

Algorithmic Fairness in Practice: Data, Models, and Interactions

David Solans Noguero

DOCTORAL THESIS UPF / Year 2022

THESIS SUPERVISORS

Prof. Dr. Carlos Castillo & Prof. Dr. Caterina Calsamiglia

DEPARTMENT

Department of Information and Communications Technologies (DTIC)



To María and Marco, I could not have undertaken this journey without their help, guidance and lovely support.

Acknowledgments

This thesis has been completed thanks to the collaboration of many people and institutions. Therefore, I would like to thank all the people who have been part of this journey.

First, I want to acknowledge my thesis advisors, Carlos Castillo (ChaTo) and Caterina Calsamiglia. I will always be thankful for their patience and support, and for having shared with me their time, experience and knowledge through these last years.

I would also like to thank all my former colleagues in CA Technologies: Oscar Rípollés, Dòrica Munell, Alberto Huélamo, David Sánchez, Jaume Ferrarons, Marc Solé, Jacek Dominiak and Michał Zasadziński. Thanks for being with me at the beginning of this journey. I really appreciate your friendliness and all your support and guidance.

I am grateful for the collaboration of Battista Biggio, Karolin Keppler, Aideen Farren, Christopher Tauchman, Francesco Bonchi, Anna Monreale and Andrea Beretta in the development of successful research projects. Additionally, I have to thank all the people involved in ALOHA-H2020 for all the effort done to keep me involved in the project.

I would also like to thank the members of the Information and Communication Technologies Department of Universitat Pompeu Fabra, including Lydia García, Ruth Temporal and Jana Safrankova.

I would like to especially thank my colleagues, collaborators and ex-members of the Web Science and Social Computing Research Group (WSSC): Ana Freire, Diana Ramírez-Cifuentes, Maria Rauschenberger, Francesco Fabbri, Valerio Lorini, Fedor Vitiugin, Marzieh Karimi-haghghi, Ricardo Baeza-Yates, Vladimir Stivill and Marina Estevez.

Special thanks to Manuel Portela, for all those meaningful conversations and all the valuable time expend together.

Additional thanks again to Diana and Ana, excellent researchers, with whom doing research has been so nice and enjoyable.

Thanks again to Francesco Fabbri, colleague, coauthor and teammate those evenings playing basketball.

Finally, and most importantly, none of this work could have been possible without the support of my family and friends, with that list being too long to be included here, thanks for everything.

Funding

This thesis has been partially funded by project ALOHA-H2020 and project So-BigData++, grant agreements 780788 and 871042 of the European Union's Horizon 2020 program.

Abstract

The fast-growing adoption of technologies based on Machine Learning (ML), in addition to the large scale at which they operate, makes them a potential source of systematic discrimination against disadvantaged social groups. This thesis is framed within the topic of Algorithmic Fairness, which aims at detecting, characterizing and mitigating those inequalities replicated, amplified or created by such autonomous systems. The wide range of domains where ML systems are being incorporated drives the multidisciplinarity of this work, with contributions in fields as varied as social sciences in the study of Artificial Intelligence (AI), detection of mental health disorders or the real state market. At the same time, it contributes to different stages of the ML life-cycle, facilitating the creation of fairer ML-based systems. With the outcomes of this thesis, we expect to contribute to the further development of tools and mechanisms to assist practitioners into incorporating Algorithmic Fairness.

Keywords: Algorithmic Fairness, Machine Learning, Discrimination, Algorithmic Assessment.

Resumen

La rápida adopción de tecnologías basadas en Aprendizaje Automático (ML según sus siglas en inglés), en adición a la larga escala a la que estos sistemas operan, los convierte en una fuente potencial de discriminación sistemática en contra de los grupos sociales más desfavorecidos. Esta tesis se enmarca dentro del campo de Justicia Algorítmica, que estudia la detección, caracterización y mitigación de desigualdades replicadas, amplificadas o creadas por sistemas automáticos. El amplio rango de dominios en los que se están incorporando sistemas basados en Aprendizaje Automático conlleva la multidisciplinariedad de este trabajo, con contribuciones al estado del arte de campos de la investigación tan variados como el estudio de la Inteligencia Artificial (AI) desde el campo de las ciencias sociales, la detección de problemas de salud mental o plataformas online que operan en el mercado inmobiliario. Al mismo tiempo, contribuye a distintas etapas del ciclo de vida de los modelos de Aprendizaje Automático, buscando facilitar la creación de modelos más justos. Con los resultados de esta tesis, esperamos contribuir al desarrollo de herramientas y mecanismos que permitan asistir a los profesionales a incorporar Justicia Algorítmica en sus desarrollos e investigaciones.

Palabras clave: Justicia Algorítmica, Aprendizaje automático, Discriminación, Evaluación Algorítmica.

Resum

La ràpida adopció de tecnologies basades en Aprendentatge Automàtic (ML segons les sigles en anglès), en addició a la llarga escala a què aquests sistemes operen, els converteix en una font potencial de discriminació sistemàtica en contra dels grups socials més desfavorits. Aquesta tesi s'emmarca dins del camp de Justícia Algorítmica, que estudia la detecció, caracterització i mitigació de desigualtats replicades, amplificades o creades per sistemes automàtics. L'ampli rang de dominis en què s'estan incorporant sistemes basats en Aprendentatge Automàtic comporta la multidisciplinaritat d'aquest treball, amb contribucions a l'estat de l'art de camps de recerca tan variats com l'estudi de la Intel·ligència Artificial (AI) des del camp de les ciències socials, la detecció de problemes de salut mental o plataformes en línia que operen al mercat immobiliari. Alhora, contribueix a diferents etapes del cicle de vida dels models d'aprendentatge automàtic, buscant facilitar la creació de models més justos. Amb els resultats d'aquesta tesi, esperem contribuir al desenvolupament d'eines i mecanismes que permetin assistir els professionals a incorporar Justícia Algorítmica als seus desenvolupaments i investigacions.

Paraules clau: Justícia Algorísmica, Aprendentatge Automàtic, Discriminació, Avaluació Algorísmica.

Contents

List of figures	xviii
List of tables	xix
List of acronyms	xx
1 INTRODUCTION	1
1.1 Motivation	2
1.1.1 Outline and contributions	4
1.2 Structure of the thesis	5
1.2.1 Chapter 2	5
1.2.2 Chapter 3	5
1.2.3 Chapter 4	6
1.2.4 Chapter 5	6
1.2.5 Chapter 6	6
1.2.6 Chapter 7	7
1.2.7 Chapter 8	7
2 BACKGROUND	9
2.1 Algorithmic Fairness in supervised classification	10
2.2 Algorithmic Fairness in two-sided markets	12
2.3 Algorithmic Fairness in industrial settings	13
2.4 Algorithmic Auditing	15
3 CHARACTERIZING BIASES IN INPUT DATA	17
3.1 Introduction	17
3.1.1 Motivation.	18
3.1.2 Contributions.	18
3.1.3 Chapter structure.	19
3.2 Related Work	19
3.3 Poisoning Fairness	20

3.3.1	Attack Formulation	20
3.4	Gradient-Based Attack Algorithm	22
3.4.1	White-Box and Black-Box Poisoning Attacks	25
3.5	Experiments	26
3.5.1	Experiments with synthetic data	26
3.5.2	Experiments with real data	28
3.6	Conclusions	31
4	PROFESSIONAL BIAS	33
4.1	Introduction	33
4.1.1	Motivation	33
4.1.2	Contributions	34
4.1.3	Chapter structure	34
4.2	Related Work	35
4.2.1	Economics of Convention (EC)	35
4.2.2	Content Analysis of Open Source Projects	35
4.2.3	Content Analysis of Scientific Articles	36
4.2.4	Content Analysis of Online Discussions	36
4.3	Economics of Convention (EC)	37
4.4	The EC Dataset	39
4.4.1	GitHub	39
4.4.2	Semantic Scholar (S2)	40
4.4.3	Reddit	40
4.5	Methods for Building the EC Model	41
4.5.1	Model Selection	41
4.5.2	Labeling of Dataset and Active Learning	42
4.6	Results	45
4.6.1	Evaluation of Conventions	47
4.7	Discussion	50
4.7.1	Limitations	52
4.8	Conclusion	52
5	EVALUATING BIASES IN A TWO-SIDED MARKET PLATFORM	55
5.1	Introduction	55
5.1.1	Motivation	55
5.1.2	Contribution	57
5.1.3	Chapter structure	57
5.2	Related Work	58
5.2.1	Access to housing	58
5.2.2	Algorithmic fairness in double-sided markets	58
5.3	Setting	59

5.4	Methodology	60
5.5	Dataset description	63
5.6	Results	65
5.7	Discussion and conclusions	75
6	DETECTING AND MITIGATING BIASES IN CLASSIFICATION	77
6.1	Introduction	77
6.1.1	Motivation	77
6.1.2	Research questions	78
6.1.3	Contributions	78
6.1.4	Chapter structure	79
6.2	Related work	80
6.2.1	Characterization and assessment of mental disorders on social media	80
6.2.2	Algorithmic fairness for detecting mental health status . .	80
6.2.3	Bias mitigation in classification	81
6.2.4	Training calibrated classifiers	82
6.3	Methods	83
6.3.1	Dataset	83
6.3.2	Bias detection	84
6.3.3	Bias characterization	85
6.3.4	Bias Mitigation	85
6.4	Results	87
6.4.1	Bias Detection	87
6.4.2	Bias Characterization	88
6.5	Bias mitigation	89
6.5.1	Training fair classifiers	90
6.5.2	Training calibrated classifiers	91
6.6	Discusion	92
6.6.1	Limitations	93
6.6.2	Ethical concerns	94
7	UNCOVERING BIASES IN USER-MACHINE INTERACTION	95
7.1	Introduction	95
7.1.1	Motivation	95
7.1.2	Research question	96
7.1.3	Contributions	96
7.1.4	Chapter structure	97
7.2	Related work	97
7.2.1	Decision Support Systems (DSS)	97
7.2.2	Trust and Reliance on Decision Support	98

7.2.3	Effects of DSS Accuracy	98
7.3	Methodology	99
7.4	Overview	100
7.4.1	Platform	100
7.4.2	Experiment variables	102
7.5	Experimental Design	103
7.5.1	Main Experiments	104
7.6	Results	105
7.6.1	Score	105
7.6.2	Time	112
7.7	Discussion	114
7.7.1	Limitations	115
8	CONCLUSIONS	117
8.1	Main contributions	117
8.2	Recommendations for practitioners	118
8.2.1	Poisoning attacks	118
8.2.2	Professional biases	119
8.2.3	Algorithmic assessments	119
8.2.4	Algorithmic unfairness mitigation	119
8.2.5	Interaction bias	120
8.3	Future work	120
8.3.1	Poisoning attacks	120
8.3.2	Professional biases	120
8.3.3	Algorithmic assessments and bias mitigation	121
8.3.4	Interaction bias	121
8.4	Reproducibility	121

List of Figures

1.1	Cyclic nature of bias, also known as vicious cycle of bias or second order bias	3
2.1	Algorithmic fairness contributions in terms of the Machine Learning lifecycle	14
3.1	Attacker’s loss $\mathcal{A}(\mathbf{x}_c, y_c)$ (<i>left</i>) and disparate impact (<i>right</i>) as a function of the attack point \mathbf{x}_c with $y_c = 1$, on a bi-dimensional classification task. Note how the attacker’s loss provides a smoother approximation of the disparate impact, and how our gradient-based attack successfully optimizes the former, which amounts to minimizing disparate impact, compromising algorithmic fairness.	23
3.2	Gradient-based poisoning attack against a logistic classifier, on a bi-dimensional classification task. The classification function and the corresponding decision regions are reported before (<i>left</i>) and after (<i>right</i>) injection of the poisoning samples (red and blue stars in the right plot).	25
3.3	(Best seen in color.) Examples of generated synthetic data sets for different values of the separation S between groups. Privileged elements ($z = +1$) are denoted by circles and unprivileged elements ($z = -1$) by crosses. Favorable labels ($y = +1$) are in green, while unfavorable labels ($y = -1$) are in red.	27
3.4	Comparison of the original model against the model generated by the White-box attack and Black-box attacks, for ten synthetic datasets generated by different separation parameters (S). Each data point is the average of ten runs of an attack. We observe that attacks have a moderate effect on the accuracy of the classifier, and can affect the classifier fairness (demographic parity and odds difference) to an extent that becomes more pronounced if the original dataset already has a large separation between classes (larger values of S).	29

3.5	Comparison of the original model against the model generated by a White-box attack and a Black-box attack, for varying percentages of poisoned samples. The main difference between both types of attack is that the black-box attack starts having more noisy behaviour also drastically reducing the accuracy of the classifier (thus being more easily detectable) when the percentage of poisoned samples exceeds a certain threshold (about 20%).	30
3.6	Transferability of the attacks from Logistic Regression to other models.	31
4.1	Active learning pipeline to collect and verify training data	43
4.2	Percentage of conventions in each data subset for AI and non-AI related items as predicted by the classifiers.	45
4.3	Confusion matrix of EC classifiers using the obtained calibration threshold	47
4.4	Co-occurrences of conventions in the predictions for AI subsets. Values in the matrices are normalized by the number of sentences in each data source.	48
5.1	Platform’s recommendation pipeline.	59
5.2	$nDCG$ score of recommendations for BSL.	66
5.3	$nDCG$ score difference between the RS and BSL across demographic groups.	66
5.4	$nDCG$ difference between ML-models and BSL.	67
5.5	Exposure distribution comparison (log-scale) between BSL and RS: total (Aggregated) and by demographics (Gender, Age, Language and Sexual-Orientation). The dashed lines in each violin plot represent the first, second and third quartile.	67
5.6	Exposure for the different models across demographic groups.	68
5.7	CR score for BSL across demographic groups	69
5.8	Conversion Rate (CR) differences between each ML model and BSL.	70
5.9	Conversion Rate (CR) differences w.r.t. BSL for each model.	70
5.10	CTR differences across different models	71
5.11	$nDCG$ -Fairness tradeoff. Listers between $nDCG$ and disparities per across groups for $nDCG$ (top). Seekers trade-off between $nDCG$ and Exposure (bottom).	73

6.1	(best seen in color) average Balanced Accuracy (bAcc) and False Negatives Ratio (FNR) compared to bAcc ratio and FNR ratio across genders on the trained models. Figures on the left show scenario (a) -unique model- and figures on the right show scenario (b) -one model per gender-	87
6.2	Labelers performance with respect to obtained ground truth.	89
6.3	Calibration curves obtained for original model and calibrators	91
7.1	Experiment interface. The terrain is represented by green (forest) and brown (desert) cells. The user has drilled in the cells in black, and a recommendation from the DSS (in yellow) is shown.	96
7.2	Selected easy, medium, and hard maps, displaying terrain (top) and oil yield distribution (bottom). The terrain is visible to participants; green cells represent forest, and yellow cells represent desert. The oil yield is hidden; darker shades indicate higher yield.	104
7.3	Distribution of scores in the three maps (easy, medium, hard), without machine assistance (left) and with machine assistance (right). In statistical significance tests, “ns” stands for no significance, and asterisks significance at: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$).	106
7.4	Score distribution by DSS quality. We compare the score obtained by human participants with machine assistance and the score that the machine would obtain.	107
7.5	Score distribution using a <u>biased</u> DSS or an <u>unbiased</u> one, for the <u>low</u> (LB vs LU) and <u>high</u> cost (HB vs HU) conditions.	107
7.6	Score comparison of <u>biased</u> (LB) and <u>unbiased</u> (LU) DSS under a <u>low</u> cost condition.	108
7.7	Score comparison of <u>biased</u> (HB) and <u>unbiased</u> (HU) DSS under a <u>high</u> cost condition.	108
7.8	Probability of “bad plays” (below median score) under different DSS accuracy.	109
7.9	Probability of “bad plays” (below median score) under different conditions of DSS bias. HB and HU are <u>biased</u> and <u>unbiased</u> DSS, respectively, in the <u>high</u> cost condition. LB and LU correspond to the <u>low</u> cost condition.	109
7.10	Centroids of clusters illustrating how the average score per play increases as players learn during each game.	111
7.11	Exploration/exploitation behavior and performance. We associate clicks near each other (first column) with exploitation, and clicks far from each other (third column) with exploration. Values indicate percentage of plays.	111

7.12	Completion time distribution by map. For representation purposes, outlier games taking more than 100 seconds have been removed.	112
7.13	Average distance between recommendation and selected cell by model quality. The horizontal line (E) is the expected distance between two random points in a 32x32 grid	113
7.14	Results of the exit survey on technology acceptance, by DSS accuracy.	113

List of Tables

2.1	Contingency table. Distribution of features and samples space in algorithmic fairness literature	11
4.1	Registers of worth in the Economics of Convention	38
4.2	A combination of the top five keywords in the dataset per convention established by manual analysis and TF-IDF frequency	39
4.3	Counts of sentences and items for AI and non-AI subsets from each data source. Depending on the specific data source, items refer to repositories, abstracts or threads.	40
4.4	Comparison of model performance per convention	44
4.5	Model performance per data source	44
5.1	Summary of the number of recommendations created with the different models through the operation of the platform. BSL is the random baseline; the other models are based on Machine Learning.	64
5.2	Percentage of seekers (S) and listers (L) belonging to different groups.	64
6.1	Base rates for each class and gender in the dataset.	83
6.2	Types of features included in the selected dataset.	84
6.3	Obtained performance in terms of Accuracy and FNR across genders for each bias mitigation technique.	90
7.1	Experimental conditions, with “L” representing low cost of drilling a forest cell, and “H” representing high cost. LB and HB correspond to cost-unaware decision support, which yields biased suggestions. LU and HU correspond to cost-aware decision support. .	105

List of acronyms

AI Artificial Intelligence

ML Machine Learning

FAccT Fairness, Accountability and Transparency

GDPR General Data Protection Regulation

DSS Decision Support System

RecSys Recomender System

LogReg Logistic Regression model

MLP Multilayer Perceptron

SVM Support Vector Machine model

bAcc Balanced Accuracy

FNR False Negatives Ratio

FPR False Positives Ratio

AN Anorexia Nervosa

EDNOS Eating Disorder Not Otherwise Specified

Chapter 1

INTRODUCTION

In recent years, we have witnessed unprecedented improvements in the automation of a broad range of processes that impact our daily life thanks to the application of Artificial Intelligence (AI) and, in particular, the adoption of Machine Learning (ML) techniques. This causes that an increasing number of decisions regarding human beings' daily routines take place in partially automated environments, what imposes that their lives and fundamental rights are often directly or indirectly affected by the functioning of such quasi-autonomous systems. Although these systems may bring innumerable benefits, their use also implies inherent risks, such as codifying biases or a reduction of accountability on their decisions. From the perspective of the academic world, the majority of the research around the field of AI is concentrated on how to build more precise, reliable and advanced models, while fewer researchers have focused on the impact and unintended harms such systems might create. In this context, the Fairness, Accountability and Transparency (FAccT) community (and its predecessors FATML and FAT^{*}) were born in the mid 2000s and early 2010s with the objective of studying the inequalities occurring in specially concerning applications. Examples of such applications are, as stated by the General Data Protection Regulation (GDPR)¹ in its article 22, cases where the predictions produce legal effects. Similarly, the AI Act² defines "High Risk" applications as such systems that pose significant risks to the health and safety or fundamental rights of persons. The object of study of this work are then those scenarios where decisions that are directly or indirectly affecting human beings and their fundamental rights are taken or assisted by automated systems, with a special interest on the protection of social disadvantaged groups.

¹<https://gdpr.eu/>

²<https://artificialintelligenceact.eu/>

1.1 Motivation

ML systems tend to exhibit important disparities of performance across demographic groups. This is often caused by biases present in the input data that is used to train such systems, which reflects historical stereotypes and prejudices while also tends to contain fewer examples about social minorities, leading to lower performance for such groups.

Before the wide spread of Algorithmic Fairness research, a dangerous but prevalent reasoning by ML practitioners was the idea that that ML systems were not incurring in subjectivity or replicating social prejudices, for the reason that they learn from objective data and features. Several cases of inequalities replicated, created or amplified by AI and more in particular, ML-based systems have been reported previously, showing that often, training data reflects social stereotypes and prejudices between other biases.

Famous examples of how data biases affect the predictions of automated systems can be found in the case of COMPAS, a recidivism prediction tool used in the U.S., where ProPublica identified a much higher false positive rate for black people (see [109]); XING, a job platform that was reported to rank less qualified male candidates higher than more qualified female candidates [120]; or the case of face recognition online services found to suffer from achieving much lower accuracy on females with darker skin color [40]. Other not so well-known examples on how biases reflect on AI-based systems include the quality of education and healthcare received [35, 101, 70, 73], news or social media people see [161, 37, 6], who receives a job [120, 162, 178], who is released from jail [109, 52] and who is subjected to increased policy [132, 204].

New legislation explicitly prohibits entirely automated processing with significant effects, similar to legal effects, over a person, as the article 22 of the GDPR. Also, the U.S. Civil Rights Act [201] defines a rule setting a maximum ratio of 80% in the probability of accessing an employment for two given classes of individuals. The topic of fairness and non-discrimination has been given a special interest in a variety of AI stakeholders.

During the past few years, the intersection between data mining, machine learning and fairness domains has received an increasing attention by the research community. Under the awareness of the potential social harm of automatic systems used to assist humans in decision-making procedures, researchers have proposed several formulations of fairness for a variety of machine learning tasks such as classification [111, 141, 84, 117, 202, 93, 217, 43], regression [5, 25], recommendation [41, 26], ranking [48, 27, 47, 218] and natural language processing [194, 148] between others.

Whereas most of the existing literature refers to *Data Bias* as the main initiator of biases, the reasons for which data reflects biases and prejudices are often com-

plex and should not be isolated from the predictive system and its environment.

