant but they have a similar meaning as *L1*, namely they are linearity measures that quantify whether the data is linearly separable, which means higher values of these attributes point to more complex problems [Lorena et al., 2019]. *N4* signifies the non-linearity of the nearest neighbor classifier and higher values are also indicative of problems of greater complexity [Lorena et al., 2019]. *F3* signifies the Maximum Individual Feature Efficiency where lower values indicate simpler problems [Lorena et al., 2019]. *JointEnt* computes the relationship of each attribute with the target variable, capturing the relative importance of the predictive attributes [Rivolli et al., 2018]. *CanCor* measures the canonical correlation between the predictive attribute and the target [Rivolli et al., 2018].

B PREDICTIVE COUNTERFACTUAL FAIRNESS

B.1 *PreSHAPF*

Alternative XAI techniques can also be employed to investigate the presence of implicit bias in a machine learning model. In this section, we use SHAP values as a means of examining disparities between two sensitive groups. The results can differ as SHAP values focuses on variations in prediction scores, rather than on decisions (which basically is the combination of a prediction score and threshold). SHAP values are a computationally efficient way to calculate Shapley values, which are defined as the average marginal contribution across all possible coalitions [Lundberg and Lee, 2017].

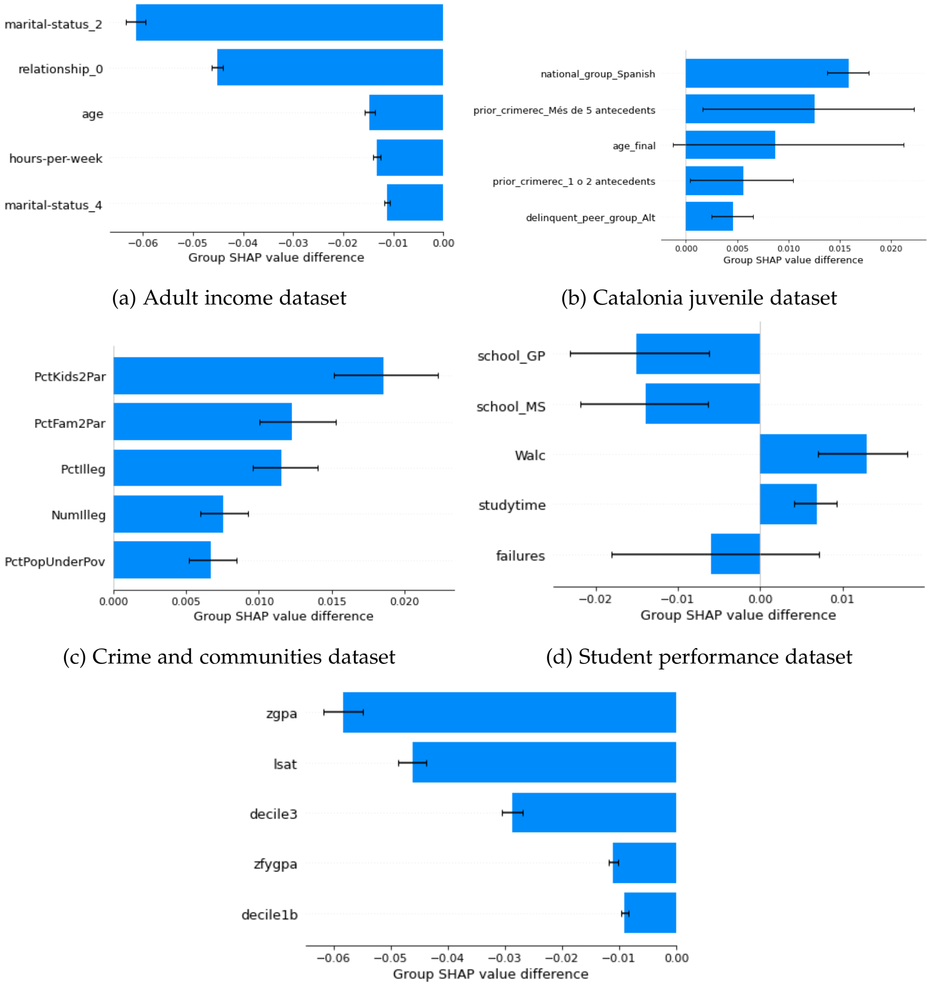
SHAP attributes to each feature the change in the expected model prediction when conditioning on that feature and thus reveals the extent to which each feature contributes to the prediction score, either positively or negatively [Lundberg and Lee, 2017]. As with *PreCoF*, we focus on the negatively affected members of both the protected and the unprotected group. We compare the mean SHAP values in both subgroups and *PreSHAPF* (*Predictive SHAP Fairness*) will reveal for which features the difference between both subgroups is the largest.[2] There are two main differences between *PreCoF* and *PreSHAPF*: First, *PreCoF* focuses on the decisions made by the model, while *PreSHAPF* focuses on the prediction scores. Second, *PreCoF* returns features (and *PreCoF*$_c$ the feature values for the categorical features), while *PreSHAPF* will always return feature values for the categorical features.

B.1.1 *PreCoF vs PreSHAPF*

The results for the datasets used in this paper can be found in Figure B.4. For each dataset, we calculate *PreSHAPF* as the discrepancy in mean SHAP values between both subgroups. As demonstrated in Figure B.4, the most salient patterns detected with *PreCoF* are also present in *PreSHAPF*, however, slight variations are observed as they measure distinct phenomena.

As depicted in Figure B.4a, the two features with the largest difference in *PreSHAPF*, namely *relationship: 0* and *marital-status: 2*, correspond to the feature values with the highest value in *PreCoF*$_c$, as can be seen in Figure 6.1b. These features, on average, negatively impact the prediction score of women (to be predicted to have a high income) compared to men. The other top features are different between *PreCoF* and *PreSHAPF*. In Figure B.4b, we observe that the feature *national group:*

[2] We use the SHAP package TreeExplainer to calculate the SHAP values (as we are explaining a random forest) [Lundberg et al., 2020]. We use the group difference plots provided by SHAP to graph the difference in mean SHAP values between the two subgroups [Lundberg et al., 2020].



(a) Adult income dataset

(b) Catalonia juvenile dataset

(c) Crime and communities dataset

(d) Student performance dataset

(e) Law admission dataset

Figure B.4: *PreSHAPF*

Spanish, which was the $PreCoF_c$ attribute in Figure 6.3b, is the feature value with the highest value in *PreSHAPF*. For foreigners, this feature, on average, has a larger positive impact on the prediction score (to be predicted to recidive) compared to locals. However, in Figure 6.3b, we see that the other values for *national group*, namely *Altres* and *Europa*, are also high ranked in *PreCoF*, but they are not among the top features in *PreSHAPF*. As illustrated in Figure B.4c, the features with the highest *PreSHAPF* value are the same as the PreCoF attributes (*PctIlleg*, *PctKids2Par*, *PctFam2Par*, *NumIlleg*) in Figure 6.5a in a slightly different order. Figure B.4d shows that both values of *School* have the highest value in *PreSHAPF*. These attributes (on average) negatively impact the prediction score (to be predicted a good student) of girls compared to boys This is in line with the results of *PreCoF*, as *School* was the *PreCoF* attribute in Figure 6.6a, but the other attributes differ. In Figure B.4e, we observe that all features, on average, have a negative impact on the prediction score (to be predicted to pass the bar) of Non-Whites compared to Whites. The two features for which this discrepancy is the largest are *zgpa* and *lsat*, which were also the two attributes with the largest value in *PreCoF* in Figure 6.8a.

Overall, we notice that the global patterns seem consistent over *PreCoF* and *PreSHAPF*. However, the less important features can vary strongly, which shows that *PreCoF* and *PreSHAPF* function differently. When we change the threshold of the machine learning classifier trained on these datasets, the results of *PreCoF* will strongly change (for some thresholds, all of the top features are different), while this will have no effect on the results of *PreSHAPF*. We add a supplementary illustration which shows the effect of the decision threshold on *PreCoF* and *PreSHAPF* in Section B.1.2.