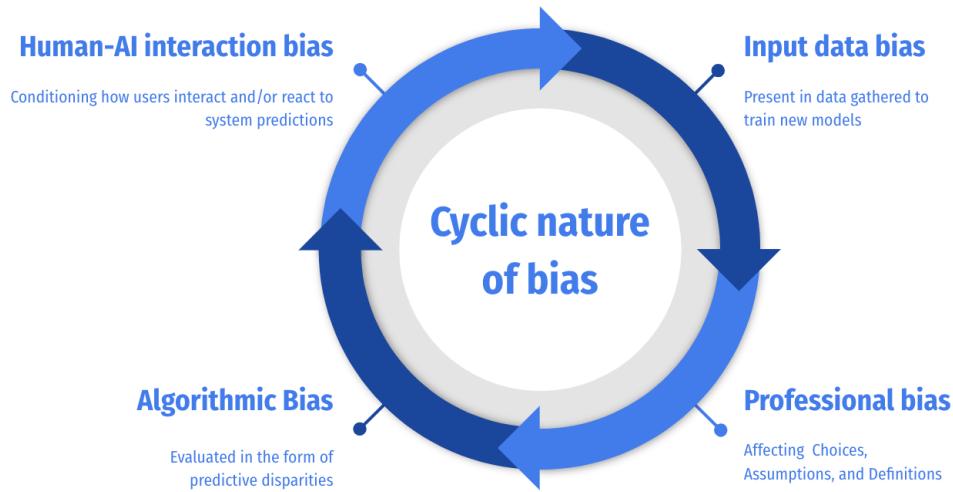


Figure 1.1: Cyclic nature of bias, also known as vicious cycle of bias or second order bias

Figure 1.1 describes the general framework we propose in this thesis. Some authors refer to similar concepts as the vicious circle of bias [203] whereas other call this second order bias [14]. In this work, we refer about this conception as the *Cyclic nature of bias*. Apart from the data itself, biases manifest during the definition, development and assessment of the systems carried out by ML practitioners. At these stages, humans might incorporate their own biases, incurring in the so-called *Professional bias*. This phenomenon has been already observed by psychologists, who reported this effect as caused by the so-called researchers degrees of freedom [184]. In their study proved the impact of four common degrees of freedom: flexibility in (a)choosing among dependent variables, (b)choosing sample size, (c)using covariates, and (d)reporting subsets of experimental conditions. As an example of the effects that professional bias can cause in the system, researchers created an experiment where twenty-nine teams with a total of 61 data-scientists took part. In the experiment, participants were all given the same data set and prompt: Do soccer referees give more red cards to dark-skinned players than light-skinned ones? As a result, and despite analyzing the same data, the researchers got a variety of results: twenty teams concluded that soccer referees gave more red cards to dark-skinned players, and nine teams found no significant relationship between skin color and red cards [10]. This type of analysis prove that, even when executed by professionals with good indent, experiments can lead to different conclusions, just because of the effects of their degrees of freedom.

Such freedom is reflected in the set of decisions taken during the execution of the experiments.

As a result of this process, algorithmic might inherit different types of biases that are then reflected on disparities in predictive performance. This known as *Algorithmic Bias* and is the source of discrimination when the disparities of performance systematically affect individuals with certain personal or inherited characteristics. After, humans interact with the predictions of such potentially biased algorithms, also reflecting their own prejudices and biases. This is the case when collecting data from social platforms [146]. This leads to a complex cycle where biases are incorporated, if not augmented, at different stages in the process.

In this dissertation, we tackle each of the stages separately, as described in the following section of this document.

1.1.1 Outline and contributions

We study various elements of Algorithmic Fairness with a particular focus on the following five, which happen in different stages of the Cyclic Nature of Bias shown in Figure 1.1.

1. We study the concept of bias in input data. For that, we show how a novel type of poisoning attack can be used to craft new samples that once added to the training set lead to more unfair learning model with minimal affection of predictive accuracy.
2. The role of professional biases through the lens of Economics of Conventions, a theory created by economists and sociologists that defines a taxonomy of normative orders of worth that are used as collective logics of coordination and evaluation. We analyze both qualitatively and quantitatively AI-related data collected from Github, Semantic Scholar and Reddit showing that the prevalence of convention varies significantly across data sources and stakeholders (developers, researchers, general public).
3. We performed an algorithmic assessment of a collection of recommender systems used in an application in the room rental market. Our analysis, focused in both sides of the market (room seekers and room owners) reveal relevant disparities in the quality of the recommendations got by different social groups (by gender, age, spoken languages or sexual orientation).
4. We detected dangerous disparities in predictive performance when training a ML model for classifying anorexia nervosa profiles from social networks. We evaluated such disparities across different types of models and show

how different repairing methods could be applied depending on the objectives of the tool, proving at the same time that there is no a perfect solution but a trade-off.

5. We studied the response of users to a decision support system (DSS). For that, we developed an experimental platform based on an online game where users get a DSS helping them to maximize their score. We use a crowd-sourcing platform to recruit more than 400 participants. Our results show different behavioral patterns that can be analyzed more in detail in further research.

1.2 Structure of the thesis

The structure of this manuscript is divided in the following chapters:

1.2.1 Chapter 2

This chapter provides an overview of the state of the art in topics related to the work developed during the execution of this thesis. In particular, we review previous work related to: i) Algorithmic Fairness in classification ii) Algorithmic Fairness in two-sided markets iii) Algorithmic Auditing iv) Algorithmic Fairness in industrial settings

1.2.2 Chapter 3

This chapter presents a study done in the context of data bias. For that, we leverage techniques used in the field of adversarial learning, where new samples are crafted with the objective of causing malfunctioning of the machine learning model. In particular, we use poisoning attacks, what aims to modify input data to affect the learned decision boundary. Using that, we propose a framework to create new samples that increase the potential data bias present in the training set. The obtained results show how the proposed framework can be used to amplify predictive disparities on the learned model, with minimal affection on predictive accuracy and with minimal previous knowledge by the attacker. This contribution corresponds to the first investigation done on Poisoning attacks for Algorithmic Fairness.

The work of this chapter was published in full at: *David Solans, Battista Biggio, Carlos Castillo: Poisoning Attacks on Algorithmic Fairness. In Proceedings of European Conference on Machine Learning and Principles and Practice of*

Knowledge Discovery in Databases (ECML/PKDD) 2020, LNCS Volume 12457, pp. 162-177.

1.2.3 Chapter 4

In this chapter we study how different stakeholders of AI: (i) researchers; (ii) developers; (iii) general-public use different values of worth when discussing AI. In order to quantify this, we use the theory of Economics of Conventions [63] that defines a taxonomy of values of worth. We gather data from Semantic Scholar, Github and Reddit and then analyze it by using both qualitative and quantitative analysis. For the quantitative analysis, we build a text classifier based on deep-learning that classifies sentences into conventions. The obtained results outline important differences in prevalence of conventions across the analyzed data sources.

The work described in this chapter was published in full at: *David Solans, Christopher Tauchmann, Aideen Farrell, Karolin Kappler, Hans-Hendrik Huber, Carlos Castillo: Learning to Classify Morals and Conventions: Artificial Intelligence in Terms of the Economics of Convention. Proceedings of the International Conference on Social Media (ICWSM) 2021, pp. 691-702*

1.2.4 Chapter 5

This chapter describes an algorithmic assessment done on a two-sided platform used in the real-state market. In our evaluation, we compare user satisfaction in both sides of the market with different versions of the recommender system, revealing trade-offs causing the highly performant classifiers to lead to higher inequalities in user satisfaction.

The work described in this chapter was published in full at: *David Solans, Francesco Fabbri, Caterina Calsamiglia, Carlos Castillo, and Francesco Bonchi. 2021. Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES). Association for Computing Machinery, New York, NY, USA, 978–988.*

1.2.5 Chapter 6

The work described in this chapter is framed within the domain of detection of mental health issues in social media posts. To do so, we use a dataset collected by Ramírez-Cifuentes et al. [163] that contains positive and negative examples gathered from twitter and labeled by professionals of psychology. Building different versions of machine learning model classifiers, we demonstrate how they

yield disparities in predictive performance, with females having higher rates of false negative rates. After that, we characterize the reasons for these disparities of performance using a dataset with higher prevalence of female samples. Finally, we use different state-of-the-art techniques to train fairer models, discussing the potential trade-offs of each of the options in different scenarios.

The work described in this chapter is described in a manuscript at: *David Solans, Diana Ramírez-Cifuentes, Esteban Ríssola, Ana Freire. 2022. Gender bias when using Artificial Intelligence to assess Anorexia Nervosa on Social Media. This manuscript was submitted to a Q1 Journal and is currently under review.*

1.2.6 Chapter 7

In this chapter we review the results obtained from a user study. To conduct the experiment, we first design an experimental platform consisting in an online game where treatment users have a machine learning-based Decision Support System (DSS) recommending them the best place for the next play whereas the control group does not have any automated help. We test different conditions: (i) levels of accuracy of the system; (ii) presence/absence of biased predictions; and evaluate the human response to them, both implicitly in their behavior and explicitly through an opinion survey after the game. The results show that in general, users are able to detect malfunctioning

The work described in this chapter is described in a manuscript at: *David Solans, Andrea Beretta, Manuel Portela, Carlos Castillo, Anna Monreale. 2022. Human Response to an AI-Based Decision Support System: A User Study on the Effects of Accuracy and Bias. At the time of this writing, we were planning to submit to a CORE-A³ conference sponsored by ACM.*

1.2.7 Chapter 8

We finally conclude this dissertation by summarizing and discussing the main findings and suggesting new directions for future research.

³<http://portal.core.edu.au/conf-ranks/>

Chapter 2

BACKGROUND

In this section, previous work related to the general objectives of this thesis is outlined. For chapter, the specific state of the art is referenced in subsequent sections of this document.

Despite this thesis referencing over 200 previous papers, it lies within an enormous, multidisciplinary field that has branched in many directions, and hence there are many more references than listed here.

Between the set of topics that could be outlined in this section, we selected four particular topics because of their relation to the general objectives of this thesis. We selected Algorithmic Fairness in supervised classification for its relation to Chapters 3 and 6 of this thesis. Additionally, supervised classification is the subdomain in Algorithmic Fairness with higher number of contributions. After, we review Algorithmic Fairness in two-sided markets, strongly related to Chapter 5 and a growing concern given the large scale at that platforms such as Amazon, Spotify, Uber or Airbnb between others do operate. Then, we review the field of Algorithmic Fairness in industrial settings, which overviews the concerns AI practitioners have while trying to incorporate Fairness. Finally, we overview the field of Algorithmic Auditing, that is closely related to Chapter 5 but at the same time, will be one of the strongest interest of industrial firms, governments and consultancies in the coming years, where the need to certify ML-based systems might become a legal requirement ¹.

¹<https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook>

2.1 Algorithmic Fairness in supervised classification

Statistical biases present in data can lead to unfairness in subsequent learning tasks. In Olteanu et al. [146], the authors describe a complete list of bias types and the existence of feedback cycles from data origins to its collection and processing. Researchers in Mehrabi et al. [138] created a taxonomy of biases, depending on the application layers that are exposed to them:

- **Input data biases.** Typically coming from social prejudices reflected in the data or caused by sampling biases between others. Examples of biases reflected in the input data are *Historical Bias*, *Representation Bias*, *Measurement Bias* or *Aggregation bias*
- **Algorithmic Biases.** Producing biases such as popularity bubbles in recommendation systems or evaluation biases. Examples of algorithmic bias are *Measurement Bias*, *Evaluation Bias*, *Ranking Bias* or *Popularity Bias*.
- **Human-Computer-Interaction biases.** Reflected as behavioral biases or presentation biases, linking biases, etc. Examples of such biases are *Linking Bias*, *Social Bias*, *Observer bias* or *Self-selection Bias*.

The presence of such biases in any of the three layers of the system can potentially lead to inequities for certain groups of individuals. In this context, the task of quantifying inequities, named algorithmic fairness, has received a lot of attention in the research community, resulting in a multitude of metrics to measure fairness.

The existing metrics of fairness for classification fall in two categories: *Individual metrics* and *Group metrics*, depending on if the goal is to require the system to give similar predictions to individuals that are considered similar or the objective is that the system treats groups equally.

Groups are often calculated on the basis of sensible attributes that are legally recognized as protected classes. Examples of such classes and of their legal protection are: i) Race (*Civil Rights Act of 1964*) ii) Color (*Civil Rights Act of 1964*) iii) Sex (*Equal Pay Act of 1963; Civil Rights Act of 1964*) iv) National origin (*Civil Rights Act of 1964*) v) Citizenship (*Immigration Reform and Control Act*) vi) Age (*Age Discrimination in Employment Act of 1967*) vii) Pregnancy (*Pregnancy Discrimination Act*)

Typically, individuals of such groups are divided between unprivileged and privileged subgroups, as identified in the historical data.

In general, approaches to Algorithmic Fairness are based on a specific distribution of the feature space. As indicated by Pedreschi et al. [153], this distribution

can be expressed in the form of a contingency table that divides the features space in terms of the groups found in the data. Groups are calculated with respect to a protected attribute A and a target variable Y , for which some possible values are considered more beneficial than others. Following these ideas, for the simple case of binary classification and two groups, the feature space is divided as depicted in Table 2.1.

		benefit		
		denied	granted	
group	unprivileged	a	b	n_1
	privileged	c	d	n_2
		m_1	m_2	n

Table 2.1: Contingency table. Distribution of features and samples space in algorithmic fairness literature

As proposed by Hardt et al. [19], fairness functions are defined by using the criteria of *independence*, *separation* and *sufficiency* depending on the final goal of the metric.

Given X , a features vector representing individuals; A , a sensitive attribute; C , a predictor and Y being a target variable, the stated criteria are defined as follows:

- **Independence:** C independent of A ($C \perp A$).

Requires that for all groups a, b and all values y , $P_a(C = y) = P_b(C = y)$. These criteria ignores possible correlations between Y and A .

The case of the perfect predictor ($C = Y$) is in general not accepted in most cases, where initial rates for Y values across groups are not equalized.

Also called *disparate Impact* in the literature [72], examples of metrics based in these criteria are: *demographic parity* and *statistical parity*.

- **Separation:** C independent of A conditional on Y ($C \perp A|Y$).

It incentives to reduce errors uniformly in all groups. The case of the perfect predictor $C = Y$ is allowed and corresponds to the ideal scenario. Some authors [215] refer to these criteria with the name of *Disparate mistreatment*. Examples of metrics based on it are: *true positives ratio*, *false positive ratio*, *false negative ratio* that for the best case scenario should be equalized for the obtained groups.

- **Sufficiency:** Y independent of A conditional on C ($Y \perp A|C$)

Also expressed as: $P(Y = 1|X = x, A = a) = P(x)$, this metric ensures

that even if correct label Y provides information about the protected attribute A , predictions Y' do not provide any additional information about A . Often measured with the Equalized odds metric, sufficiency can be achieved by model calibration.

The impossibility theorem of satisfying any two of the three criteria at the same time is already known. In Choulechova et al.[52] and Kleinberg et al. [114], the authors demonstrated the unfeasibility of achieving separation and sufficiency. The work presented by Hardt et al. [19] proves the theorem for the rest of pairs.

Even in the case of trying to create fair systems, given those theorems, developers, decision makers and society in general need to specifically decide the metric or group of metrics to be optimized. In his genial tutorial, the author of [144] explains shows how conflicting objectives of different stakeholders can lead to the selection of distinct fairness metrics, demonstrating that the selection also depends on the specific role of each user in the system.

There are different methods to ensure that fairness criteria are satisfied in classification algorithms; they can be divided into pre-processing of training data [110, 45, 165], post-processing of the outcomes of the algorithm [93, 153], or formulating fairness criteria as constraints or part of the objective function optimized during training, i.e., in-processing [4, 215].

2.2 Algorithmic Fairness in two-sided markets

Two-sided (or more generally multi-sided) markets are defined as markets in which one or several platforms enable interactions between end-users, and try to get the two (or multiple) sides “on board” by appropriately charging each side. That is, platforms court each side while attempting to make, or at least not lose, money overall [171].

In these settings, stakeholders can be often differentiated depending on their role in the platform. On one hand, *Providers* develop, create or own items/services exposed on the platform. On the other hand, *Consumers* seek to acquire items/services from the providers. Additionally, the platform itself, which intermediates and matches providers and customers based on their preferences.

The concept of Algorithmic Fairness has been extended to address multiple stakeholders [42, 2], with the problem becoming how to balance fairness demands on multiple sides of the market.

- **C-Fairness.** Focuses in the satisfaction of consumers through the quality of the presented results, often referred as the received utility.
- **P-Fairness.** Addresses the quality of the platform for producers by considering inequalities in terms of exposure.

- **CP-Fairness.** Considers both consumers and producers' perspectives.

The Authors of Shur et al. [192] analyze a double-sided market in the context of ride hailing platforms, giving an special emphasis to the role of the riders (producers). Additionally, Hutson et al. [104] analyze a similar setting, in their case online dating apps, revealing different inequities based on race and/or sexual orientation.

As an example of *CP-Fairness*, the authors of Patro et al. [151], addressed it by considering individual fairness for both consumers and producers. From the producer side, their approach aims to reduce the exposure inequality among items, whereas and from the user-side, the authors argue that the platforms should fairly distribute the utility among the customers.

2.3 Algorithmic Fairness in industrial settings

Due to an increasing awareness about the potential for autonomous systems to amplify social inequities and unfairness, recent efforts on creating guidelines for *Ethical AI* [147] mention the requirement of considering at least four different aspects: (a) Algorithmic Fairness, (b) Privacy, (c) Explainability, (d) Robustness when building autonomous systems. Also, understanding data as the main element affecting such aspects, there have been attempts to create a standard for datasets description [80].

However, and even after the creation of such guidelines and standards, industrial solutions are facing a set of challenges when trying to produce fairness-aware learning algorithms by intervening at different stages of a decision-making pipeline to produce "fair" outcomes.

In the attempt to understand industrial practitioners' needs to assess and mitigate such unfairness, the authors of Holstein et al. [102] collected a set of challenges that difficult the creation of fairness-aware products. Those challenges are:

1. *Fairness-aware data collection*

The algorithmic fairness literature emphasizes the central role of data set quality [209, 81]. However, enhancing the data collection to reduce representation inequities is a complex task whose solution highly depends on the domain.

2. *Blind spots*

The existence of potential blind spots which might stand in the way of effectively addressing fairness issues, or even thinking to monitor some forms of unfairness in the first place.

3. Need of more proactive auditing processes

With the detection of fairness issues presenting many unique auditing challenges, practitioners seem to be reactive to user complaints in contrast to the proactive approaches for detecting potential security risks.

4. Needs for more holistic auditing methods

Most of the existing literature has focused in domains where fairness can be at least partially understood in terms of well-defined quantitative metrics. However, applications involving richer, complex interactions between the user and the system reported the necessity of more holistic, system-level auditing methods.

5. Addressing detected issues

Revealing a set of challenges and needs around debugging and remediation of fairness issues. This includes, among others, needs to support of the most effective strategies to address particular issues; methods to estimate how much data is additional required for particular subgroups; processes to anticipate potential trade-offs between fairness definitions and other model quality metrics; and frameworks to help navigate complex ethical decisions

Aligned to these findings, recent contributions such as Cramer et al. [53] have also framed industrial requirements in terms of the Machine Learning life cycle, revealing areas and processes that are not being addressed from the perspective of algorithmic fairness literature.

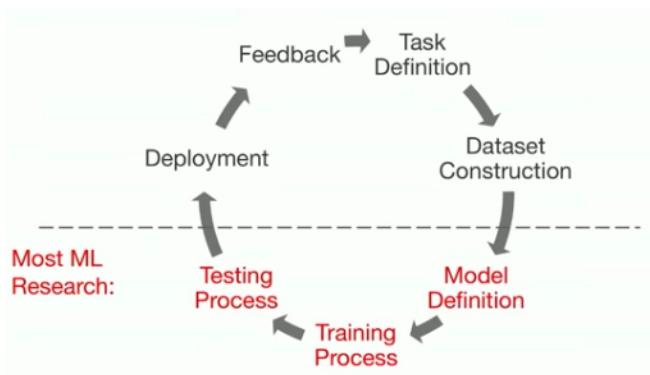


Figure 2.1: Algorithmic fairness contributions in terms of the Machine Learning lifecycle

In concrete, as shown in Figure 2.1, the report outlines a lack of assistance for evaluating performance in deployment, where the system potentially affects

millions of users, specially interesting due to the impossibility of ensuring that training data is equally representative for the whole population.

The task definition, or how to correctly define the problem that the system is trying to solve, is also an arduous task, where defining the intended effects but also the unintended and possible biases, together with the fairness requirements is a clear necessity.

After, the construction of the dataset is a critical phase where biases can manifest due to different reasons, as described by Baeza-Yates [14], all of them contributing to the creation of biased sources of information to be used in the next phases of the cycle.

2.4 Algorithmic Auditing

In many cases, perhaps in most cases, designers of computational systems fail to include accountability and transparency mechanisms “by design” [150]. In this context, Algorithmic Auditing allows us to uncover and understand potential sources of discrimination driven by such algorithmic-based decision-making, as a post-hoc solution to audit the system behavior in its past executions.

Early research on this topic includes a set of methods to detect discrimination in online platforms [179]. Eslami et al. [69] show how to detect and quantify a rating algorithm’s bias using cross-platform audit techniques in the context of hotel rating platforms. Their work identified systematic differences of ratings between three existing platforms and revealed how bias awareness can shift users’ attention from their own experience to the system as a whole, even trying to open the black-box by gaming the rating system. This work also introduces a taxonomy of methods for algorithmic auditing that frames our work as a *within-platform* study. The authors of Galdon et al., [78] describe an auditing of an application used to promote well-being among its users. This work discusses the issue of not collecting sensitive data, as required by the GDPR and the data minimization principle. This might prevent researchers from detecting biases against protected groups. Baracas et al. [18] provided a taxonomy of the choices that are involved in the designing of algorithmic evaluations.

In the specific context of search engines auditing, Mehrotra et al. [139] proposes a methodology for measuring differential satisfaction across demographics. Based on the proposed methodology, they conduct an external auditing of a search engine based on a dataset collected by a third party composed of search queries issued to the platform in a short period of time. Their analysis shows significant differences in usage patterns and evaluation metrics for different demographic groups, mainly based on age and gender. In the domain of access to housing, Asplund et al. [11] perform an algorithmic auditing of received user ads and of

the ordering of recommendations in different housing portals in the U.S. They use a strategy based on “sock puppets,” creating automatic systems that interact with the platform under fake user profiles, concluding that there are not statistical significant differences between results shown for profiles simulating different age or race. In the topic of Policy Learning in Raking, Singh et al. [186] proposes a theoretical methodology to optimize not only for the utility of the rankings for the users, but also considering fairness constrains of exposure with respect to the ranked items. Their work studies the relation between an allocation metric (normalized cumulative gain) and group disparity, measured in terms of item exposure, proposing the inclusion of exposure-allocation constrains in the learning.