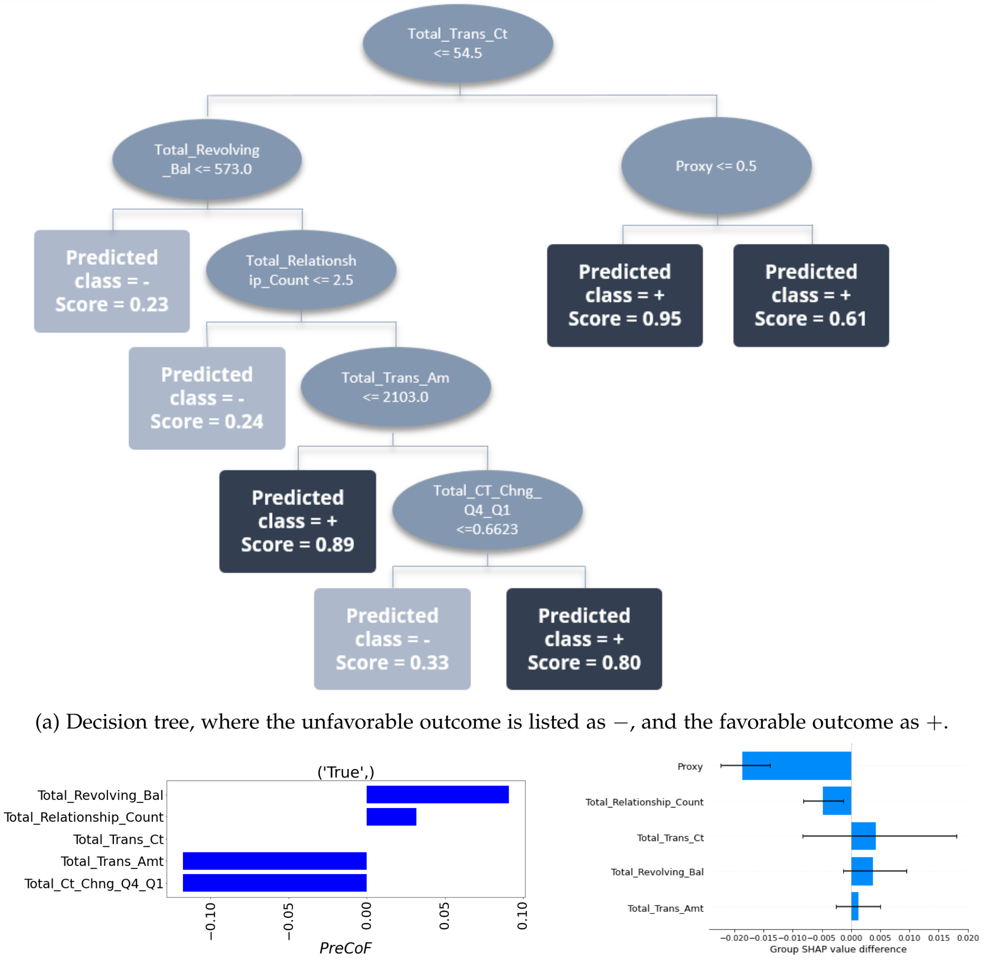
B.1.2 *PreCoF versus PreSHAPF on a transparent model*

To further demonstrate the functionality of *PreCoF* and its distinction with SHAP values, we present an additional illustration using an existing churn dataset set.[3] This dataset aims to predict whether a bank customer will churn or not, where the unfavorable outcome is that the customer will attrite, and the favorable outcome that the customer will remain loyal. The dataset does not contain a sensitive attribute, but we artificially introduce this aspect to the dataset, randomly assigning half of the instances the gender of male and half of the instances the gender of female. In contrary to our previous experiments, we use an interpretable decision tree (with a restricted number of 7 leaf nodes) to provide insight into the model's functioning and to facilitate a comparison of how counterfactual explanations and SHAP values detect bias within the model.

To investigate the implicit bias, we add a proxy that is correlated with the target outcome and gender. This action is likely to result in the model picking up this biased pattern, even after we remove the sensitive attribute (gender) and may result in gender discrimination in the model's predictions. As in our previous experiments to detect implicit bias, we remove the sensitive attribute (*gender*) from the data, split the data into a training and test set, and fit a machine learning on the training set. However, in this scenario, we use a simple decision tree, as opposed to a Random Forest model, to compare the results from *PreCoF* and *PreSHAPF* with the actual model, as depicted in Figure B.5a.