Chapter 3

CHARACTERIZING BIASES IN INPUT DATA

3.1 Introduction

Algorithmic Fairness is an emerging concern in computing science that started within the data mining community but has extended into other fields including machine learning, information retrieval, and theory of algorithms [91]. It deals with the design of algorithms and decision support systems that are non-discriminatory, i.e., that do not introduce an unjustified disadvantage for members of a group, and particularly that do not further place at a disadvantage members of an already disadvantaged social group. In machine learning, the problem that has been most studied to date is supervised classification, in which algorithmic fairness methods have been mostly proposed to fulfill criteria related to parity (equality) [216]. Most of the methods proposed to date assume benevolence from the part of the data scientist or developer creating the classification model: she is envisioned as an actor trying to eliminate or reduce potential discrimination in her model.

The problem arises when dealing with malicious actors that can tamper with the model development, for instance by tampering with training data. Traditionally, *poisoning attacks* have been studied in *Adversarial Machine Learning*. These attacks are usually crafted with the purpose of increasing the misclassification rate in a machine learning model, either for certain samples or in an indiscriminate basis, and have been widely demonstrated in adversarial settings (see, e.g., [29]).

In this work, we show that an attacker may be able to introduce algorithmic discrimination by developing a novel poisoning attack. The purpose of this attacker is to create or increase a disadvantage against a specific group of individuals or samples. For that, we explore how analogous techniques can be used to compromise a machine learning model, not to drive its accuracy down, but with

the purpose of adding algorithmic discrimination, or exaggerating it if it already exists. In other words, the purpose of the attacker will be to create or increase a disadvantage against a specific group of individuals or samples.

3.1.1 Motivation.

The main goal of this chapter is to show the potential harm that an attacker can cause in a machine learning system if the attacker can manipulate its training data. For instance, the developer of a criminal recidivism prediction tool [109] could sample training data in a discriminatory manner to bias the tool against a certain group of people. Similar harms can occur when training data is collected from public sources, such as online surveys that cannot be fully trusted. A minority of ill-intentioned users could *poison* this data to introduce defects in the machine learning system created from it. In addition to these examples, there is the unintentional setting, where inequities are introduced in the machine learning model as an undesired effect of the data collection or data labeling. For instance, human annotators could systematically make mistakes when assigning labels to images of people of a certain skin color [40].

The methods we describe on this chapter could be used to model the potential harm to a machine learning system in the worst-case scenario, demonstrating the undesired effects that a very limited amount of wrongly labeled samples can cause, even if created in an unwanted manner.

3.1.2 Contributions.

This work first introduces a novel optimization framework to craft poisoning samples that against algorithmic fairness. After this, we perform experiments in two scenarios: a “black-box” attack in which the attacker only has access to a set of data sampled from the same distribution as the original training data, but not the model nor the original training set, and a “white-box” scenario in which the attacker has full access to both. The effects of these attacks are measured using impact quantification metrics. The experiments show that by carefully perturbing a limited amount of training examples, an skilled attacker has the possibility of introducing different types of inequities for certain groups of individuals. This, can be done without large effects on the overall accuracy of the system, which makes these attacks harder to detect. To facilitate the reproducibility of the obtained results, the code generated for the experiments has been published in an open-source repository.¹.

¹<https://github.com/dsolanno/Poisoning-Attacks-on-Algorithmic-Fairness>

3.1.3 Chapter structure.

The rest of this chapter is organized as follows.

Section 3.2 provides further references to related work. Section 3.3, describes the proposed methodology to craft poisoning attacks for algorithmic fairness. Section 3.5 demonstrates empirically the feasibility of the new types of attacks on both synthetic and real-world data, under different scenarios depending on the attacker knowledge about the system. Section 3.6 presents our conclusions.3

3.2 Related Work

Adversarial Machine Learning Attacks. This work is based on Gradient-Based Optimization, an optimization framework widely used in the literature on Adversarial Machine Learning for crafting poisoning attacks [28, 140, 142, 105, 59]. Such framework is used to solve the bilevel optimization given by Eqs. (3.1)-(3.3), and requires computing the gradient of the classification function learned by the classifier. As a result, poisoning samples can be obtained by iteratively optimizing one attack point at a time [211].

Measuring Algorithmic Fairness. Many different ways of measuring algorithmic fairness have been proposed [144]. Among those that can be applied in an automatic classification context we find two main types: individual fairness metrics and group fairness metrics [91]. The former seek *consistency* in the sense that similar elements should be assigned similar labels [67]. The latter seek some form of *parity*, and in many cases can be computed from a contingency table indicating the number of privileged and unprivileged samples receiving a positive or negative outcome [153]. Popular group fairness metrics include disparate impact, equalized odds [93], and disparate mistreatment [215].

Optimization-Based Approaches to Increase Fairness. Algorithmic fairness can and often is compromised unintentionally, as discrimination in machine learning is often the result of training data reflecting discriminatory practices that may not be apparent initially [19]. When this is the case, training data can be modified by a type of poisoning attack, in which so-called “antidote” samples are added to a training set to reduce some measure of unfairness. One such approach proposes a method to be applied on recommender systems based on matrix factorization [165]; another is based in the Gradient-Based Optimization framework used in this work [118].

In addition to methods to mitigate unfairness by modifying training data (something known as a pre-processing method for algorithmic fairness [91]), other methods modify the learning algorithm itself to create, for instance, a fair classifier [217, 215] In these works, the trade-off between accuracy and fairness is

approached through an alternative definition of fairness based in covariance between the users sensitive attributes and the signed distance between the feature vectors of misclassified users and the classifier decision boundary.

3.3 Poisoning Fairness

In this section we present a novel gradient-based poisoning attack, crafted with the purpose of compromising algorithmic fairness, ideally without significantly degrading accuracy.

Notation. Feature and label spaces are denoted in the following with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \in \{-1, 1\}$, respectively, with d being the dimensionality of the feature space. We assume that the attacker is able to collect some training and validation data sets that will be used to craft the attack. We denote them as \mathcal{D}_{tr} and \mathcal{D}_{val} . Note that these sets include samples along with their labels. $L(\mathcal{D}_{val}, \theta)$ is used to denote the validation loss incurred by the classifier $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, parametrized by θ , on the validation set \mathcal{D}_{val} . $\mathcal{L}(\mathcal{D}_{tr}, \theta)$ is used to represent the regularized loss optimized by the classifier during training.

3.3.1 Attack Formulation

Using the aforementioned notation, we can formulate the optimal poisoning strategy in terms of the following bilevel optimization:

$$\max_{\mathbf{x}_c} \quad \mathcal{A}(\mathbf{x}_c, y_c) = L(\mathcal{D}_{val}, \theta^*) , \quad (3.1)$$

$$\text{s.t.} \quad \theta^* \in \arg \min_{\theta} \mathcal{L}(\mathcal{D}_{tr} \cup (\mathbf{x}_c, y_c), \theta) , \quad (3.2)$$

$$\mathbf{x}_{lb} \preceq \mathbf{x}_c \preceq \mathbf{x}_{ub} . \quad (3.3)$$

The goal of this attack is to maximize a loss function on a set of untainted (validation) samples, by optimizing the poisoning sample \mathbf{x}_c , as stated in the outer optimization problem (Eq. 3.1). To this end, the poisoning sample is labeled as y_c and added to the training set \mathcal{D}_{tr} used to learn the classifier in the inner optimization problem (Eq. 3.2). As one may note, the classifier θ^* is learned on the poisoned training data, and then used to compute the outer validation loss. This highlights that there is an implicit dependency of the outer loss on the poisoning point \mathbf{x}_c via the optimal parameters θ^* of the trained classifier. In other words, we can express the optimal parameters θ^* as a function of \mathbf{x}_c , i.e., $\theta^*(\mathbf{x}_c)$. This relationship tells us how the classifier parameters change when the poisoning point \mathbf{x}_c is perturbed. Characterizing and being able to manipulate this behavior is the key idea behind poisoning attacks.

Within this formulation, additional constraints on the feature representation of the poisoning sample can also be enforced, to make the attack samples stealthier or more difficult to detect. In this work we only consider a box constraint that requires the feature values of \mathbf{x}_c to lie within some lower and upper bounds (in Eq. 3.3, the operator \preceq enforces the constraint for each value of the feature vectors involved). This constraint allows us to craft poisoning samples that lie within the feature values observed in the training set. Additional constraints can be additionally considered, e.g., constraints imposing a maximum distance from an initial location or from samples of the same class, we leave their investigation to future work. Our goal here is to evaluate the extent to which a poisoning attack which is only barely constrained can compromise algorithmic fairness.

The bilevel optimization considered here optimizes one poisoning point at a time. To optimize multiple points, one may inject a set of properly-initialized attack points into the training set, and then iteratively optimize them one at a time. Proceeding on a greedy fashion, one can add and optimize one point at a time, sequentially. This strategy is typically faster but suboptimal (as each point is only optimized once, and may become suboptimal after injection and optimization of the subsequent points).

Attacking Algorithmic Fairness.

We now define an objective function $\mathcal{A}(\mathbf{x}_c, y_c)$ in terms of a validation loss $L(\mathcal{D}_{\text{val}}, \theta)$ that will allow us to compromise algorithmic fairness without significantly affecting classification accuracy. To this end, we consider the *disparate impact* criterion [20]. This criterion assumes data items, typically representing individuals, can be divided into unprivileged (e.g., people with a disability) and privileged (e.g., people without a disability), and that there is a positive outcome (e.g., being selected for a scholarship).

Although one might argue that there are several algorithmic fairness definitions [144] that could be used for this analysis, we selected this criterion for its particularity of being incorporated in legal texts in certain countries [72, 217]. Apart of that, recent studies [76] show how fairness metrics are correlated in three clusters what means that targeting this criterion will also affect a set of other metrics with similar strength. In addition to this, authors of [19] used this metric to illustrate the first of the three historical fairness goals that have been used to define fairness metrics. Disparate impact is observed when the fraction of unprivileged people obtaining the positive outcome is much lower the fraction of privileged people obtaining the positive outcome. Formally, to avoid disparate impact:

$$D = \frac{P(\hat{Y} = 1 | G = u)}{P(\hat{Y} = 1 | G = p)} \geq 1 - \epsilon, \quad (3.4)$$

where \hat{Y} is the predicted label, and $G = \{u, p\}$ a *protected attribute* denoting the group of unprivileged (u) and privileged (p) samples within a set \mathcal{D} . Disparate impact thus measures the ratio between the fractions of unprivileged and privileged samples that are assigned to the positive class. Typically, one sets $\epsilon \approx 0.2$ which suggests $D \geq 0.8$ for a fair classifier, as stated by the four-fifths rule of maximum acceptable disparate impact proposed by the US Equal Employment Opportunity Commission (EEOC) [72, 217]. Thus, in general, we should have D values closer to one to improve fairness.

For our poisoning attack to work, we aim to minimize such a ratio, i.e., decreasing the fraction of unprivileged samples for which $\hat{y} = 1$, while increasing the fraction of privileged users which are assigned $\hat{y} = 1$. For numerical convenience, we choose to maximize the difference (instead of the ratio) between the mean loss computed on the unprivileged and the privileged samples:

$$L(\mathcal{D}_{\text{val}}, \theta) = \underbrace{\sum_{k=1}^p \ell(\mathbf{x}_k, y_k, \theta)}_{\text{unprivileged}} + \lambda \underbrace{\sum_{j=1}^m \ell(\mathbf{x}_j, y_j, \theta)}_{\text{privileged}}. \quad (3.5)$$

Note that the parameter λ here is set to p/m to balance the class priors (rather than dividing the first term by p and the second by m).

To minimize D , we would like to have unprivileged samples classified as negative (lower numerator) and privileged classified as positive (higher denominator). As we aim to maximize $L(\mathcal{D}_{\text{val}}, \theta)$, we can label the unprivileged samples as positive ($y_k = 1$), and the privileged samples as negative ($y_j = -1$). Maximizing this loss will enforce the attack to increase the number of unprivileged samples classified as negative and of privileged samples classified as positive.

In Fig. 3.1, we report a comparison of the attacker’s loss $\mathcal{A}(\mathbf{x}_c, y_c) = L(\mathcal{D}_{\text{val}}, \theta^*)$ as given by Eq. (3.5) and the disparate impact D , as a function of the attack point \mathbf{x}_c (with $y_c = 1$) in a bi-dimensional toy example. Each point in the plot represents the value of the function (either \mathcal{A} or D computed on an untainted validation set) when the point \mathbf{x}_c corresponding to that location is added to the training set. These plots show that our loss function provides a nice smoother approximation of the disparate impact, and that maximizing it correctly amounts to minimizing disparate impact, thus compromising algorithmic fairness.

3.4 Gradient-Based Attack Algorithm

Having defining our (outer) objective, we are now in the position to discuss how to solve the given bilevel optimization problem. Since our objective is differentiable, we can make use of existing gradient-based strategies to tackle this problem. In

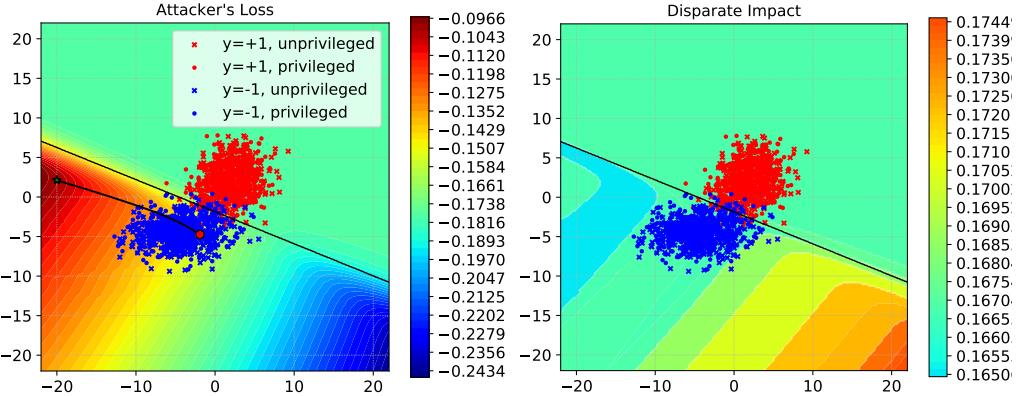


Figure 3.1: Attacker’s loss $\mathcal{A}(\mathbf{x}_c, y_c)$ (left) and disparate impact (right) as a function of the attack point \mathbf{x}_c with $y_c = 1$, on a bi-dimensional classification task. Note how the attacker’s loss provides a smoother approximation of the disparate impact, and how our gradient-based attack successfully optimizes the former, which amounts to minimizing disparate impact, compromising algorithmic fairness.

particular, we will use a simple gradient ascent strategy with projection (to enforce the box constraint of Eq. 3.3). The complete algorithm is given as Algorithm 1. In Fig. 3.1 we also report an example of how this algorithm is able to find a poisoning point that maximizes the attacker’s loss.

Attack Initialization. An important remark to be made here is that *initialization* of the poisoning samples plays a key role. In particular, if we initialize the attack point as a point which is correctly classified by the algorithm, the attack will not even probably start at all. This is clear if one looks at Fig. 3.1, where we consider an attack point labeled as positive (red). If we had initialized the point in the top-right area of the figure, where positive (red) points are correctly classified, the point would have not even moved from its initial location, as the gradient in that region is essentially zero (the value of the objective is constant). Hence, for a poisoning attack to be optimized properly, a recommended strategy is to initialize points by sampling from the available set at random, but then flipping their label. This reduces the risk of starting from a flat region with null gradients [28, 211].

Gradient Computation. Despite the simplicity of the given projected gradient-ascent algorithm, the computation of the poisoning gradient $\nabla_{\mathbf{x}_c} \mathcal{A}$ is more complicated. In particular, we do not only need the outer objective to be sufficiently smooth w.r.t. the classification function, but also the solution θ^* of the inner optimization to vary smoothly with respect to \mathbf{x}_c [28, 142, 59, 29]. In general, we need \mathcal{A} to be sufficiently smooth w.r.t. \mathbf{x}_c .

Under this assumption, the gradient can be obtained as follows. First, we

Algorithm 1 Gradient-based poisoning attack

Require: \mathbf{x}_c, y_c : the initial location of the poisoning sample and its label; η : the gradient step size; $t > 0$: a small number.

Ensure: \mathbf{x}'_c : the optimized poisoning sample.

- 1: Initialize the attack sample: $\mathbf{x}'_c \leftarrow \mathbf{x}_c$
 - 2: **repeat**
 - 3: Store attack from previous iteration: $\mathbf{x}_c \leftarrow \mathbf{x}'_c$
 - 4: Update step: $\mathbf{x}'_c \leftarrow \Pi(\mathbf{x}_c + \eta \nabla_{\mathbf{x}_c} \mathcal{A})$, where Π ensures projection onto the feasible domain (i.e., the box constraint in Eq. 3.3).
 - 5: **until** $|\mathcal{A}(\mathbf{x}'_c, y_c) - \mathcal{A}(\mathbf{x}_c, y_c)| \leq t$
 - 6: **return** \mathbf{x}'_c
-

derive the objective function w.r.t. \mathbf{x}_c using the chain rule [28, 211, 142, 29, 140]:

$$\nabla_{\mathbf{x}_c} \mathcal{A} = \nabla_{\mathbf{x}_c} L + \frac{\partial \theta^*}{\partial \mathbf{x}_c}^\top \nabla_\theta L, \quad (3.6)$$

where the term $\frac{\partial \theta^*}{\partial \mathbf{x}_c}$ captures the implicit dependency of the parameters θ on the poisoning point \mathbf{x} , and $\nabla_{\mathbf{x}_c} L$ is the explicit derivative of the outer validation loss w.r.t. \mathbf{x}_c . Typically, this is zero if \mathbf{x}_c is not directly involved in the computation of the classification function f , e.g., if a linear classifier is used (for which $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$). In the case of kernelized SVMs, instead, there is also an explicit dependency of L on \mathbf{x}_c , since it appears in the computation of the classification function f when it joins the set of its support vectors (see, e.g., [28, 59]).

Under regularity of $\theta^*(\mathbf{x}_c)$, the derivative $\frac{\partial \theta^*}{\partial \mathbf{x}_c}$ can be computed by replacing the inner optimization problem in Eq. (3.2) with its equilibrium (Karush-Kuhn-Tucker, KKT) conditions, i.e., with the implicit equation $\nabla_\theta \mathcal{L}(\mathcal{D}_{\text{tr}} \cup (\mathbf{x}_c, y_c), \theta) \in \mathbf{0}$ [140, 142]. By deriving this expression w.r.t. \mathbf{x}_c , we get a linear system of equations, expressed in matrix form as $\nabla_{\mathbf{x}_c} \nabla_\theta \mathcal{L} + \frac{\partial \theta^*}{\partial \mathbf{x}}^\top \nabla_{\mathbf{w}}^2 \mathcal{L} \in \mathbf{0}$. We can now compute $\frac{\partial \theta^*}{\partial \mathbf{x}_c}$ from these equations, and substitute the result in Eq. (3.6), obtaining the required gradient:

$$\nabla_{\mathbf{x}_c} \mathcal{A} = \nabla_{\mathbf{x}_c} L - (\nabla_{\mathbf{x}_c} \nabla_\theta \mathcal{L})(\nabla_\theta^2 \mathcal{L})^{-1} \nabla_\theta L. \quad (3.7)$$

These gradients can be computed for various classifiers (see, e.g., [59]). In our case, we simply need to compute the term $\nabla_\theta L$, to account for the specific validation loss that we use to compromise algorithmic fairness (Eq. 3.5).

Finally, in Fig. 3.2, we show how our poisoning attack modifies the decision function of a linear classifier to worsen algorithmic fairness on a simple bi-dimensional example. As one may appreciate, the boundary is slightly tilted, causing more unprivileged samples to be classified as negative, and more privileged samples to be classified as positive.

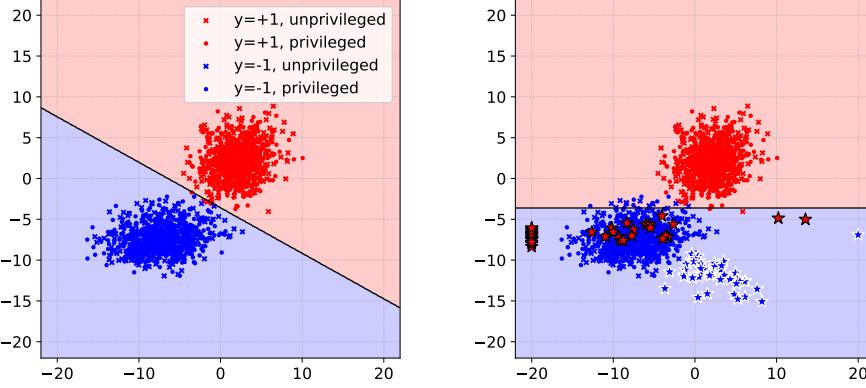


Figure 3.2: Gradient-based poisoning attack against a logistic classifier, on a bi-dimensional classification task. The classification function and the corresponding decision regions are reported before (*left*) and after (*right*) injection of the poisoning samples (red and blue stars in the right plot).

3.4.1 White-Box and Black-Box Poisoning Attacks

The attack derivation and implementation discussed throughout this section implicitly assumes that the attacker has full knowledge of the attacked system, including the training data, the feature representation, and the learning and classification algorithms. This sort of *white-box* access to the targeted system is indeed required to compute the poisoning gradients correctly and run the poisoning attack [29]. It is however possible to also craft *black-box* attacks against different classifiers by using essentially the same algorithm. To this end, one needs to craft the attacks against a *surrogate model*, and then check if these attack samples *transfer* successfully to the actual target model. Interestingly, in many cases these black-box transfer attacks have been shown to work effectively, provided that the surrogate model is sufficiently similar to the target ones [149, 59]. The underlying assumption here is that it is possible to train the surrogate model on samples drawn from the same distribution as those used by the target model, or that sufficient queries can be sent to the target model to reconstruct its behavior.

In our experiments we consider both white-box attacks and black-box transfer attacks to also evaluate the threat of poisoning fairness against weaker attackers that only possess limited knowledge of the target model. For black-box attacks, in particular, we assume that the attacker trains the substitute models on a training set sampled from the same distribution as that of the target models, but no queries are sent to the target classifiers while optimizing the attack.

3.5 Experiments

This section describes the obtained results for two different datasets, one synthetic set composed of 2000 samples, each of them having three features, one of them considered the sensitive attribute, not used for the optimization. The second dataset corresponds to one of the most widely used by the *Algorithmic Fairness* community, a criminal recidivism prediction dataset composed by more than 6000 samples, with 18 features describing each individuals. For each dataset, we consider both the white-box and the black-box attack scenarios described in Section 3.4.1.