These results clearly illustrate how counterfactual explanations and SHAP values function differently. The *proxy* has a large impact on the prediction score, but will

[3] https://www.kaggle.com/datasets/syviaw/bankchurners



(a) Decision tree, where the unfavorable outcome is listed as −, and the favorable outcome as +.

(b) *PreCoF*

(c) *PreSHAPF*

Figure B.5: Additional illustration with a transparent machine learning model to show the difference between *PreCoF* and *PreSHAPF*

not have an effect on the decision for any of the instances (both leaf nodes after the biased feature split result in the same outcome as the threshold is 0.5). When using *PreSHAPF*, we see in Figure B.5c that the feature with the largest value is *proxy*. This makes sense, as we see in the decision tree, that it has a large effect on the prediction score and we know that it is correlated to gender. On the other hand, in Figure B.5b, *PreCoF* does not report this feature as it does not change the decision for any of the instances. If the threshold of the machine learning classifier changes to 0.7 or 0.8, *PreCoF* does report *proxy* as the top feature.

These results indicate that both SHAP values and counterfactual explanations are well-suited to identify patterns of indirect discrimination, but that they measure distinct phenomena. Their outcomes may vary as counterfactual explanations focus on decisions and SHAP values on prediction scores. In this paper, we use counterfactual explanations as our focus is on the actual decisions people receive, but using SHAP values is a good alternative when the focus is on fair scoring (for example with a varying or unfixed threshold). Our main argument that a deeper understanding of the nature of the bias is necessary before deciding on a method to address it, remains valid when using both XAI techniques. Finally, our experiments further confirm that both *PreCoF* and *PreSHAPF* are detecting bias in the model, and not in the underlying data. If a biased feature is added to the dataset but not picked up by the model, neither *PreCoF* or *PreSHAPF* will show this biased feature.