3.5.1 Experiments with synthetic data

The first round of experiments uses synthetic data set to empirically test the impact of the attacks with respect to varying levels of disparity already found in the (unaltered) training data. Data is generated using the same approach of Zafar et al. [217]. Specifically, we generate 2,000 samples and assign them to binary class labels ($y = +1$ or $y = -1$) uniformly at random. Each sample is represented by a 2-dimensional feature vector created by drawing samples from two different Gaussian distributions: $p(x|y = +1) \sim N([2; 2], [5, 1; 1, 5])$ and $p(x|y = -1) \sim N([\mu_1; \mu_2], [10, 1; 1, 3])$ where μ_1, μ_2 are used to modify the euclidean distance S between the centroids of the distributions for the privileged and unprivileged groups so that different base rates [52] can be tested in the experiments. Then, a sample's sensitive attribute z is assigned by drawing from a Bernoulli distribution using $p(z = +1) = \frac{p(x'|y=+1)}{p(x'|y=+1)+p(x'|y=-1)}$ where $x' = [\cos(\phi) - \sin(\phi); \sin(\phi), \cos(\phi)]x$ corresponds to a rotated version of the feature vector x .

Using the generator we have described, datasets such as the ones as depicted in Figure 3.3 can be obtained. In this figure, the feature vector x is represented in the horizontal and vertical axes, while the color represents the assigned label y (green means favorable, red means unfavorable) and the symbol the sensitive attribute z (circle means privileged, cross means unprivileged).

We generate multiple datasets by setting $S \in \{0, 1, 2, \dots, 9\}$. We then split each dataset into training D_{tr} (50% of the samples), validation D_{val} (30%) and testing D_{test} (20%) subsets. In each run, a base or initial model \mathcal{M} is trained. This model \mathcal{M} corresponds to a Logistic Regression model in the first setting and to a Support Vector Machine with linear kernel in the second scenario. The regularization parameter C is automatically selected between $[0.5, 1, 5, 10]$ through cross validation. In the *White-Box* setting, the attack is optimized for \mathcal{M} so that Eq. 3.1 is minimized in the training set D_{tr} and Eq. 3.3 is maximized in the validation set D_{val} . In the *Black-Box* setting, the attack is optimized against a surrogate model

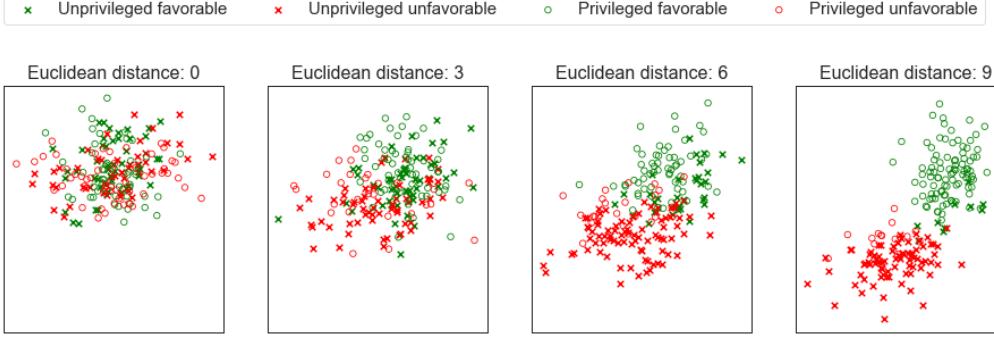


Figure 3.3: (Best seen in color.) Examples of generated synthetic data sets for different values of the separation S between groups. Privileged elements ($z = +1$) are denoted by circles and unprivileged elements ($z = -1$) by crosses. Favorable labels ($y = +1$) are in green, while unfavorable labels ($y = -1$) are in red.

$\hat{\mathcal{M}}$, a Logistic Regression classifier, trained with another subset of data generated for the same value of the parameter S . Each of these attacks generates a number of poisoning samples. The poisoned model is the result of retraining the original model with a training set that is the union of D_{tr} and the poisoned samples.

The attack performance is measured by comparing the model trained on the original training data with a model trained on the poisoned data. The evaluation is done according to the following metrics, which for each dataset are averaged over ten runs of each attack:

- **Accuracy** The accuracy on test obtained by the poisoned model is similar and correlated with the accuracy obtained by a model trained on the original data. It is important to note that the separability of the generated data is also highly correlated with the separation between the groups in the data, creating this effect.
- **Demographic parity** Measures the allocation of positive and negative classes across the population groups. Framed within the Disparate impact criteria that aims to equalize assigned outcomes across groups, this metric is formulated as:

$$P(\hat{Y} = 1|G = \text{unprivileged}) - P(\hat{Y} = 1|G = \text{privileged})$$

It tends to zero in a fair scenario and is bounded between $[1, -1]$ being -1 the most unfair setting. This metric is correlated with the *Disparate impact* metric introduced in Section 3.3 and has been selected for convenience in the visual representation of the results.

- **Average odds difference** The average odds difference is a metric of disparate mistreatment, that attempts for Equalized odds [93], it accounts for differences in the performance of the model across groups. This metric is formulated as:

$$\frac{1}{2}[(FPR_p - FPR_u) + (TPR_p - TPR_u)]$$

It gets value zero in a fair scenario and is bounded between $[1, -1]$ being -1 the most unfair setting.

- **FNR privileged** False Negative Rate for the privileged group of samples.
- **FNR unprivileged** False Negative Rate for the unprivileged group of samples.
- **FPR privileged** False Positive Rate for the unprivileged group of samples.
- **FPR unprivileged** False Positive Rate for the unprivileged group of samples.

Results shown on Figure 3.4 show the obtained performance of the attacks for the generated data. In this figure, the horizontal axis is the separation S between classes in each of the ten datasets. Analyzing the results, we observe that the poisoned models increase disparities in comparison with a model created on the unaltered input data, across all settings. Additionally, they yield an increased FPR for the privileged group (privileged samples that actually have an unfavorable outcome are predicted as having a favorable one), increasing significantly the observed unfairness as measured by the fairness measurements. We note that the attacks also decrease the FNR of the unprivileged group (unprivileged samples that actually have a favorable outcome are predicted as having an unfavorable one). This is most likely a consequence of the attack’s objective of maintaining accuracy and show that this attack is not trivial. If the attack were only to increase disparities, it would also increase the FNR of the unprivileged group with a larger decrease in accuracy than what we observe. The decrease of FNR for the unprivileged group, however, is smaller than the increase of FPR for the privileged group, as the average odds difference plot shows, and hence the attack succeeds.

3.5.2 Experiments with real data

To demonstrate the attacks on real data, we use the COMPAS dataset released by ProPublica researchers [109], which is commonly used by researchers on Algorithmic Fairness. This dataset contains a prediction of criminal recidivism based

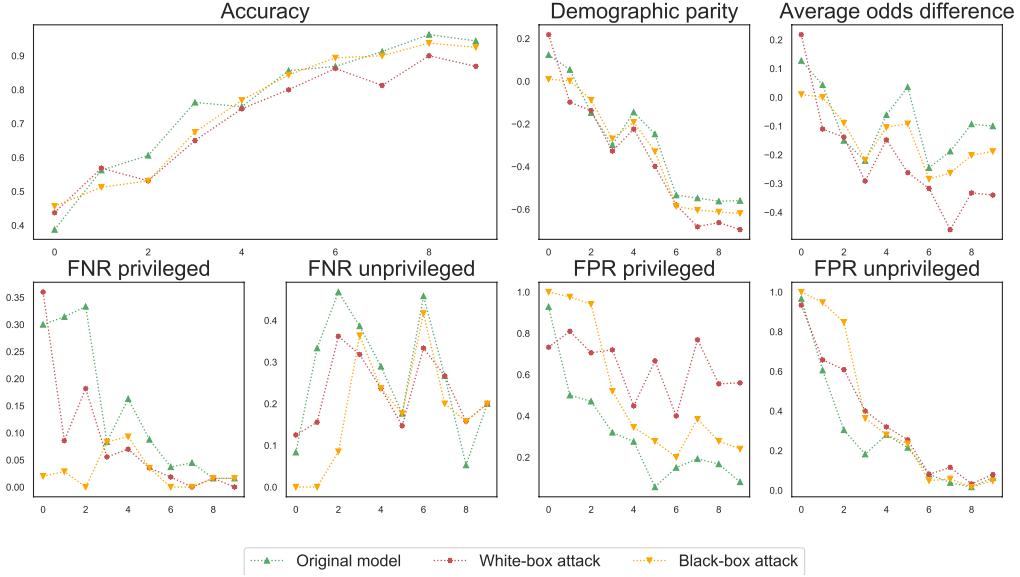


Figure 3.4: Comparison of the original model against the model generated by the White-box attack and Black-box attacks, for ten synthetic datasets generated by different separation parameters (S). Each data point is the average of ten runs of an attack. We observe that attacks have a moderate effect on the accuracy of the classifier, and can affect the classifier fairness (demographic parity and odds difference) to an extent that becomes more pronounced if the original dataset already has a large separation between classes (larger values of S).

on a series of attributes for a sample of 6,167 offenders in prison in Broward County, Florida, in the US. The attributes for each inmate include criminal history features such as the number of juvenile felonies and the charge degree of the current arrest, along with sensitive attributes: race and gender. For each individual, the outcome label (“recidivism”) is a binary variable indicating whether he or she was rearrested for a new crime within two years of being released from jail.

We use this dataset for two different types of experiments. First, we show how the attacks demonstrated on synthetic data can also be applied to this data, and demonstrate the effect of varying the amount of poisoned samples. Second, we evaluate the transferability of the attack to other classification models.

White-Box and Black-Box poisoning attacks with varying amounts of poisoned samples. This experiment compares the original model against the model obtained under the two attack models.

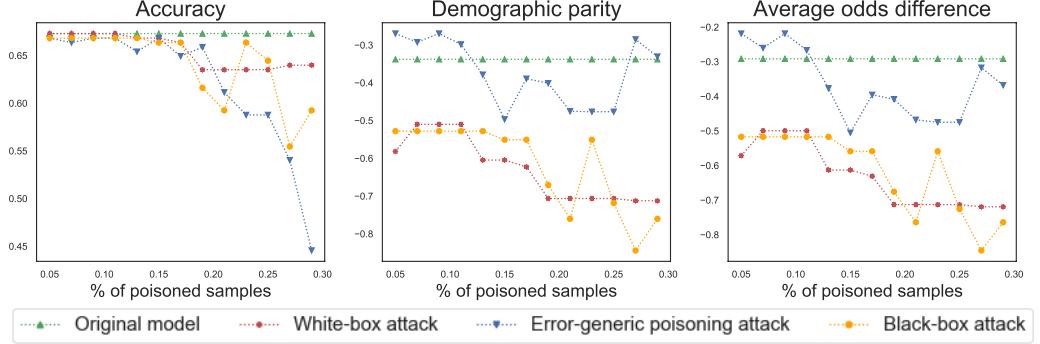


Figure 3.5: Comparison of the original model against the model generated by a White-box attack and a Black-box attack, for varying percentages of poisoned samples. The main difference between both types of attack is that the black-box attack starts having more noisy behaviour also drastically reducing the accuracy of the classifier (thus being more easily detectable) when the percentage of poisoned samples exceeds a certain threshold (about 20%).

Figure 3.5 shows the results, which are in line with the findings of the experiments on synthetic data. According to the obtained results, both types of poisoning attacks are able to increase unfairness of the model with a more modest effect on the accuracy. Also, an interesting finding is the stability of the *White-Box* attack as opposite to the *Black-Box* attack. Whereas the first keeps the same trend with the growing number of samples, the later starts having a unstable and noisy behaviour after adding the 20% of samples, causing for some cases a more unfair model but also affecting the accuracy of the system in a manner that could be easily detected.

In Figure 3.5 we also include an Error-Generic Poisoning Attack [59] for the Logistic Regression model , which is designed to decrease the accuracy of the resulting model. We observe that this type of generic adversarial machine learning attack does not affect the fairness of the classifier nearly as much as the attacks we have described on this paper.

As expected, computing the obtained performance for all the stated metrics, (Figure omitted for brevity) can be observed that the effect of any attack increases with the number of poisoned samples. In general, these attacks increase the False Negatives Rate (FNR) for the unprivileged samples, and increase the False Positives Rate (FPR) for the privileged samples.

Transferability of the attack. We study how an attack would affect the performance of other type of models, simulating different scenarios of *Zero Knowledge* attacks.

Specifically, the attacks we perform is optimized for a Logistic Regression

model, and its performance is tested for other models: (a) Gaussian Naive Bayes. (b) Decision Tree; (c) Random Forest; (d) Support Vector Machine with linear kernel; and (e) Support Vector Machine with Radial Basis Function (RBF) kernel.

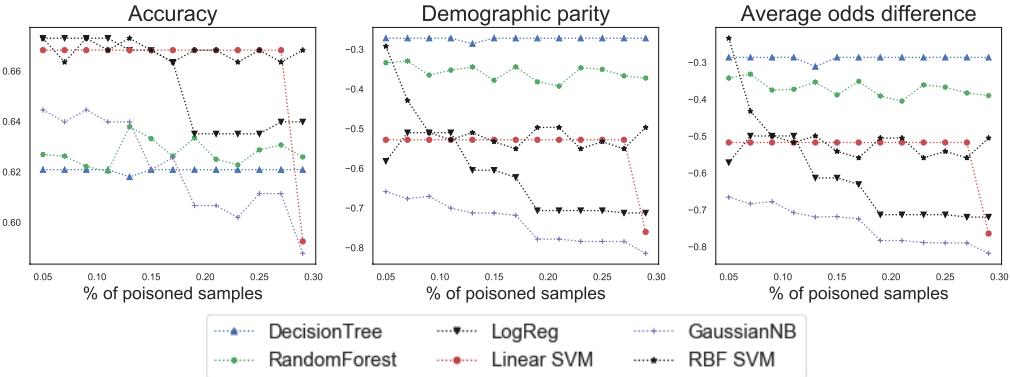


Figure 3.6: Transferability of the attacks from Logistic Regression to other models.

Results are shown on Figure 3.6, in which each data point corresponds to the average of five experimental runs. We observe that the attack optimized on a Logistic Regression classifier has a stronger effect on the Logistic Regression, Support Vector Machine (for both types of kernel tested) and Naive Bayes models. In contrast, while it can introduce unfairness through demographic disparity and average odds difference on a Decision Tree or Random Forest classifier, its effects are more limited.

3.6 Conclusions

The results show the feasibility of a new kind of adversarial attack crafted with the objective of increasing disparate impact and disparate mistreatment at the level of the system predictions. We have demonstrated an attacker effectively alter the Algorithmic Fairness properties of a model even if pre-existing disparities are present in the training data. This means that these attacks can be used to both introduce algorithmic unfairness, as well as for increasing it where it already exists. This can be done even without access to the specific model being used, as a surrogate model can be used to mount a black-box transfer attack.

Chapter 4

PROFESSIONAL BIAS

4.1 Introduction

The term Artificial Intelligence (AI) describes a broad concept related to the ability of machines to carry out tasks in a way that might be perceived as “smart”. Machine Learning (ML) constitutes a subfield of AI that studies algorithms that improve automatically through experience and have been used to infer meaning, generalise and learn patterns from data and thus discover “knowledge” that was not explicitly programmed by the creator.

In recent years, the AI domain has experienced an impressive growth.¹

4.1.1 Motivation

The main objective of this paper is to analyze advancements in and discussions around AI from a social sciences’ point of view. From this perspective we examine which ‘conventions’ or moral orders are employed during the creation of these models based on dialogue and justifications between individual(s) and the collective. By studying the research on, the design of, the development of and the public opinion on AI related systems, we focus on the interim process reasoned to be a key contributor to the subsequent interactions between humans and machines. To this end, we employ the Economics of Convention (EC) – a general social science theory – which proposes a pragmatic perspective to study coordination and conflicts, analyzing the underlying justifications and conventions. Through the theoretical lens of the EC, we analyze how distinct moral registers represented by conventions within the EC are reflected in this domain. Having a better understanding of the conventions guiding the perceptions and advancements in the

¹<https://www.forbes.com/sites/louiscolumbus/2018/01/12/10-charts-that-will-change-your-perspective-on-artificial-intelligences-growth/>

field of AI is considered to be a necessary preliminary step to a) understand the conventions reflected by these autonomous systems in their interactions with societies thereafter and b) shed light on ongoing conflicts around transparency or human vs. AI.

The Economics of Convention (EC) provide the framework for this study, which are described in detail in the first part of this paper. For the analysis of conventions, we create a real-world text dataset with subsets from three different text sources and examine the distribution of conventions in these subsets. We use an iterative training process based on active learning as proposed in [180] to build a supervised ML model with one binary classifier per convention and show results for each convention. The dataset along with the code is released to the research community.²

4.1.2 Contributions

This work employs the theoretical framework of the EC to study written dialogues and research abstracts in 1) AI software design and development, 2) AI research and 3) social discussions about AI. Either researchers describe their findings to different communities (GitHub, Semantic Scholar (S2)) or AI is discussed in a community (Reddit). We aim to reveal the conventions, which these different communities follow. We assume that documents in open-source ML and AI software repositories, and the conversations within, reflect the conventions guiding decisions taken during the AI development phase. Research articles in the domain of ML and AI describe findings to the research community, and as such should reflect the conventions followed by scientists working in the field of AI research and design. For discussions in online forums where individuals with varied levels of expertise on the topic of AI exchange information and discuss recent advancements in the field, we assume a broader and more general use of conventions.

4.1.3 Chapter structure

This chapter is structured as follows: First, we provide in Section 4.2 an overview of the related work in relevant areas closely related to the work in this paper. After that, the theoretical framework of our analysis is described in Section 4.3. The next section (Section 4.4) provides an overview of the creation of the dataset and the different subsets in Section 4.5, where we also outline the architecture to train the ML models. In the subsequent section, we describe the results of the analysis of the dataset as we evaluate the performance of our classifiers and analyze the use of conventions in the different subsets of our dataset in Section 4.6. Finally,

²<https://github.com/dsolanno/AIVC>

Section 4.7 discuss limitations of our approach and 4.8 summarizes our work in a conclusion and .

4.2 Related Work

Let us start off by providing an overview of the state of the art on the EC field. Research efforts focused in the analysis of each of the data sources considered in this work are summarized.

4.2.1 Economics of Convention (EC)

Although there is a large body of literature on understanding the motivation of open source software developers, none of them examines the use of the EC. [103] apply the EC in order to explain inter-organizational relationships in the coordination process of platform-based multi-sourcing in the general context of software development. Non-technical approaches such as [60], [82] or [116] use the EC to explain the coordination of pluralism and contradictory strategies in organizations. Replacing the term “Economics of Convention” with “motivation” leads to additional results in the domain of software development. Especially in open source software development, several studies focus on motivation [96, 170]. Accordingly, previous research identifies five primary categories of motifs [36]:

- **Intrinsic motivation**, i.e., fun or self-efficacy [173].
- **External rewards**, i.e., monetary incentives or career opportunities [122].
- **Ideology**, i.e., altruism [190].
- **Community recognition**, i.e., fame or reputation [145].
- **Learning**, i.e., development of personal skills or knowledge [207].

However, these categories only partially relate to the EC, as the EC shifts the research perspective; the above mentioned along with most previous works rely on agent-based approaches, which focus on the agents or actors, while the EC studies situations, in which agents, objects, technologies, etc. interact.

4.2.2 Content Analysis of Open Source Projects

GitHub has been widely studied as a source of information for software development projects. Most of the existing contributions based on the analysis of open source project content fall under the following four categories: **user analysis**,

programming language prevalence, project quality analysis and project evolution predictability. Due to the vast amount of studies on open source project content, this review is limited to contributions which are closely related to the work described in this paper.

Besides technical approaches, previous work on the study of project content often applies mathematical and statistical modelling to understand behaviour [51]. This approach is also sometimes combined with qualitative studies based on automated processes. [181] combine automated topic extraction with manual validation to categorise GitHub repositories based on the content of README files. Furthermore, [94] propose the use of both qualitative and quantitative approaches to automatically detect instructions for software development in project description files.

Apart from these efforts, [158] automatically structure the content of GitHub README files. In order to do so, they combine manual annotation with automated text classification approaches. [219] perform a qualitative analysis of software projects related to scientific articles in the field of AI in work which analyzes content specifically related to ML and/or AI in GitHub. Although there are studies on the content of GitHub project description files, these studies have different objectives. In contrast, our work proposes for the first time the categorisation of AI and ML related projects based on the content of the README file according to the EC paradigm.

4.2.3 Content Analysis of Scientific Articles

Although there is indeed much work in quantitative analysis on scientific articles, this body of work is mainly focused around the extraction of various entity and relation types such as named entities [12], co-references [90] and semantic roles [95]. Accordingly, previous work analysing Semantic Scholar (S2) focuses on those types [131]. Although there is work on the identification of patterns within the research community, this work is concerned with structural analysis such as citations and gender and not with discourse patterns [205]. In recent work on language modeling in scientific texts, [22] report state of the art results on several standard NLP tasks. However, such a model is generally not directly feasible for convention classification as this complex task requires in depth control of the iterative labeling and classification process.

4.2.4 Content Analysis of Online Discussions

Online forums and discussion sites are widely used to study social interaction. Different research communities study a variety of aspects such as the evolution and predictability of interactions in general [83] and popular posts in particular

[55], [39] study the evolution of user communities and social roles. [24] and [92] focus on the reliability and correctness of the information.

[135] perform a sentiment analysis of public perception of AI for expert and non expert groups of users on Twitter and [108] compare opinions of the public and media on robots and autonomous systems. [71] study the evolution of media perception of AI, and [134] study privacy concerns of users about intelligent assistants by performing a survey and analysing public reviews. While [57] study inter-community conflicts and common patterns, they define the conflicts as anti-social behaviour and do not consider the EC theory or other types of conflicts.

All this work proves that online social sites are valuable sources of knowledge for the understanding of social behaviours and opinions. Along this line, our work enhances the understanding of society's perception of AI through the EC framework.

4.3 Economics of Convention (EC)

The main focus of our work lies at the intersection of the EC theory and the research, design, development and public opinions of AI-related systems. The EC, as a general social science theory developed by [34], proposes consistent pragmatic and situative concepts for the sociological analysis of behavioral coordination. It relies on justifications observed during ordinary disputes. This framework of justification is conceived as a theoretical research lens to empirically study co-operation and conflicts. In conflict situations, human actors mobilise arguments to defend their perspective. Based on field surveys and Western political philosophy, Boltanski and Thévenot develop a taxonomy of various conventions, or registers, of the so called “common good” the actors mobilize. The common good – or the benefit or interests of all – directly refers to specific perceptions of justice and fairness [34, 61]. Hence, (potential) conflicts arise when a view of the common good that is based on one principle of justification is criticised according to criteria which underlie another principle of justification. This theoretical approach has been already used in many different fields, e.g. the production of consumer goods [191, 32] and health [56, 182, 21]. It is found to be useful for gaining more insight into what is at stake in emerging conflicts. [34] identify six justification registers, each based on different philosophical foundations in Western liberal societies and conceptions of justice and what is fair: **Civic, Industrial, Market, Domestic, Inspired, and Renowned**. [33] and [119] expand it with two more registers: the **Project** and the **Green** register. [182] introduce a further **Vitalist** register based on the ‘googlization of health research’.