C Beyond accuracy-fairness

C.1 *Score distributions for the other data*

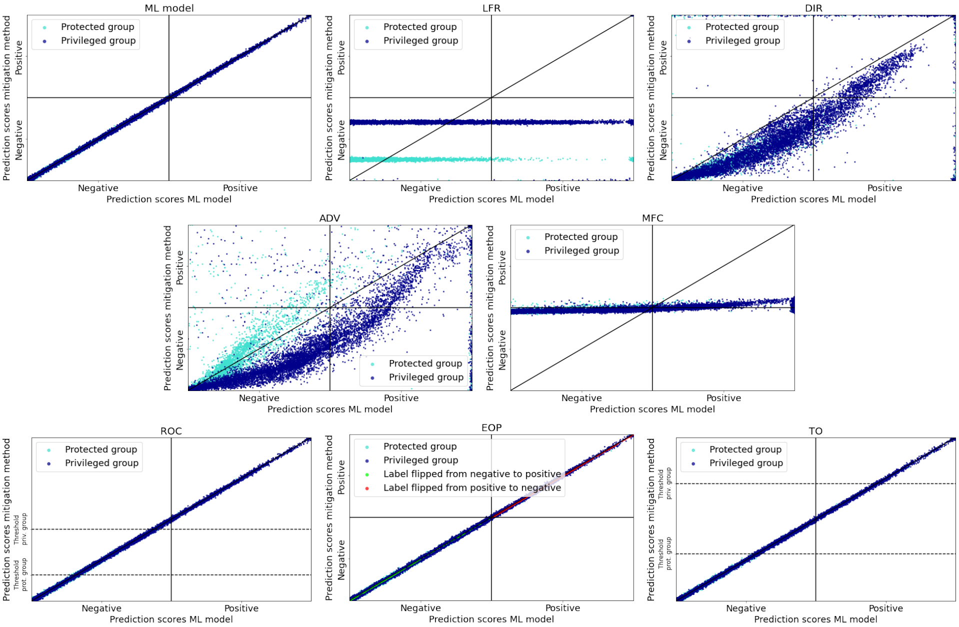


Figure C.6: Score distributions for the Adult dataset

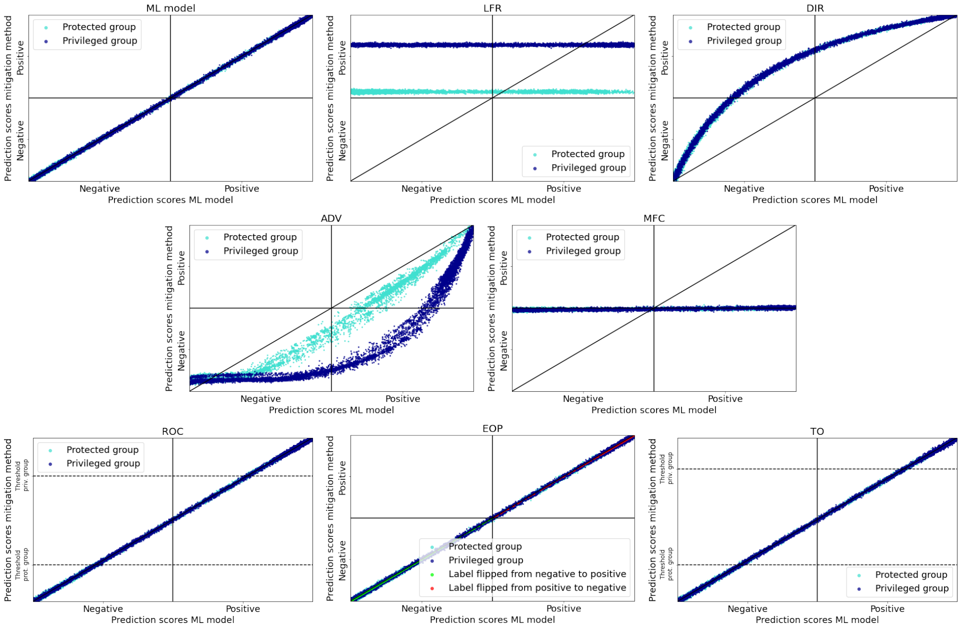


Figure C.7: Score distributions for the Dutch dataset

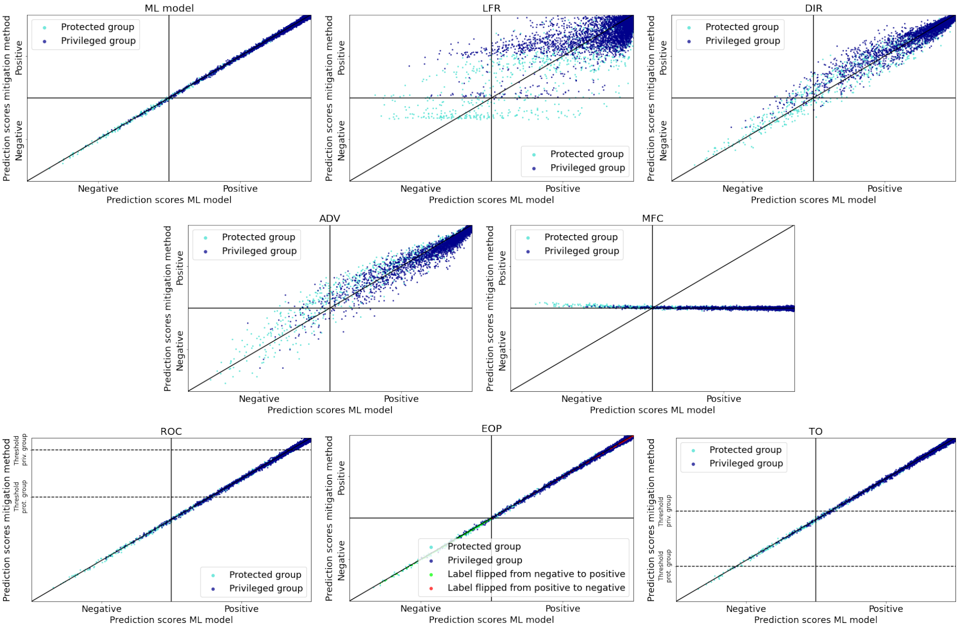


Figure C.8: Score distributions for the Law dataset

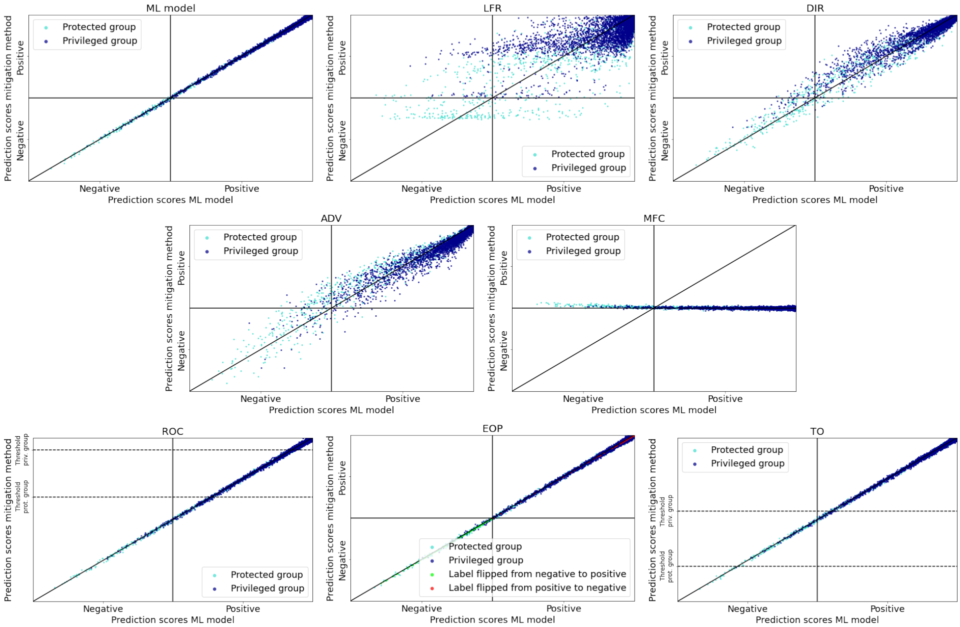


Figure C.9: Score distributions for the Student dataset

D THE PRIVACY ISSUE OF COUNTERFACTUAL EXPLANATIONS

D.1 *Results with different values for k*

Table D.4: Results of CF-K (k = 5).

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.18%	2.07%	14.53%	0.93%	1.97%	7.34%
Pureness (mean)	99.69%	96.24%	99.85%	100%	95.31%	89.16%
Execution time (mean)	16.82s	12.67s	6.25s	1.92s	14.93s	25.04s
C_{DM}	83,990	3,584	576	450	12,809	4,755
$\frac{C_{DM}}{\#explanations}$	106.72	8.83	9.6	8.04	17.17	7.19
CM	0.81	0.24	0.07	0.25	0.71	0.11

Table D.5: Results of Mondrian (k = 5)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	9.28%	4.95%	42.02%	24.44%	14.72%	29.40%
Pureness (mean)	92.30%	79.85%	93.82%	100%	71.05%	74.39%
Execution time (mean)	16.65s	1.56s	0.56s	0.36s	2.48s	2.64s
C_{DM} **(mean)**	116,132	4,348	402	486	12,524	4,607
$\frac{C_{DM}}{\#explanations}$	147.56	10.71	6.07	8.68	16.79	6.97
CM (mean)	0.82	0.24	0.22	0.32	0.73	0.31

Table D.6: Results of CF-K (k = 20).

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	0.84%	7.46%	27.38%	5.02%	5.35%	12.86%
Pureness (mean)	99.61%	88.85%	99.08%	100%	88.85%	82.46%
Execution time (mean)	32.69s	19.29s	25.44s	4.15s	27.61s	86.101s
C_{DM}	93,597	9,509	1,541	1,329	24,442	15,976
$\frac{C_{DM}}{\#explanations}$	118.93	23.42	25.68	23.73	32.76	24.17
CM	0.84	0.25	0.02	0.39	0.85	0.14

Table D.7: Results of Mondrian (k = 20)