Table 4.1 provides an overview of each of these registers with their principles of justification. It shows that there is a plurality of possible conventions or regis-

Convention	Common good	Values
Industrial	Increased efficiency	Functionality, expertise, optimization
Project	Innovation and the network	Activity, experimentation, connection
Market	Economic growth	Competition, consumer choice, profit
Inspired	Inspiration	Spontaneity, deliberation, emotion
Civic	Collective will	Inclusivity, solidarity, equality
Domestic	Tradition	Hierarchy, trust
Green	Protection of environment	Environmental activism
Renown	Public opinion	Popularity, fame

Table 4.1: Registers of worth in the Economics of Convention

ters. The EC defines a ‘convention’ or ‘register’ not merely as a habit or custom [197, 34]; the concept of conventions in the EC is more complex. Conventions and registers form interpretative frameworks which actors develop and manage to evaluate and coordinate ‘action situations’ [62]. However, this does not imply that each individual is part of a particular convention, or that individuals consciously act according to the precepts of any of these mentioned [56]. On the contrary, depending on interactions with others, actors can easily pass ‘from one convention to another’ [56]. Similarly, the justifications for each of the actor’s activities are implicit; individuals only make them explicit in a conflict. Coordination of these conflicts requires either agreement on a common principle or that the actors find a common understanding, which can then emerge between different registers of justification. All conventions refer to a legitimate and immeasurable conception of the collective so that no convention is more rational than any other. The decision for a certain convention or register is not merely a matter of calculation but a choice between several possible common traits the actors share in their interactions [61]. Each register or convention acts as a logical, harmonious order of statements, objects and people that provide a general sense of justice. Hence, the typology of [34] offers an applicable framework to identify the conventions, which guide researchers, developers and their moral orientations in the field of AI.

Convention	Top keywords
Industrial	Performance, standard, tests, learning, reliable
Project	City, projective, connections, links, networks
Market	Customized, goods, license, sell, billion
Inspired	Inspiration, inspired, visual, passion, method
Renown	Opinion, press, fame, audience, influence
Civic	Collective, civic, interests, license, children
Domestic	Superiors, upbringing, trust, dependence, origin
Green	Green, economy, growth, carbon, sustainable

Table 4.2: A combination of the top five keywords in the dataset per convention established by manual analysis and TF-IDF frequency

4.4 The EC Dataset

The dataset contains subsets from three main data sources: Semantic Scholar (S2) research chapter abstracts³, GitHub README files⁴ and Reddit forums⁵.

To pre-filter documents we use a combination of two sets of keywords: First, we use a keywords list manually created by domain experts, including one of the authors and based on the registers introduced in Table 4.1. Second, we perform keyword matching after a first iteration of labeling based on ‘Term Frequency-Inverse Document Frequency’ (TF-IDF) [177] to extract keywords that are more common for each convention and not so common for the rest. Table 4.2 shows the five most important (of more than 30) keywords for each convention.

4.4.1 GitHub

GitHub is a web-based interface and cloud-based service that provides tools to effectively store and manage code in addition to tracking and controlling changes in the code base. GitHub stores the code and metadata of more than 100 million projects with involvement from more than 31 million developers.⁶ More than 8,500 projects related to AI topics are collected using the official GitHub API. We collect the content of the README file along with creation and last update timestamps in addition to statistics about the popularity of a repository. To avoid bias, repositories from all different levels of popularity (measured with the GitHub star rating) are gathered. In order to compare the use of conventions in GitHub AI

³<https://semanticscholar.org>

⁴<https://github.com>

⁵<https://reddit.com/>

⁶<https://github.blog/2018-11-08-100m-repos>

Data source	Sentences	Items
GitHub AI	127,236	8,609 repositories
GitHub non-AI	71,706	5,358 repositories
S2 AI	22,742	2,954 abstracts
S2 non-AI	69,694	5,970 abstracts
Reddit AI	38,296	2,455 threads
Reddit non-AI	219,916	3,875 threads
Total size	549,590	29,221

Table 4.3: Counts of sentences and items for AI and non-AI subsets from each data source. Depending on the specific data source, items refer to repositories, abstracts or threads.

related repositories with those in non-AI related repositories, data from an equivalent number of repositories similar to AI related topics is collected. Similarity is calculated on the basis of the number of stars. Table 4.3 shows the no. of sentences and the no. of repositories in the GitHub subset.

4.4.2 Semantic Scholar (S2)

Semantic Scholar (S2) is a search engine for peer-reviewed articles, which provides an open research corpus with more than 40 million chapters from computer science and bio-medicine in machine readable JSON format [8]. For the analysis of the conventions, we select a sample of entries that appear in one of the AI conferences listed in [112] and which are published after the year 2016. This list helps us to analyze the use of conventions in different sub-fields of AI, such as robotics, computer vision and natural language processing. We only select publications from 2016 onward because during this time, research in AI and applications of ML in particular received a significant boost with the release of TensorFlow [1]. This sample is further narrowed down by pre-filtering documents with the help of a list of keywords that belong to either of the registers in Table 4.1. Table 4.2 shows some of the most important keywords from this list.

4.4.3 Reddit

Reddit is a website centered around social news, web content rating, and discussion. Communities are named ‘subreddits’ and created around topics. We collect different threads from ML and AI ‘subreddits’. In detail, the text from the title of post which starts a thread, its body and the first level answers are collected by using the Reddit API. Samples from the AI domain are collected from a ‘ subreddit’

called ‘r/artificial’, whereas the non-AI examples were gathered from a variety of ‘subreddits’ related to the computer science field: ‘Javascript’, ‘DataBase’, ‘Python’, ‘Android’. We only use threads with a minimum of 4 upvotes (positive votes by readers from the community) to ensure that only relevant threads are considered in the analysis.

4.5 Methods for Building the EC Model

In order to build an EC ML model and analyse the predictions on our dataset, we define the EC classification as a multi-label task whereby each sentence in our dataset may have multiple associated conventions and hence multiple labels.

To the best of our knowledge, this is the first attempt to build a text-based EC classifier and no existing datasets can be used to train such an ML classifier. We regard the creation of a dataset for this purpose as a valuable contribution to the scientific community. Due to the complexity of the EC theory, the labeling of the dataset facilitated by the authors of this chapter was a time consuming task necessitating expertise and care. To optimize the labeling effort we use an active learning approach [180] focused on the labelling of items most beneficial to the training of the models. The quality of the predictions are thus incrementally improved while at the same time new samples are labeled to train successive versions of the classifiers.

4.5.1 Model Selection

The EC model should cover the following:

- Support multi-label classification, where one sentence can have multiple labels and the number of labels per sentence is not fixed.
- Support multi-class classification, where sentences can belong to 1 out of multiple categories

To this end, the classifiers are trained using a strategy commonly known as one vs. rest (or one vs. all) [168]. This strategy involves the training of one binary classifier per class (i.e. convention) to model a multi-class problem.

As such, the eight binary class-labels show multiple classes per item (i.e. sentence) along with a confidence score between 0 and 1 for each predicted label. This in effect represents a multi-label architecture because one item can belong to multiple classes (i.e. one sentence can belong to more than one convention). We decompose a multi-label, multi-class problem into a set of binary classifiers.

The upside of the one vs. all strategy is that it enables classifier calibration in terms of precision. Selecting a classification threshold with equal levels of precision for all classifiers allows a balanced comparison of the results from the different classifiers. A classifier only outputs a positive label when this threshold is exceeded, otherwise the label is negative. Furthermore, the architecture based on classifiers that are combined into one big model facilitates the building and testing of individual convention classifiers which offers individual performance checks. This lightweight approach also eases the data handling process in the active learning scenario.

We use convolutional neural network (CNN) classifiers following the architecture proposed by [113] with the standard parameters. The network uses an input sequence of 32 vectors per sample to represent a sentence, where each of the vectors is encoded with a 100-dimensional word embedding vector. The network is composed of 14 layers, four of them convolutional layers, with over 10 mio. parameters of which $\sim 300k$ are trainable. It uses *categorical cross entropy* as loss and a *relu* activation function for the hidden layers.

Accordingly, one individual classifier C_c is trained per convention C . Given a sentence S , the classifier C_c is trained such that it assigns a probability score P for that sentence being part of the convention C . Therefore: $C_c(S, C) = P$ where $P = [0, 1]$. A combination of $N = 8$ binary classifiers (one per convention) predicts the probability of an item (sentence) to belong to each possible class label (convention). We set the calibration threshold to 0.9 precision during training to ensure meaningful labels. We classify conventions on sentence level because sentences correspond to the minimal units which reflect conventions in text.

As a ML classifier requires data input in the form of numeric values rather than continuous or discrete variables, a method to numerically represent the training text in the form of a vector is required. The most common approach to date to solve this problem is the use of word embeddings. Words are transformed into n-dimensional vector representations and projected into a new multidimensional space. The contextual relationship of words with similar context is reflected in the n-dimensional space by distance (e.g. similar words are close to one another). To this end we use pre-trained GloVe word embeddings [154] for the vector representation of words in this n-dimensional space.

4.5.2 Labeling of Dataset and Active Learning

Due to the complexity of the EC, labeling the dataset demands both time and expertise. That is why an active learning model with a focus on uncertainty sampling is implemented. Uncertainty sampling prioritizes correctly labelling items based on classifier confidence. One objective is to enhance the training data by correctly labelling items that are classified with a low confidence score below 0.2

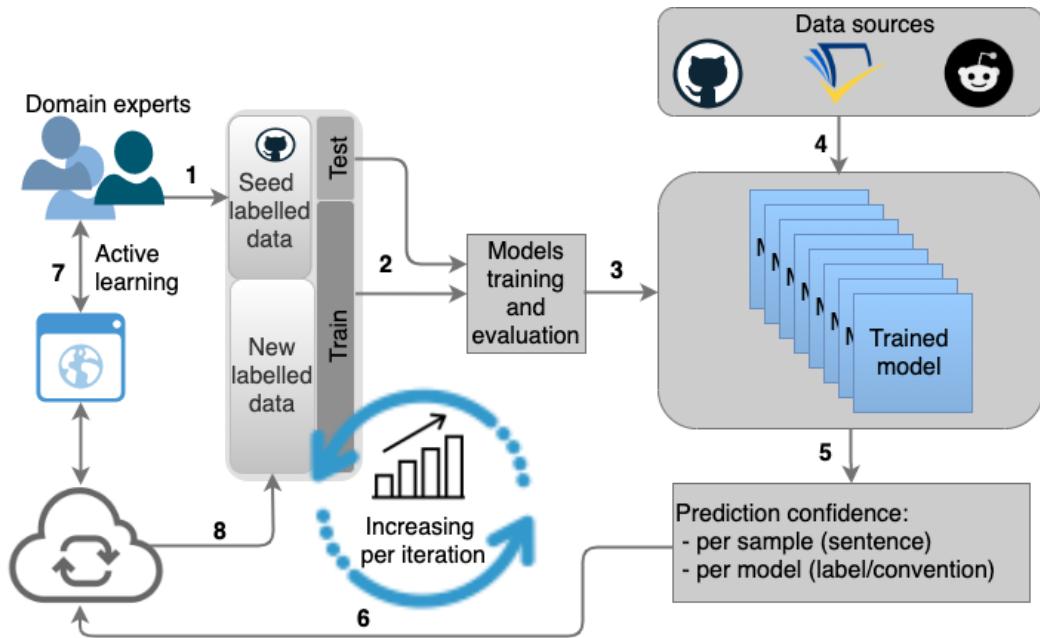


Figure 4.1: Active learning pipeline to collect and verify training data

and improve classifier performance like that. Further focus is on correctly labeling items classified with a confidence close to the classifier's decision boundary (i.e. between 0.4 and 0.6) and a strong focus lies on confirming the models' belief in items with a confidence score above 0.8). A total of 60% of the labeled samples in our dataset come from high confidence predictions, 35% are (re-)labeled from the low confidence predictions and the remaining 5% come from the interval around the decision threshold.

The models are updated with an iterative active learning pipeline. After each iteration the model is evaluated on a fixed labeled set of items of 20% of the (growing) entire dataset. A fixed set is suitable for fast evaluation. The pipeline illustrated in figure 4.1 includes the following steps:

1. The classifiers are pre-trained with seed data. To this end, domain experts labeled a random set of sentences from the GitHub subset.
2. In the first iteration, the eight classifiers are trained with the seed data, new labels are incorporated in succeeding iterations.
3. The performance of the trained classifiers is evaluated on labeled data and they are ready for predictions on unseen data.
4. Sentences from GitHub, S2 and Reddit are classified.

Convention	Accuracy	AUC	N	$E_{prevalence}$
Industrial	0.750	0.708	1289	1/10
Project	0.801	0.828	521	1/100
Market	0.870	0.931	1082	1/100
Renown	0.812	0.859	301	1/100
Civic	0.902	0.897	477	1/1000
Inspired	0.801	0.895	355	1/1000
Domestic	0.866	0.901	475	1/1000
Green	0.901	0.931	280	1/10000

Table 4.4: Comparison of model performance per convention

Data source	Accuracy	AUC
GitHub	0.792	0.823
S2	0.748	0.749
Reddit	0.789	0.765

Table 4.5: Model performance per data source

5. The classification outputs eight confidence scores per sentence (one per classifier).
6. The aggregated data containing sentences and the associated confidence scores is pushed to a centralised cloud service and consumed by our web based active learning tool⁷. Since the labeled data should be representative of the available unlabeled data, The active learning tool shows a histogram to provide insight to the most beneficial areas of focus for the domain experts.
7. Domain experts validate or relabel sentences with a confidence score or label unseen sentences.
8. The labeled sentence is added to the training data for the next iteration. A separate algorithm ensures equal numbers of positive and negative examples per classifier to avoid imbalance. Steps (2) to (8) are repeated until training data suffices.

We ensure label quality with quality checks using a Qualitative Data Analysis (QDA) software⁸ following the principle of deductive procedure for content analysis [137] parallel to the iterative active learning pipeline approach. We ensure

⁷A Python-based interactive GUI

⁸<https://atlasti.com/>

the validity and reliability of the qualitative analysis by means of investigator triangulation. Investigator triangulation involves the use of multiple researchers in an empirical study [9].

Our investigator triangulation involves three authors of this chapter from different disciplines in the coding and labelling process and external EC-experts, with whom codes and labels are contrasted and discussed. The final coding iteration is performed on a random sample of 100 threads per data set, including context information such as links to the original posts in order to account for the situational approach of the EC.

4.6 Results

This section evaluates the performance of the classifiers on the entire dataset as well as on each subset. Furthermore, we present a quantitative and qualitative analysis of the predicted conventions.

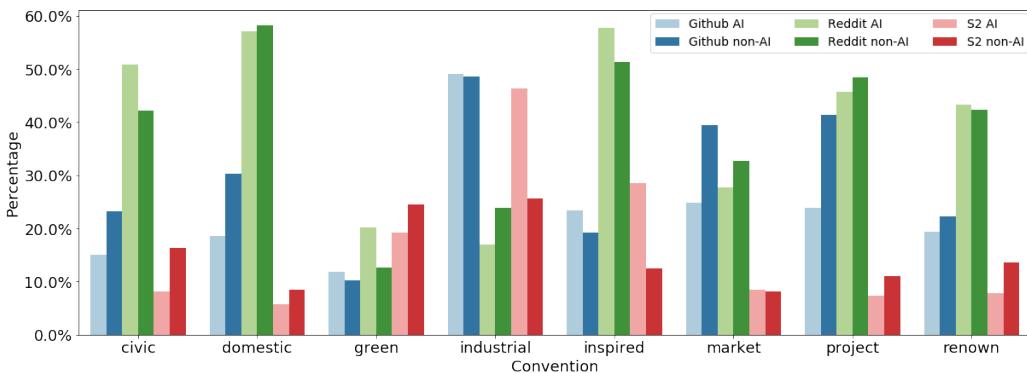


Figure 4.2: Percentage of conventions in each data subset for AI and non-AI related items as predicted by the classifiers.

Performance of Classifiers

We evaluate the performance of the classifiers with the following metrics:

- **Accuracy:** Accuracy is the ratio of correctly predicted elements between all the samples. Accuracy measures the ability of the classifier to identify elements from the positive and the negative classes and also considers the ability to differentiate positive samples from the negative ones.
- **Area under curve (AUC):** The AUC score provides an aggregate measure of performance across all possible classification (confidence) thresholds.

AUC can be interpreted as the probability for a model to rank a random positive example higher than a random negative example.

- **Precision:** Precision is the ratio $tp/(tp + fp)$ where tp is the number of true positives and fp the number of false positives. Precision is intuitively the ability of the classifier not to label as positive a sample that is negative. Precision is used to set the performance acceptability threshold for the built classifiers.

Each of the models is independently evaluated on the test set with both metrics using leave-one-out cross validation. For each classifier, a classification threshold with value $T_{calibration}$ is selected so that at least precision of 90% in test is obtained. Having similar precision for all of them facilitates the comparison of their predictions and ensures a limited amount of false positives.

Table 4.4 contains the average score for each classifier according to the following metrics: the number N of training samples for each convention and a value $E_{prevalence}$ referring to the estimated prevalence of each convention in the dataset, which we determine in a manual analysis. Only a small number of conventions with a high discrepancy between N and $E_{prevalence}$ are in the dataset, so we collect samples from other data sources to train such classifiers. Learning curves provide insight about the amount of labeled data which the classification models require to achieve satisfactory results and the amount they need to improve the results. We use ten fold cross-validation to split the whole dataset $k = 10$ times in training and test set. Accordingly, the classifier is trained repeatedly on all but one of the subsets and evaluated on each one of the other subsets and a score for each training subset size and the test set is computed. Afterwards, the scores are averaged over all k runs for each training subset size.

In order to show that the classifiers generalize across all data sources, we calculate their performance for each individual data source. Table 4.5 shows average scores on equal numbers of positive and negative examples per convention. We see very similar performance across data sources.

A confusion matrix illustrates how well each classifier differentiates between positive and negative samples. The diagonal represents the ratio of true positives whereas the rest of the matrix corresponds to false negatives. Rows of the confusion matrix are normalized by using the total number of examples having a certain true label, so numbers represent the percentage of samples from each convention matched by each classifier.

Figure 4.3 shows the confusion matrix for each classifier using the $T_{calibration}$ threshold. To create the confusion matrix we select only sentences with a single label. Values in the cells represent the amount of sentences matched by each classifier for each convention. High values between 0.6 and 0.92 accuracy are in

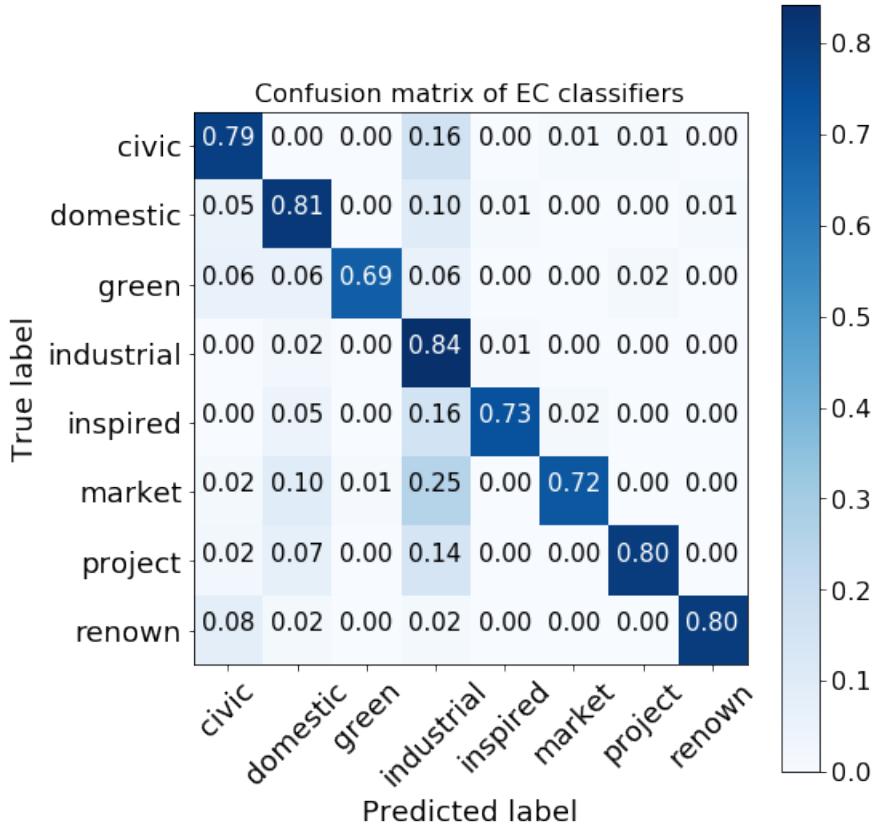


Figure 4.3: Confusion matrix of EC classifiers using the obtained calibration threshold

the diagonal axis of the matrix – the classifiers are correctly differentiating. The Classifiers for the *Civic* and *Market* conventions are performing best.

4.6.1 Evaluation of Conventions

In the following evaluation, we discuss our EC classification results, compare the conventions in AI and non-AI subsets of our dataset, and present the co-occurrences of conventions.