Dataset	Adult	CMC	German	Heart	Hospital	Informs
NCP (mean)	23.32%	12.35%	71.39%	57.44%	38.63%	42.62%
Pureness (mean)	85.01%	59.40%	89.52%	100%	58.08%	72.88%
Execution time (mean)	16.39s	2.21s	0.72s	0.42s	2.77s	3.06s
C_{DM} **(mean)**	126,238	14,557	1,832	2,003	25,566	20,242
$\frac{C_{DM}}{\#explanations}$	160.40	35,85	30.53	35.77	34.27	30.62
CM (mean)	0.82	0.20	0.13	0.43	0.87	0.33

E THE IMPACT OF CLOAKING DIGITAL FOOTPRINTS ON USER PRIVACY AND PERSONALIZATION

E.1 *Results of other cloaking strategies*

E.1.1 *Using domain-based metafeatures*

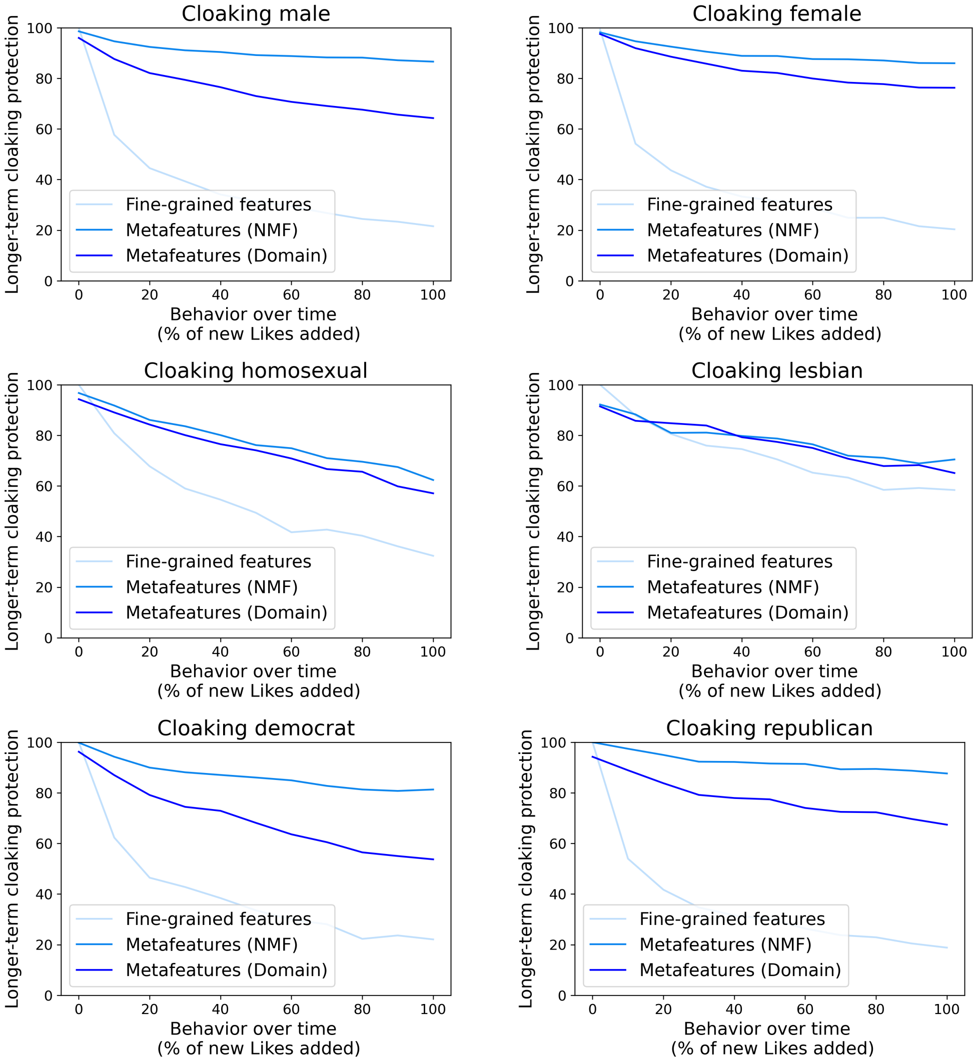


Figure E.10: Longer-term cloaking protection over time when using domain-based metafeatures.