Figure 4.2 shows the distribution of both AI and non-AI related sentences for each data subset. In general, the prevalence of the different conventions is fairly aligned with the estimated ones. Regarding the different conventions, the *Industrial* convention is very dominant in Github (AI and non-AI) and S2 (AI) with a proportion of about 50%. As Github consists mainly of technical descriptions and standards and S2 of scientific abstracts, this is in line with our expectations. In S2,

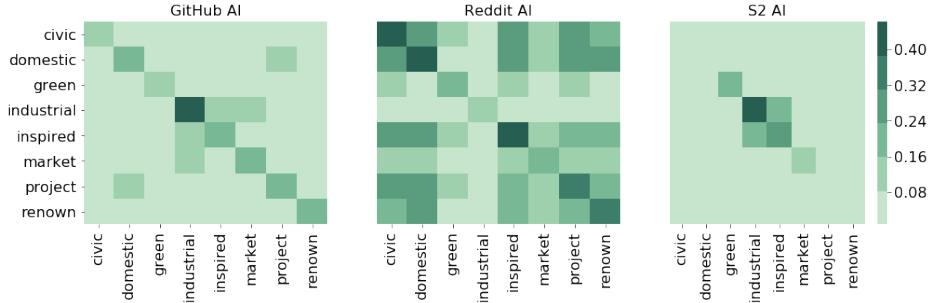


Figure 4.4: Co-occurrences of conventions in the predictions for AI subsets. Values in the matrices are normalized by the number of sentences in each data source.

the *Civic*, *Domestic*, *Market*, *Project* and *Renown* conventions are rarely present, while the *Inspired* convention refers to innovative approaches and the *Green* convention links with ecological projects. In Github, the *Market* and *Project* conventions – somehow stronger in the non-AI texts – are quite dominant, referring to licensing or commercialization for the first one and to the field of computer science, programming, and software for the second one. In contrast with these two subsets, the *Industrial* convention shows a lower percentage in Reddit, together with the *Green* convention, while it is dominated by a cluster consisting of the *Inspired*, *Domestic*, *Civic* (at least for the AI-texts), *Project*, and *Renown* conventions. Therefore, Reddit seems to be more balanced, due to the presence of a different set of conventions, reflecting the variety of topics and approaches in its discussions, while Github and S2 are dominated by one or two conventions. Generally speaking, the *Green* convention is scarcely found (at least in Github and S2), showing that ecological and sustainable considerations are of little importance in these two subsets. The *Market* convention often refers to questions of (commercial) licensing or business models, it was not excepted in the scientific articles, while it should be more present in software development.

The comparison of conventions of AI and non-AI samples reveals interesting tendencies for all three sources. By carefully looking at the results shown in figure 4.2, a positive ratio can be observed between AI and non-AI domains for two conventions: the *Domestic* and the *Project* one. Only the *Inspired* convention shows a negative ratio for all three subsets, confirming that AI related texts are more related to innovative and inspired approaches than non-AI ones. Interestingly, the ratio for the *Industrial* convention differs between the three subsets with nearly no difference in Github, a positive ratio in Reddit and a negative one in S2, highlighting the importance of standardization and scientific methods

Figure 4.4 shows the co-occurrences of conventions in the AI related items. The most interesting finding is the dominant correlation between the *Industrial*

and *Inspired* conventions in the S2 subset, confirming its specific scientific character. In Reddit, validating the findings from figure 4.2, we can observe a rather balanced proportion and co-existence of conventions, with slightly higher correlations in the combination of the *Domestic* and *Inspired* as well as the *Domestic* and *Project* conventions. This is in line with reflections on traditional and experienced-based ways of doing, as well as discussions on power and hierarchy, present in the Reddit subset. In contrast, Github shows a slight surplus in the combination of *Industrial* and *Inspired*, as well as *Industrial* and *Market* with percentages over $\sim 10\%$, showing the content alignment of this subset. In none of the subsets, we find significant co-occurrences with the *Civic* convention, indicating a certain disconnection between civic values and the other dominant conventions in the AI domain.

Qualitative sentence evaluation

Automatic convention classification goes beyond merely detecting significant buzzwords. The correct attribution of a label has to include the buzz words, which refer to the ‘worth’ of each convention. Additionally and more important it also must include a corresponding practical test (see [34]), which checks the corresponding ‘worth’. In the case of the *Industrial* convention that is a procedural test, as any process can be only classified as *Industrial* - in the sense of the EC - if it develops or produces something efficiently and productively in a standardized way. A label is only correct if this test is passed.

To illustrate this procedure and show the reliability of our classifiers on the basis of these requirements we compare a list of three sentences pairs (one pair per data source). The sentence pairs consist of one high accuracy ('good example') and one low accuracy ('bad example') sentence per data source from the *Industrial* convention:

In example (2) from S2, the buzz-word “effective” does not automatically mean that this sentence belongs to the *Industrial* convention. Simple technical descriptions such as example (4) from GitHub does also not imply any convention, although technical, scientific or industrial words are used. In contrast, (1) (extracted from S2) or (3) (extracted from GitHub) include buzz-words, such as “approximation”, “significantly”, or “optimization” and they refer to standardized processes. Accordingly, they belong to the *Industrial* convention. The Reddit example (5) implies modelling as the central process for obtaining efficiency (corresponding with the industrial convention), while the example (6) from the same data source does not refer to an industrial standardized process and therefore corresponds to the *Domestic* convention.

We carry out several iterations of labelling, training and qualitatively analyzing the conventions. The analysis of sentences based on these conventions in-

S2

-
- (1) *Graph partition can then be formulated as searching an optimal interface in the node weighted directed graph without user initialization.*  
 - (2) *Effective soil mapping on farms can enhance yields reduce inputs and help protect the environment.*  
-

GitHub

- (3) *It is often able to determine a good approximation of the true pareto front in significantly less iterations than genetic algorithms.*  
 - (4) *Full documentation is available at: docs.syph.com repository is an apache licensed java reference client implementation for working with the api.started to get started you'll need some api credentials i.e a 'client-id' and 'client-secret'.*  
-

Reddit

- (5) *They use it to model things like large scale particle interactions in a more computationally efficient way.*  
- (6) *I would actually prefer if it generated Java code so I could tweak it by hand.*  

cludes context information of the coded threads in order to determine the ‘practical test’ and achieve a first step in grasping the social complexity of the EC in an automated classification.

4.7 Discussion

The EC and the automatic classification of the conventions offer a comprehensive insight into the dominant conventions and moral orders in the AI-field, partly linking and explaining the functioning of the five primary categories of motifs listed in Related Work (Section 4.2)s. For instance, the *Inspired* convention can be associated with the categories of intrinsic motivation and learning (e.g. development of personal skills or knowledge), whereby the latter is also partly represented by the *Domestic* convention. Furthermore, the category of external rewards can be attributed to the *Market* convention and community recognition to the *Renown* convention. An important finding in this regard is that the *Industrial* convention, which turned out to be one of the most dominant ones in the subsets investigated (see section "Evaluation of conventions"), is not reflected by any of these motifs.

There are ongoing discussions and research on the backgrounds and moral orders, which influence the development of the digital world. In this context, [46] refers to the evolution of the internet as the result of the intersection of diverse cultures, from the purely ‘geek’ and technocratic to the out most capitalist, melded with that of hackers and libertarians. The present study of the prevailing conventions in AI research, development and discussions continues and deepens this

reflection, showing that there is a certain dominance of a techno-meritocratic culture (reflected in the *Industrial* convention), at least in the scientific and technical descriptions of the AI projects. Less influence – depending on the specific project and topic – of the virtual communitarian culture (the *Civic* and partly *Project* and *Green* conventions), the entrepreneurial culture (reflected in the *Market* convention) and the hacker culture (the *Domestic* and *Inspired* culture). In contrast, the Reddit subset includes blog posts, conversations and discussions on a variety of issues related to the field of AI, including ethical reflections, historical analysis, utopian and dystopian views. Hence, in the qualitative analysis done by Mayring et al. [137] of the randomized sample of Reddit subset, pre-classified by the automatic classifiers and focusing on the concurrence of conventions (in the same sentence or in consecutive sentences), no dominance of one or two conventions is observable. Rather, Reddit seems to be characterized by a couple of specific co-occurring conventions, which seem to be central to the discussions about AI, indicating possible (ethical) conflicts. There seems to be, e.g., an ongoing conflict between the *Industrial* and *Domestic* convention around AI, reflecting discussions about the desirability and possibility to develop human-like machines or machine-like humans, and the superiority of human vs. AI. The EC and the automatic classifiers with its underlying concepts of standardization and optimization (in the case of the *Industrial* convention) and trustworthiness, hierarchy and experience (in the case of the *Domestic* convention) illustrates these conflicts. The automatic detection of conventions, as proposed by the classifiers, is able to shed light on the underlying moral assumptions in the AI (and other) fields. By this, it supports a deepened and mutual understanding of different points of view and moral backgrounds.

Our work involves a large amount of human knowledge and interaction. Accordingly, different types of bias might occur. Olteanu et al. [146] report a list of biases in areas such as data acquisition and data querying; data filtering and also biases in results interpretation and issues with the evaluation and interpretation of findings. We briefly discuss the measures we take in this work to promote neutrality and mitigate those biases to the best of our capacities. Due to the size of the content of both Github and Reddit, strong preselection is necessary. This is not the case for S2, where we gather the complete publicly available dataset and perform subsequent steps on the whole dataset. We attempt to gather data from GitHub and Reddit in an equal manner. To ensure extensive discussion and good quality, we collected data from repositories of all different levels of popularity (GitHub) and all the threads with more than 4 up votes (Reddit). To limit the bias in individual researchers' labeling in the active learning pipeline, the researcher triangulation and the sampling process from different levels of confidence both aim to mitigate this problem. We evaluated the EC model with well-known performance metrics by convention and by data source to study potential systematic

differences and incorporated qualitative analysis. We aim to foster reproducibility as well as discussion on methodological approaches, so we release our dataset models and experiments to the research community.

4.7.1 Limitations

We assume similar classifier performance on the AI and non-AI portions of the dataset, although we do not carry out an empirical evaluation of non-AI portions of our dataset; the results for both the AI and non-AI portions in figure 4.4 support this assumption. Furthermore, we assume the wording to be similar in the AI and non-AI portions of the dataset. Even as each data source belongs to a different text type, all data sources for both portions come from the computer science related technical domain. However, this assumption remains speculative and as such it would benefit from empirical evaluation on labeled sentences.

In the approach of this chapter, items in the dataset are analyzed on sentence-level. According to the EC literature, conventions are better reflected on discussions where individuals need to defend their positions. Future work can focus in using the current shape of the EC classifiers to analyze other data sources that, if having a conversational nature, will be better confronting and reflecting the conventions.

Further, we have observed that the proposed techniques are highly dependent on the collection of high quality training data. Although an approach to facilitate such gathering has been proposed, further advances might be required to reduce the amount of manual work to be done by human annotators.

The EC is a social theory based on and therefore limited to Western political philosophy. Further, non-Western ‘moral orders’ are not reflected by the EC and the current analysis. But with further training of the models with non-Western-centric datasets, further conventions might be found, enriching not only the EC, but widening a global comprehension of morals.

4.8 Conclusion

In this work, we described approaches both to analyze and predict conventions according to the EC. We created a dataset mainly from three text sources of scientific research: chapter abstracts from scientific conferences and software development, and analyzed the distribution of conventions in each subdomain. We developed an interactive architecture based on active learning both to support domain experts in data labeling and select the most valuable items to train ML classifiers. Preliminary results on the ML classifiers trained on the EC showed promising results. In an additional study, the results were contrasted with the results from a

classifier trained on software conventions, and we have shown comparable and understandable results on both theoretic frameworks.

Chapter 5

EVALUATING BIASES IN A TWO-SIDED MARKET PLATFORM

5.1 Introduction

5.1.1 Motivation

Two-sided sharing economy platforms have changed how business is conducted in a multitude of domains. They have been particularly disruptive in the real-estate sector where platforms such as Airbnb have changed the status quo. These platforms typically involve three types of stakeholders: (i) providers of items/services, (ii) customers seeking to acquire from the providers, and (iii) the platform itself, which intermediates and matches providers and customers based on their preferences. The explosive growth of these platforms in the real estate sector has been at the core of various political battles at some of the largest cities in the world. Advocates of the sharing economy argue about the benefits they can bring to societies, such as extra income, better distribution and allocation of resources, and the creation of new opportunities for cities and municipalities.¹ On the other hand, critics argue that the costs generated by the platforms surpass their benefits by far: they are very appealing business options so that the main side effect of their wide adoption is that they worsen what is an already troublesome housing shortage in particularly attractive areas, driving up rental prices and, ultimately, boosting gentrification. Concerns also exist about the potentially discriminatory impact of their

¹Airbnb study: Airbnb related activities contributed with up to 175M\$ to the city of Barcelona. <https://www.airbnb.es/press/news/new-study-airbnb-community-contributes-175-million-to-barcelona-s-economy>

algorithms.

The main goal of our algorithmic evaluation is to identify and quantify existing biases in different versions of the platform, showing the trade-offs and potential harms of introducing a machine learning based functionalities, also accounting for the different recommender systems used during the application life-cycle. In contrast with most previous work, our research focuses on the biases exhibited by the system through its recommendations, instead of analyzing how the users behave on the platform [11].

The particular design of the platform, with a baseline method running permanently, executed together with ML-based methods evolving over time, allowed us to extract conclusions in comparison to the baseline.

Our findings show that the introduction of a ML-based algorithm increases the probability of *matching* for the majority of users. This means that the recommendation system effectively facilitate the finding of room-mates or flat-mates.

At the same time, the ranking algorithms utilized in the application exhibit various types of inequalities in terms of performance, significantly affecting the experience and opportunities of some groups of users. Among other differences, the system performance varies across demographic groups based on self-declared gender, sexual orientation, age, and main spoken language. Moreover, we observe that minority groups – groups already disadvantaged or with smaller prevalence in the population – experience lower performance of the system or more differences on its functioning, depending on the particular model they are exposed to.

Research questions

Our research questions are related to the stages of the pipeline depicted in Figure 5.1 and are the following:

RQ1. How effective are the different recommendation methods? If we consider the baseline *random* recommender as a *control*, and each of the ML-based systems as a *treatment*, we would like to answer this question considering both average effects (treatment versus control) as well as heterogeneous effects (different treatments). In the next section we describe suitable metrics for measuring effectiveness.

RQ2. Are there any disparities arising from the usage of ML-based rankings? This is also a question that we address both at the level of average effects as well as heterogeneous effects through appropriate metrics.

5.1.2 Contribution

In this chapter we focus on the latter problem. Specifically, we present a comprehensive and independent algorithmic evaluation of a recommender system of a platform used in the real state market,² designed specifically for finding shared apartments in metropolitan areas. Our examination enjoys full access to the internals of the platform, including details on algorithms and usage data during a period of 2 years. More in detail, the platform aims to help *listers*, i.e., landlords/landladies or room owners, find appropriate *seekers*, i.e., users looking for a room to rent. The recommender system facilitates matching and interaction between seekers and listers, with profile-based matching functionalities resembling those of dating platforms [104]. Listers can “like” the profiles of seekers and send a request to them. Seekers can accept such requests in case they like the offered room. If a lister sends a request to a given seeker and the latter gives a positive response to it, then a *match* occurs, which lets them talk through an in-app chat service to arrange a meeting and potentially sign a rental contract.

The platform mediates the connections between providers (listers) and customers (seekers), and as a mediator it has the potential to either facilitate or hamper the emergence of societal biases. Indeed, the bias against certain minorities, if left unmitigated, can be amplified through its recommendations [160]. These biases are particularly dangerous in this sector, where the fundamental right to adequate housing [200] might be compromised.

5.1.3 Chapter structure

This chapter is structured as follows:

Section 5.3 provides the details about the platform and the setting for our analysis. It also present the specific research questions that are addressed in the remainder of this paper. Section 5.4 presents the methodology and the specific utility metrics adopted. Section 5.5 describes our dataset and provides some general statistics. Finally, Section 5.6 present in details our experimental results and findings.

The next section describes previous work related to the analysis presented here and provides some background.

²Company name omitted.

5.2 Related Work

5.2.1 Access to housing

Experiments conducted throughout the last decades reveal discriminatory behaviors and practices that negatively affect minorities when trying to buy or rent property. Chambers et al. [49] debate the idea of *sustainable livelihoods*, that as they explain, require social equity among other things to achieve sustainability. They expose their ideas with a special focus on the rural poor and other minorities. Turner et al. [199] describes a series of experiments in 23 metropolitan areas in the United States, revealing serious differences between white and minority citizens on different aspects related to access to housing for renting or buying. Wachter et al. [208] show that there are persistent differences in homeownership rates across racial and ethnic groups in the US.

More recently, an experiment conducted by the Barcelona city hall showed how prejudices decrease the opportunities of finding housing to buy or rent for some groups. In particular, it was observed that LGBTQ seekers or those with Arabic sounding names had a lower chance of being scheduled for visiting a flat [74].

In contrast with these previous works, our experiments are based on an online platform in which the contact between users is mediated and influenced by a recommendation algorithm. Although the observed behaviour in the system could be a mirror of societal biases contained in the training data of the machine learning system, those biases, if not mitigated, can be amplified by an algorithm.

5.2.2 Algorithmic fairness in double-sided markets

Analyzing the case of Airbnb, Quattrone et al. [160] outlined the difficulties of creating regulatory policies in a changing environment. They collected a set of recommendations for regulating Airbnb, contributing to the general idea of “algorithmic regulation”, which advocates for the analysis and use of large sets of data to produce evidence-based regulations that are responsive to real-time demands. Shur et al. [192] analyze a double-sided market in the context of ride hailing platforms, giving an special emphasis to the role of the riders (producers). Hutson et al. [104] analyze a similar setting, in their case online dating apps, revealing different inequities based on race and/or sexual orientation.

Our work contributes to this research as the first work that studies how different versions of a system facilitate the goals and preferences of users in different sides of the market. Also, we quantify the effects of using a ML-based algorithm in comparison with a rules-based random baseline.

5.3 Setting

The platform analyzed in this work corresponds to a system that aims to help matching users having available rooms in their flats, with potential new tenants or flat-mates/room-mates. This setting can then be described as a two sided market, where *listers* supply rooms that are in demand by *seekers*. Most of the interactions are done through a mobile app that offers a recommendation list for the listers.

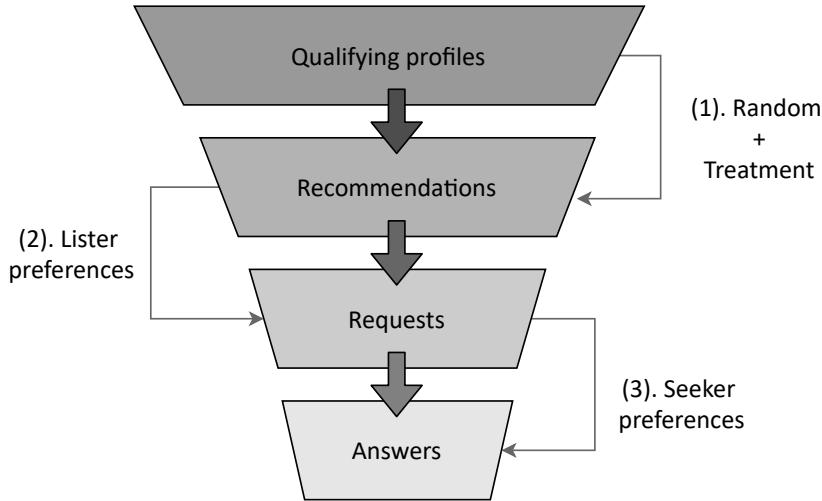


Figure 5.1: Platform's recommendation pipeline.

As depicted in Figure 5.1, listers receive recommendations in the form of an ordered list of ~ 20 recommended seeker profiles. These profiles come from a pool of qualifying profiles (e.g., seekers searching for a room in the area where the lister's room is located), from where recommendations are selected. This selection might include profiles selected by the baseline method interleaved with profiles prioritized by a ML-based recommender system. This allows the platform to monitor in real time the performance of each of the RecSys versions compared to the random group, using a within-subjects [176] A-B testing [115]. The same user can be exposed to A or B treatments in different visits or receive recommendations given by A and B in a given recommendation list.

In the following, we refer to the baseline system as *random* and to each of the ML-based systems as a *ranking*, given that their main difference is how they rank qualifying seekers. We analyze the performance of *random* and *ranking* separately to understand their differences, and the implications of introducing a ML-based system.

Once the list of recommendations is shown to the listers, they select and can send a request to a subset of seekers according to their own preferences. After the listers send a request, seekers receive a notification for each of them. These requests can, in turn, be accepted or rejected by seekers.

In the following, we consider all the ranking systems together as opposed to the random system. However, we note that different recommendation models that were developed at different points within the life cycle of the platform are used, and their training sets are slightly different. Although it is something outside the scope of the present chapter, and a limitation of our work, each new version of the ranking system may have been influenced by the behavior of older versions, and this could lead to feedback loops amplifying biases for each new version of the system, as an example of the cyclic nature of bias 1.1. For the purpose of this study, we compare the performance of the different models, including the random system, as isolated instances, whereas their recommendations can appear together in the recommendation lists. However, each ranking model is optimized for the same objective: to maximize the expected probability of a match.

In this setting, biases can be observed directly (i) in the ranking produced by each system, (ii) in the lister preferences when selecting among the ranked items, or (iii) in the seeker preferences within the received requests.

5.4 Methodology

The methodology that we use to analyze whether the system leads to biased or discriminatory outcomes follows previous studies [175, 78] and consists of four main steps:

1. Identification of potentially disadvantaged groups.
2. Selection of effectiveness and disparity metrics.
3. Computation of relevant metrics for each stage.
4. Comparative analysis of treatment and control settings across groups.

Identification of protected groups

The main purpose of this initial step is to identify potentially discriminated [172], disadvantaged groups whose lack of privileges might be replicated or amplified within the platform. We consider four groupings that can lead to discrimination in this scenario and that we can evaluate with the available data: i) gender, ii) age, iii) languages spoken, and iv) “gay friendly” profiles. We remark that the data made available to us did not include any identifier that allows us to link these attributes to individuals, nor we made any attempt to do so. We also maintained data security by keeping the dataset within our research infrastructure, which can only be accessed by researchers in our team directly involved in this research.

Self-declared gender. Users specify their gender in a binary form (male/female) when registering for the app. The cases where the user did not inform their gender are discarded from the analysis.

Self-declared age. Users also can specify their age. Following [139] we consider the individuals in the range $[18 - 75]$ and then, looking at the distribution of data, create 4 different groups: (i) 18-34 (millennials), (ii) 35-54 (generation X), (iii) 55-75 (boomers) and (iv) < 18 or > 75 (outlier)

Languages spoken. The city from which we use data (Barcelona) is a cosmopolitan city hosting people from a variety of places. The main languages declared by users of the platform in this city are Catalan, Spanish, English, and Italian. Basically all Catalan-speakers users of the platform in Barcelona also speak Spanish. Hence, we compare these majority languages against cases in which the listers indicated other languages (such as Arabic).

“Gay friendly” profiles. Many descriptions of listed rooms, as well as profiles of individuals, included phrases such as “gay-friendly” or even “only gay-friendly people are welcome.” Users are not asked to declare sexual orientation in this platform, but as sexual orientation had been found to be one determinant in access to housing [74], we consider that analyzing this “gay friendly” signal was important. We use a set of phrases that are variants of “gay friendly” to detect descriptions fitting this category.