E.1.2 *Using explanations with a tolerance*

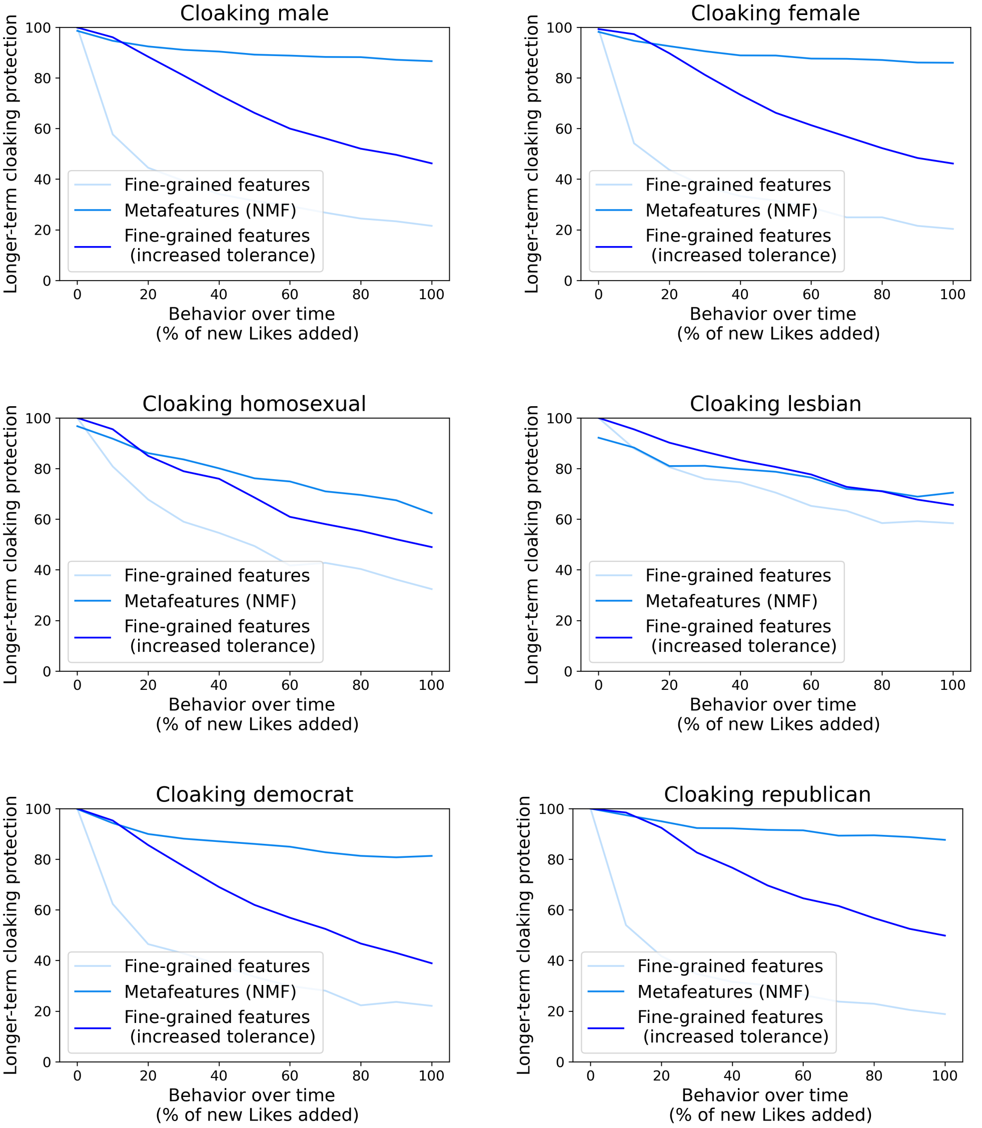


Figure E.11: Longer-term cloaking protection over time when using explanations with an additional tolerance level.

COLOPHON

This document was typeset in LaTeX using the typographical look-and-feel classicthesis. Most of the graphics in this thesis are generated using pgfplots and pgf/tikz. The bibliography is typeset using natbib.