Understanding that users in each side of the market might have different goals and/or preferences, we additionally *separate people according to their role within the platform* (lister or seeker).

Utility metrics

To define the metrics, we first need to introduce some notation. Let \mathcal{U} represent the set of all users, with \mathcal{U}_L corresponding to listers, and \mathcal{U}_S corresponding to seekers, in such a way that $\mathcal{U} = \mathcal{U}_L \cup \mathcal{U}_S$. We remark that a small fraction of users ($\approx 4\%$), are listed as both room-owners as well as room-seekers.

Let \mathcal{H} represent the set of rooms, and $H : \mathcal{H} \rightarrow \mathcal{U}_L$ associate each room with its lister. Let $\mathcal{R} \subseteq \mathcal{H} \times \mathcal{U}_S$ describe the recommendations presented to the listers, i.e., the different seekers selected for each room. Let $\mathcal{X} \subseteq \mathcal{R}$ be the requests created from such recommendations, i.e., the instances in which the recommendation was followed by a lister who contacted a seeker, and finally let $\mathcal{A} \subseteq \mathcal{X}$ the instances in which the contacted seeker answered positively to the request.

From the identification of the protected groups we can generate several partitions (e.g by gender, age, language spoken and sexual orientation). Users can be partitioned: (i) by gender (\mathcal{G}), (ii) by age (\mathcal{Y}), , (iii) by language spoken (\mathcal{N}) and

(iv) by “gay friendly” (\mathcal{F}). We use the symbol \mathcal{P} to reference the complete set of partitions.

Some of our utility metrics are independent of the role that a user has in the system. For instance, we assume that users in both sides want to minimize the effort required to find a roommate. Other metrics recognize that in some cases users may have opposite goals. For instance, listers want to minimize the income they obtain by renting their rooms at the highest possible price, while seekers seek to rent a room at the lowest possible price, all other things being equal.

DCG - Discounted Cumulative Gain (for listers)

This is a measure of ranking quality, which in our case measures the value of a list of recommendations given to a lister. The metrics consider the positions in the ranking list of the items that a user finds relevant [107]. In its more general form, given a list of recommendations $R = \langle (r, u_1), (r, u_2), \dots, (r, u_{|R|}) \rangle$ for a room $r \in \mathcal{H}$:

$$DCG_R = \sum_{i=1}^{|R|} w_i \cdot v_i$$

where w_i is a discounting factor that decreases with i , and v_i is the relevance of the i -th recommendation in R .

A common choice for the discounting factor is logarithmic discount: $w_i = 1 / \log_2(1 + i)$. The relevance of the i -th recommendation can be defined as the extent to which $H(r) \in \mathcal{U}_L$, the lister of room r , will consider $u_i \in \mathcal{U}_S$ an appropriate candidate for renting the room. The discounting factor stresses the requirement that the most useful recommendations should appear near the top of the list.

We use the normalized version of DCG that is divided by its maximum possible values, so the resulting $nDCG$ is in the range $[0, 1]$.

CR - Conversion Rate (for listers)

A “conversion” in online marketing indicates a successful traversal through a funnel, e.g., becoming a purchasing customer. In our case, success for a lister means finding of a suitable seeker, hence CR measures the probability that a request sent by a lister is accepted. If \mathcal{X}^ℓ are all the requests performed by lister $\ell \in \mathcal{U}_L$, and \mathcal{A}^ℓ are all the requests that are accepted by the recipient seekers, then:

$$CR_\ell = \frac{\mathcal{A}^\ell}{\mathcal{X}^\ell}$$

CTR - Click Through Rate (for seekers)

This indicates the probability that a seeker is contacted after being shown to a lister. Similar metrics have been used before to approximate item relevance for users [139], and CTR is a common metric used to evaluate, for instance, the relevance of web pages in personalized advertisement [167]. In our case, for a generic

seeker $s \in \mathcal{U}_S$, we consider the fraction of listers who click on him/her over the total number of listers that saw him/her. Let \mathcal{R}^s be the set of recommendations containing the seeker s and \mathcal{X}^s the set of requests created from such recommendation by the listers:

$$CTR_s = \frac{\mathcal{X}^s}{\mathcal{R}^s}$$

e_s – Exposure (for seekers)

Differences in exposure have been recently studied to evaluate whether ranking models used in search and recommendation treat people from different groups similarly [185]. In our setting, we consider \mathcal{R}^s , which are all the recommendations including a particular seeker s , and the position $p(s, r)$ of the seeker s within a particular recommendation $r \in \mathcal{R}^s$.

$$e_s = \sum_{r \in \mathcal{R}^s} w_{p(s, r)}$$

where w_i is a discounting factor that decreases with i , as in the computation of DCG .

Assuming to consider a subset $S_a \subseteq \mathcal{U}_S$, where all the seekers considered in the subset are characterized by the property a (e.g. a sensitive attribute), we can quantify the **disparate exposure** received by the group as:

$$DT(S_a) = \frac{\sum_{s \in S_a} e_s}{\sum_{s \in \mathcal{U}_S} e_s} \times \frac{|\mathcal{U}_S|}{|S_a|}$$

Where $|S_a|$ and $|\mathcal{U}_S|$ corresponds to the size of the two sets. This index is inspired by the metrics already introduced by [185]. This non-negative metric $DT(S_a)$ is equal to 1 when the exposure generated for the group S_a is proportional to its relative size, if $DT(S_a) < 1$ then the group is *under-exposed* while for $DT(S_a) > 1$ the group is *over-exposed*.

5.5 Dataset description

The platform that we study operates in several large cities across the world. We select the city in which the platform has its largest use base, Barcelona. The dataset gathered for this research contains 4,296,000 rows describing recommendations issued during a contiguous 30-months period from January 2017 through June 2019. It contains information about 61,997 unique users. Each recommendation includes a *lister* and *room* for which the recommendation is created, and the *seeker* that is recommended for that room and lister. Including the position in which each seeker was listed and the utility score assigned by the ranking system

to it. Additionally, when the lister initiates a *request* from the recommendation, we have information that a request was initiated and about the response from the seeker addressed by the request. Responses by the seekers include accepting or rejecting the request, or leaving it pending, which means the request expires when the room is rented or becomes unavailable. The dataset also contains demographic information about the *age*, *(binary) gender*, *level of studies*, *work occupation*, and *spoken languages* for both seeker and listers.

Model	Listers	Seekers	Recommendations	Requests	Conv. Rate
BSL	35.72K	6.76K	1.78M	343.82K	19.37%
CF	15.35K	794.00	200.59K	45.72K	22.79%
MF	9.78K	7.83K	2.54M	568.37K	22.37%
XGB-1	4.02K	1.21K	396.54K	80.45K	20.29%
XGB-2	10.47K	5.07K	237.66K	84.54K	35.57%
XGB-3	3.80K	3.18K	384.31K	101.22K	26.34%

Table 5.1: Summary of the number of recommendations created with the different models through the operation of the platform. BSL is the random baseline; the other models are based on Machine Learning.

General statistics

The dataset contains baseline and ML-based recommendations. The baseline recommendations (BSL) are based on a random selection of available seekers for a room. They have always been provided by the platform, throughout its entire operation, and are used as a control. The ML-based recommendations have gone through several re-design iterations, including the following models:

- **Collaborative filtering (CF).** A collaborative filtering model trained to maximize the probability that listers send requests to the recommended seekers.
- **Matrix factorization (MF).** It corresponds to an instance of a Factorization Machine inspired by the model proposed by [166]. It included features from

Table 5.2: Percentage of seekers (S) and listers (L) belonging to different groups.

Model	Male		Female		Baby-boomer		Generation-X		Millennial		Outlier		Eng-Ita-Spa		Other		No-gay-friendly		Gay-friendly	
	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S
CF	44.04	50.41	55.96	49.59	3.78	1.53	42.58	25.00	53.26	73.3	0.39	0.17	88.55	82.84	11.45	17.19	99.46	99.67	0.54	0.36
MF	43.80	56.22	56.20	43.78	4.13	1.44	37.59	22.7	57.96	75.71	0.32	0.15	98.14	99.05	1.86	0.95	99.30	99.19	0.70	0.81
XGB-1	45.00	59.86	55.00	40.14	3.72	1.15	36.75	22.61	59.15	76.00	0.38	0.25	94.45	96.94	5.55	3.06	99.26	99.15	0.74	0.85
XGB-2	46.69	43.23	53.31	56.77	4.03	1.33	39.49	22.67	55.98	75.90	0.50	0.10	98.88	99.50	1.12	0.50	99.47	99.27	0.53	0.73
XGB-3	43.76	55.22	56.24	44.78	4.03	1.54	37.53	23.4	57.98	75.04	0.46	0.02	99.8	99.76	0.20	0.24	99.52	99.18	0.48	0.82
BSL	43.49	51.22	56.51	48.78	4.10	1.67	38.23	25.02	57.29	73.15	0.38	0.15	94.75	89.52	5.25	10.49	99.42	99.43	0.58	0.58

the rooms.

- **XG-Boost** During the operation of the platform, different versions of XG-Boost (gradient boosted decision trees) have been used: (i) **XGB-1**, first version of the model, which optimizes the probability of sending a request; (ii) **XGB-2**, second version, which optimizes the probability of a match, following the approach introduced by [206], (iii) **XGB-3**, third version, which optimizes the probability of matches leading to actual rentals.

The number of recommendations generated by each method, as well as the time periods in which they were generated, are presented in Table 5.1. A summary of demographic information is reported in Table 5.2. In the following, we will use the acronym RS to refer to all the ML-based ranking systems together, in contrast with the baseline BSL.

5.6 Results

In this section we report our analysis and our findings with respect to the research questions introduced in the subsection 5.1.1.

Observed performance and disparities in the recommendations

We begin the evaluation by analyzing the first step in the recommendation pipeline (Figure 5.1). This part of the funnel selects a set of qualifying profiles, i.e. the list of suitable seekers according to the preferences selected by the lister for a room, then it ranks them and shows the top 20.

Lister side. We first compare, from the perspective of the listers, the relevance of recommendations selected by the random baseline (BSL) against the performance of recommendations created by any of the ML-based ranking system (RS). We assess the quality of the recommendations, computing the $nDCG$ by considering that the relevant items are the seekers to whom the listers send a request. This utility metric is computed at individual level and then aggregated for each demographic group.

The random BSL exhibits an average $nDCG$ score of 0.42. The performance by demographic groups is shown in Figure 5.2.

The introduction of the ML-based ranking system leads to an increase in the overall performance, with an average $nDCG$ score of 0.49. However, the increase in performance of the ranking system is not equal across the different demographic groups, as shown in Figure 5.3.

The observed differences in the $nDCG$ score indicate that most groups obtain better recommendations, except for the “Gay-friendly” group, one of the minorities considered in our analysis, who got a decrease of 2.3% of the $nDCG$ score.

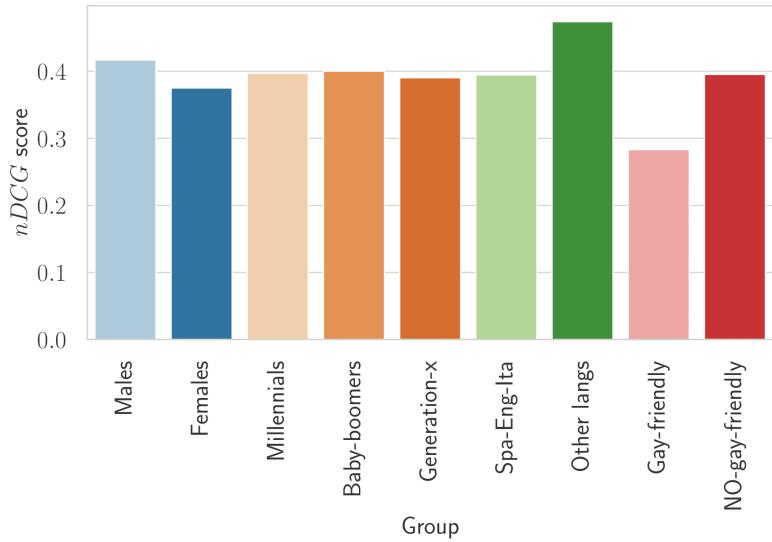


Figure 5.2: $nDCG$ score of recommendations for BSL.

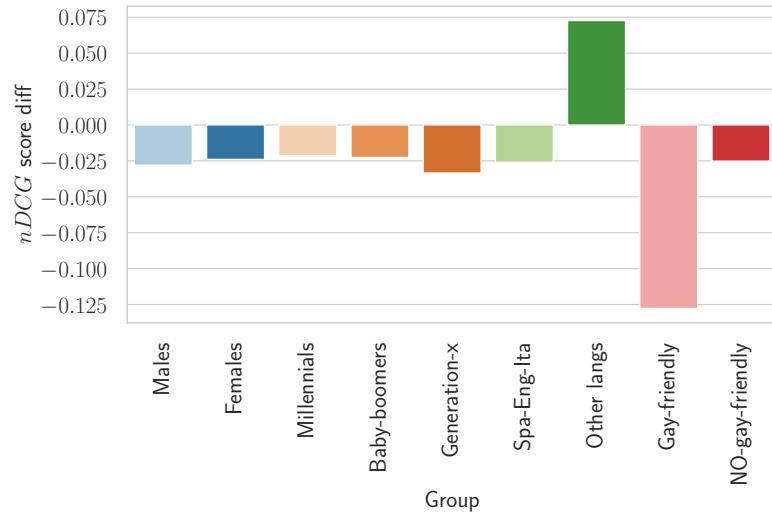


Figure 5.3: $nDCG$ score difference between the RS and BSL across demographic groups.

We next compare the performance of individual models. Figure 5.4 reports the difference in performance between ML-models and BSL across demographic groups.

We observe that *XGB-2* is one of the best in terms of $nDCG$ for most of the groups. On the other hand, the *MF* model is the more robust, since the differences in performance among groups are minimal. It is also the only model reporting a gain of performance w.r.t. the random baseline for the “Other languages” group.

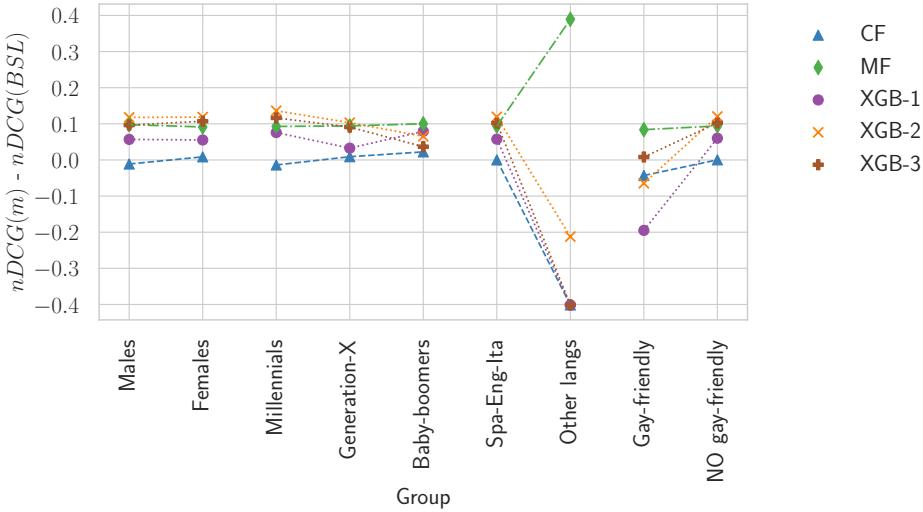


Figure 5.4: $nDCG$ difference between ML-models and BSL.

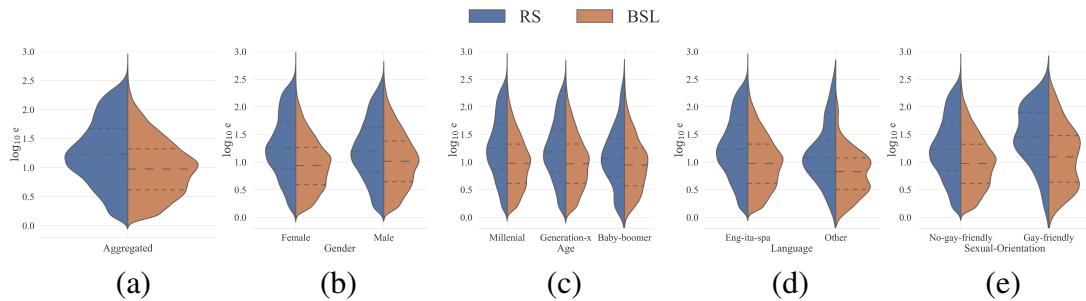


Figure 5.5: Exposure distribution comparison (log-scale) between BSL and RS: total (Aggregated) and by demographics (Gender, Age, Language and Sexual-Orientation). The dashed lines in each violin plot represent the first, second and third quartile.

CF is the one showing the larger differences in performance by groups. In particular “Males”, “Millenials”, “Spa-Eng-Ita” and “No Gay-friendly” obtain better recommendations with the random baseline than with *CF*.

Observation 1 *ML-based ranking models have in general a positive average effect in recommendation performance, but different models lead to heterogeneous effects in terms of quality of recommendations for different groups.*

Seeker side. After analyzing the performance obtained by the listers to whom recommendations are presented, we consider the experience of the recommended

users, i.e., the seekers. To evaluate the recommender systems from the seekers' side, we focus on the *exposure* they receive. As in the previous section, we first look at the average effect of the ML-based ranking systems (BSL vs RS), then perform an analysis per model.

In Figure 5.5 we report the exposure distribution for RS and BSL. Consistently in all the plots we can observe a heavy tail for RS on the larger values of exposure. This indicates that introducing the ML-based model leads to larger disparities in exposure among seekers. This effect results to be stronger for the groups of "Females", "Millennials", "Other" (language) and "No Gay-friendly".

Observation 2 *The introduction of the ML-based recommendations increases the disparity in the exposure distribution: some people get much more exposure than the rest.*

We next analyze the exposure for each model across demographic groups.

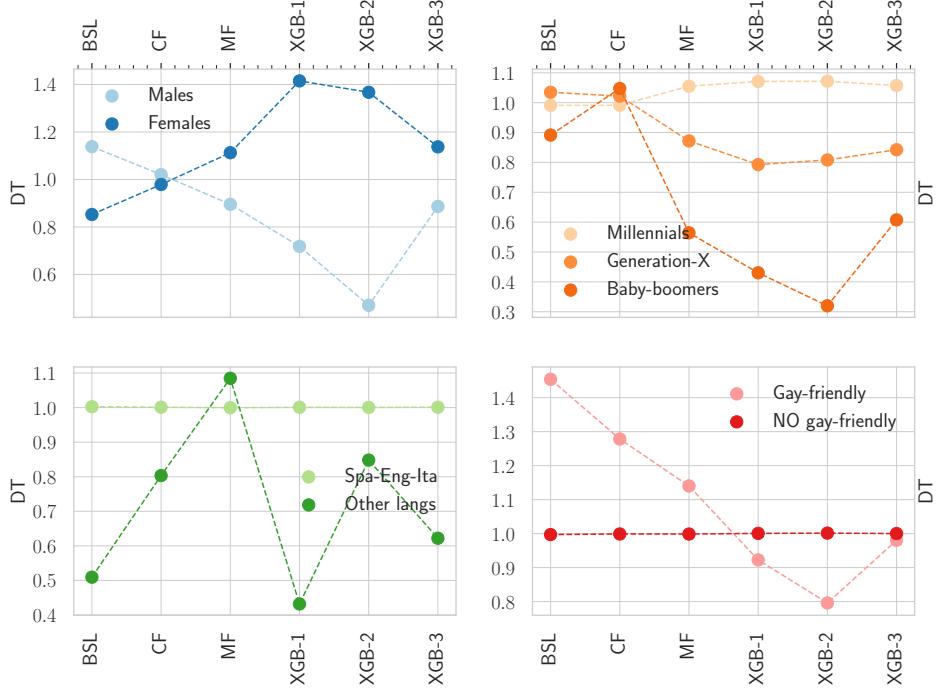


Figure 5.6: Exposure for the different models across demographic groups.

Figure 5.6 reports the exposure that different models give to different demographic groups. It shows that on average the exposure is larger for "Females" than for "Males" for most of the models. Regarding the "Age" partition, the group with more members, "Millennials", obtains a fairly constant average exposure with a little increment for last versions of the models where as the other groups obtain lower exposure in general. For the remaining two partitions ("Spoken languages"

and “Gay-friendly”), we observe how the majority groups obtain an exposure close to 1, meaning that they are shown a numbers of times closely correlated to the size of their group, where as the two minorities experience more variance on their exposure, depending on the individual model that is recommending them.

Observed performance and disparities in the requests

Lister side. We next use the CR (Conversion Rate) metric to quantify the performance of the system for the listers. In general, the random baseline had a CR score of 10.36, which implies that on average, a generic lister needs to send ≈ 10 messages to recommended seekers about a room to get at least one seeker to accept it. By analyzing the CR score aggregated by groups, we obtain the results reported in Figure 5.7. In such plot, we first observe that the system does not present relevant differences of performance along the different subgroups. We can also observe that male listers have lower CR score than females, inverting the trend observed for the $nDCG$ metric used to evaluate the quality of the recommendations. This means the recommendations show to men appear to be more relevant than those shown to women as they click on the top ones more, but once men issue a request to a seeker they have smaller chances than women of getting their request accepted.

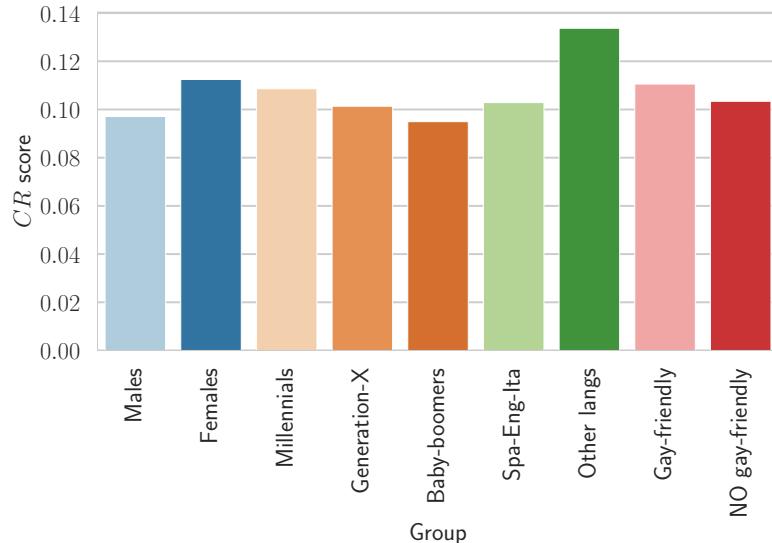


Figure 5.7: CR score for BSU across demographic groups

Looking how the CR score changes (Figures 5.8 and 5.9) with the addition of the different ML-based models, we observe heterogeneous variation of performances along the groups. In Figures 5.8 we notice how two subgroups do not benefit from the use of the RS (“Baby-boomers” and “Other languages”). Figure 5.9 shows the relevant differences in CR across models. None of the models

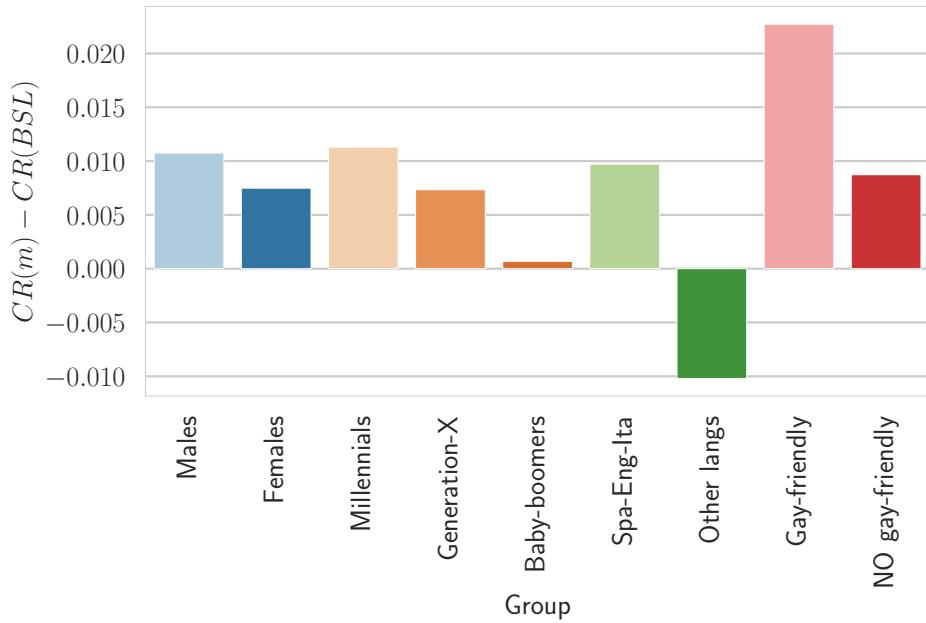


Figure 5.8: Conversion Rate (CR) differences between each ML model and BSL.

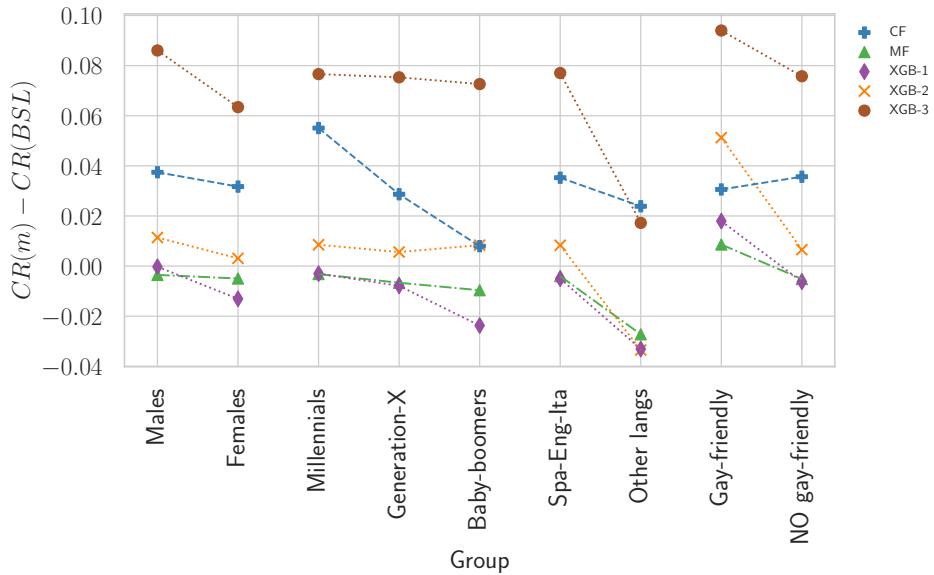


Figure 5.9: Conversion Rate (CR) differences w.r.t. BSL for each model.

is consistent in terms of CR along the groups, where XGB-3 and CF result to be the ones improving the most the baseline. CF is also the one which leads to biggest difference in performance in "Age" group. XGB-2 is capable to generates the smaller differences in CR for the subgroups within "Gender" and "Age". MF and XGB-1 are consistently worse than the baseline, since we observe a CR larger

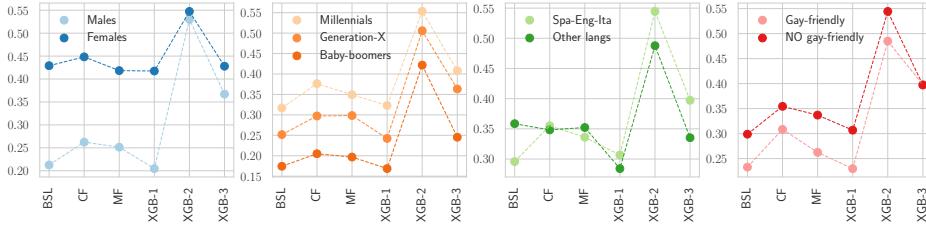


Figure 5.10: CTR differences across different models

than the baseline, only for the “Gay-Friendly” group.

Observation 3 *The addition of ML-based rankings leads on average to improvements in terms of conversion rate, but the levels of improvement are substantially different from one model to another.*

Seeker side. To analyze the effectiveness for the seekers we adopt the Click Through Rate (*CTR*): as usual, such metric is evaluated for each model and across demographic groups. Figure 5.10 reports the results of such analysis. As the baseline model (BSL) is a random selection of seekers, it may better reflect the “raw” preferences of the listers. When comparing against the ML-based models, we notice how the *CTR* does not improve equally across groups. In all the models, except for *XGB-2*, we observe a significant difference between women and men: different strategies and models do not reduce the gap in the two values of *CTR*. Only *XGB-2* is able to obtain the same benefits for both.

Additionally, focusing on the “Age” attribute, we see how the “Millenials” obtain higher *CTR* along all the models while, “Generation-X” and “Baby-boomers” subgroups are always less clicked. The gap between the three categories is in some cases partially mitigated (*XGB-2* and *XGB-3*) but never reduced completely to zero. In the partition by “Spoken languages”, while the baseline model shows a slightly higher *CTR* for the “Other languages” group, this distance is strongly reduced along the other models. We also notice how all the *XGB* models flip the order of the two subgroups, in particular this phenomenon appears stronger in *XGB-2*. Eventually, we observe a systematic gap of preferences between the two subgroups in the “Gay-Friendly” partition. The “No Gay Friendly” subgroup experiences an average positive difference in *CTR* of 5%, except for the case of *XGB-3*, which leads to same *CTR*. Finally, we also see how *XGB-2*, which is optimized for matches, is reflected here with a gain of *CTR* for all the groups, probably explained by the fact that such model is doing a better work on recommending seekers to the listers that will be interested on them.

Observation 4 Increases in Click Through Rate (CTR) by the ML-based recommenders are not consistent across groups. The changes in CTR are not aligned with the changes in exposure across groups and models.

Performance and equity trade-off.

In the context of Learning-to-Rank (LTR) [186] defended the necessity of considering not only ranking utility to the users but also enforce the need of utility-aware metrics. Adapting this framework, we evaluate the quality of the recommendations that each model provides for the listers, in comparison to a measure of algorithmic fairness, which we define next for both sides of the market. As a measure of utility or quality, we compute the average of the $nDCG$ scores measured for the different groups. We call this new metric **Balanced** $nDCG$, which corresponds for a generic model m :

$$Balanced_{nDCG}(m) = \frac{\sum_{\mathcal{P}_i \in \mathcal{P}} \sum_{a \in \mathcal{P}_i} nDCG(m, a) / |\mathcal{P}_i|}{N}$$

where N corresponds to the cardinality of all the possible partitions defined by demographics. In our specific case, considering all the subgroup, we have $N = 4$. On the other side, we want also to highlight potential discrepancies in performance between demographic groups, in order to compare with the quality measure defined above. To do so, we quantify the algorithmic fairness of each model using a metric inspired by the notion of demographic parity, which states that *each demographic group should receive the positive outcome at equal rates* [67]. We translate this context into two different metrics, one for the listers, one for the seekers. For the listers, we define a measure of disparity based on the average standard deviation of $nDCG$ scores across the demographic groups, called σ_{nDCG} . To define the disparity metric for the seekers, inspired by [186], we look at the ratio of the exposure and CTR by groups:

$$D_{ind}(a, m) = \frac{e_s(a, m)}{CTR(a, m)},$$

Where m is the model selected and a the specific demographic attribute. This new measure D_{ind} express the alignment between the two measures of quality for the seekers. It is a non-negative index that gets lower values when the exposure given to the group is lower than the merit observed (CTR). Then, for each model we compute D_{model} , which corresponds to the average standard deviation of $D_{ind}(a, m)$, computed as follows:

$$D_{model}(m) = \frac{\sum_{\mathcal{P}_i \in \mathcal{P}} \sigma_{\mathcal{P}_i}}{|\mathcal{P}|}$$

Where $\sigma_{\mathcal{P}_i}$ is the standard deviation of D_{ind} observed for all the sensitive attributes belonging to the \mathcal{P}_i . Through the use of this three new metrics we can focus on two different trade-offs concerning accuracy and fairness: (i) one for the listers, which is measured comparing $Balanced_{nDCG}$ and σ_{nDCG} and (ii) the other for the seekers comparing $Balanced_{nDCG}$ and $D_{model}(m)$. In both cases the measure of accuracy is the same, since it is the metric which quantifies most the quality of performance for the platform.

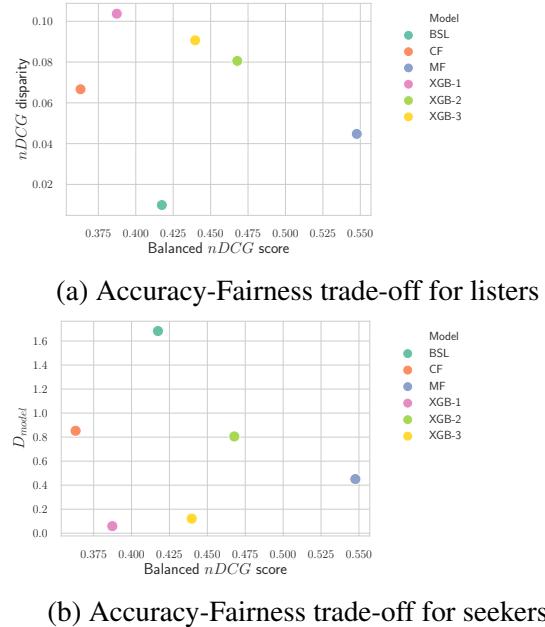


Figure 5.11: $nDCG$ -Fairness tradeoff. Listers between $nDCG$ and disparities per across groups for $nDCG$ (top). Seekers trade-off between $nDCG$ and Exposure (bottom).

Figure 5.11 reports the resulting trade-off plots. We observe in Figure 5.11a that for most of the models improving the average accuracy decreases also the level of unfairness for the listers. Only the BSL shows lower disparity with lower performances. While, observing (Figure 5.11b), we notice that improving the average accuracy of the system tends to slightly increase the level of unfairness for the seekers. CF model is the only one which results to be lower in accuracy but also unfair towards the seekers. The BSL model is interestingly unfair towards the seekers too.

Observation 5 *Improving the accuracy of the system benefits the listers, increasing fairness (as defined above) among groups. On the other side, the disparity in exposure seems to be weakly affected by the improvement in accuracy.*

Disparities in the answers: inequality of incomes

To evaluate the performance and equity for the last step of the recommendations pipeline, we use a different perspective. Understanding that this last step might result in an economic transaction in the form of a rental contract, we evaluate potential inequalities in the incomes users get across different models.

Listers side. This analysis is three-fold: first, we consider the room price assigned to the rooms uploaded by listers, then, we evaluate for potential differences in such distribution with respect to the room price distribution in the requests that were accepted; finally, each model is considered individually to evaluate for potential differences between them. For this analysis, we discard the rooms with prices that are outliers (> 1000 EUR or < 200 EUR).

From this assessment, we first can observe that the ranking system does not imply a significant difference of room price in the accepted requests, with an average difference of $\approx 2\%$ between the ML-based ranking and the random baseline: the average for the random system is 414.77 EUR per month, whereas lister make 425.04 EUR per month when exposed to the ranking system in average.

Nevertheless, as result of the analysis, we detect few cases with differences bigger than 3% in average between the room price in the uploaded rooms and the resulting room price in the accepted requests (figures were omitted for brevity). Among these few cases, none of them reach more than 10% of relative difference in average. Executing a statistically significance test, we find that none of those detected case is statistically significant for $p - value = 0.05$.

Seekers side. We then assess whether there are potential differences in the average price of the rooms accepted by seekers across different models. This analysis does not reveal any significant difference across any of the demographic groups or models. The performed evaluation reports an average of 412.77 EUR and 423.78 EUR per month for the random baseline and ML-based ranking systems respectively.

Observation 6 *The observed disparities in the quality of recommendations shown to listers, probability of listers sending a requests to seekers, and probability of seekers accepting those requests, do not seem to lead to substantial differences in the prices at which rooms are rented.*

5.7 Discussion and conclusions

Understanding the performance of a sharing-economy platform across all its users involved is an arduous task that requires to consider multiple aspects in the assessment. In this paper, we approach this task by first considering the role of different users inside of the platform. To perform our analysis, we need to consider the different goals that users might have depending on the side of the market where they are located. In this context, we conduct a layer-by-layer analysis, evaluating not only the system performance but also potential inequities created by such system for each of the steps in the recommendation pipeline. We also evaluate different versions of ranking system used during the platform life-cycle, and compare them to a baseline model based on random recommendations.

Our results show that compared to the random baseline, ML-based ranking systems on average increase the relevance of the provided rankings for the direct consumers of them, i.e., listers, according to the $nDCG$ score. Splitting this analysis across demographic groups, we observe how certain groups do not benefit equally from average increases in the system performance, and even may be served worse than the baseline in some cases. Focusing on the other side of the market, i.e., seekers, we observe how incorporating the ML-based system increases disparities in exposure among them, leading to a small fraction of users receiving larger exposure, resulting in yet another example of disparate exposure caused by a ML-based system.

Then, we analyze the requests issues by listers when they find a relevant seeker among the recommendations. From there, we first show disparities in the Conversion Rate (CR) metric, a measurement of how easy it is to get accepted by the contacted seekers. During the assessment of the request driven by the random system, we observed small inequities between demographic groups that perhaps merit further analysis. In general, those sub-groups of the population which benefit more according to the $nDCG$ (i.e., they find more suitable seekers to contact among the top recommendations), are also the ones with lower gains in the CR metric (i.e., they are not accepted as much by the seekers they contact).

After that, we observe how the addition of the ranking system created unevenly distributed gains of performance for the CR score across demographic

groups. That, after being analyzed per model, showing significant gains for more sophisticated models. However, once again, minorities or already disadvantaged groups, obtained lower performance for that metric.

On the seekers side, we observed that inequities in Click Through Rate (*CTR*), a metric of interest of ranked users for the listers, were generally consistent across demographic groups. This fact can be interpreted as a systematic failure of recommending certain groups to the listers that would really be interested on them or, as an example of biased user preferences altering a performance metric. Most probably, it could be due to a combination of both aspects.

After analyzing the first two layers in the system, we wanted to empirically validate some ideas introduced by [186], where authors claimed that ranking systems optimized for the utility of the rankings to users, tend to be oblivious on their impact to the ranked items. We assessed this in a fairness-utility analysis for both sides of the market. First, we observed how increasing the utility lead to lower disparities in the same metric for the listers. Nevertheless, we also observed how higher accuracy led at the same time to slightly larger inequities for the seekers exposure, validating the hypothesis. From this analysis, we also observed how the random system was not following the general trend of the rest of the models, most probably because it was not really optimized for the utility of the rankings.

Finally, we assessed whether different models related to inequalities in the amount of the economic transactions facilitated by them (rentals). From this final analysis, we can claim that generally there were no significant differences either for the listers or seekers for each of the models. In other words, while the different inequalities we have observed impact the probability that a user finds a rental, they do not seem to change substantially the price at which rooms are rented, for the cases in which a rental is found.

As a result of this analysis, we conclude that when analyzing such a system, measuring average effects may be quite insufficient, and it is necessary to consider each stage in the process, each algorithm, and each sub-group of people.

Chapter 6

DETECTING AND MITIGATING BIASES IN CLASSIFICATION

6.1 Introduction

6.1.1 Motivation

On a daily basis, many people are increasingly using social media platforms to share their feelings and moods. This creates a unique opportunity to proactively identify linguistic patterns that correlate with mental disorders [163, 174]. Early risk prediction of depression and anorexia [129] and suicide risk-assessment [221] are just some examples of different initiatives which have fostered the research on the interaction between language and mental health disorders on online social media and the application of ML to address such challenge.

However, as ML becomes more pervasive in sensitive domains, special care should be paid to a recent issue that has drawn scholars' attention: algorithmic bias. The great success of ML algorithms resides in their ability to indiscriminately learn latent nuances in the input data, even if they are not explicitly instructed to do so. Yet, human data encodes human biases by default [44] and therefore, these algorithms are prone to replicate and even amplify such biases in their outcomes, leading to unfair decisions.

In the context of risk-assessment and decision-making systems, fairness is defined as the “absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics” [138]. Hence, an algorithm whose decisions are skewed towards a particular group of people, often a certain minority, is considered unfair. In recent years, several cases have been identified as examples of inequalities created or amplified by AI-based systems. Such systems are trained by using data extracted from society, often reflecting several stereotypes in the form of biases. Examples of how data biases are then re-

flected in the predictions of automated systems can be found in the case of COMPAS, a recidivism-prediction tool used in the U.S., where ProPublica identified a much higher false positive rate for black people [109]; XING, a job platform that was reported to rank less qualified male candidates higher than more qualified female candidates [120]; or the case of face recognition online services found to suffer from achieving much lower accuracy on females with darker skin color [40]. These biases can be particularly harmful when dealing with health-related data. As argued by Walsh et al. [210], health disparities contribute to algorithmic bias. For example, women have a higher prevalence of major depression and anxiety disorders [127]. Prevailing societal notions about several groups' susceptibility to mental disorders contribute to incorporating bias in the underlying data and model specification. Furthermore, this issue, along with other factors, might prevent most of the risk-assessment and decision-making technological developments from ever being used in real-life setting [210].

6.1.2 Research questions

In this context, we define the following research questions that guide our experimental setup, analysis and contributions:

1. RQ1: To what extent Machine Learning based predictive models exhibit performance disparities across Anorexia Nervosa (AN) demographic groups?
2. RQ2: What are the causes of the existing biases when assessing AN on social media using Machine Learning algorithms?
3. RQ3: How can we mitigate the aforementioned biases?

6.1.3 Contributions

In this chapter, we perform an exploratory analysis that considers algorithmic fairness in characterizing eating disorders on social networks. In particular, we study a classification problem where an automated system predicts, given a set of posts authored by a user on a social media platform, whether such a person might suffer from an eating disorder (positive prediction) or not (negative prediction). We observe, quantify and characterize different types of inequalities that are leading to an unfair predictive system.

Our results show how a ML model trained on the collected data exhibits relevant biases in the form of higher False Negative Rate (fraction of false negatives between the number of positives, FNR) for females if compared to the performance obtained by male users. In the context of AN diagnosing on online forums,

False Negatives (FN) require special consideration, as they might cause the professionals to not detect such cases, which could lead afterwards to an omission of treatment.

Although the dataset considered in our experiments contained a higher proportion of female examples, the model was not able to accurately separate the positive from the negative examples equally well for both genders. In particular, the overall performance (accuracy) obtained for males was significantly higher (~10%). This is opposed to the fact that male cases tend to be more difficult to diagnose in practice [75].

Given that the set of characteristics contained in the dataset were mostly based on linguistic features, the root cause of this effect might rely on the fact that body dissatisfaction tends to be higher in women than in men [97]. Thus, it is expected for women to be more open and talkative about diets and food-related conversations, regardless of the fact that they suffer an eating disorder or not. However, men usually do not talk about these concepts. This causes that both positive and negative examples contain similar distributions for certain features in the female case, whereas those distributions are strictly different for the male case.

To the best of our knowledge, this is the first work that applies algorithmic fairness in detecting eating disorders on social networks. This contribution focuses on how inequalities can be easily created and/or amplified by predictive systems in the domain of eating disorders. Additionally, we use state-of-the-art techniques to mitigate detected biases, showing that finding a solution that removes all existing biases becomes a very arduous task. We conclude that the preference for one solution over the other might depend on the concrete application system.

6.1.4 Chapter structure

The remainder of this chapter is structured as follows: Section 6.2 reviews previous works related to characterization of mental disorders on social media and Algorithmic Fairness; Section 6.3 describes the proposed methods for bias detection, characterization and mitigation; on Section 6.4 we discuss the main results found, while Section 6.5 closes the chapter with additional discussion pointing the limitations and ethical concerns of this research.

6.2 Related work

6.2.1 Characterization and assessment of mental disorders on social media

Traditionally, mental health practitioners have collected and integrated information from various instruments to characterize the mental state of individuals [196]. These include direct observation, focused questions on the current symptoms and formalized psychological tests. Such instruments have been used to assess several mental-related variables, such as the appearance, mood and attitudes of the subjects to determine the presence of any irregularity. The proliferation of online social media platforms is changing the dynamics in which mental health state assessment is performed (Skaik, 2020, Chancellor, 2020, [169]). Individuals are using these platforms on a daily basis to share their thoughts as well as to disclose their feelings and moods (Prieto, 2014, Coppersmith, 2014). These platforms have become promising means to detect different mental health disorders since the language used as well as the emotions expressed in the text (e.g., social media posts) and shared with followers or friends on a daily basis may pinpoint feelings like worthlessness, guilt, or helplessness. This can provide a characterization of symptoms of psychological disorders, such as Anorexia Nervosa. In this regard, [163] characterizes different stages of Anorexia Nervosa on Spanish-speaking Twitter users by combining the analysis of text, images and social interactions.

6.2.2 Algorithmic fairness for detecting mental health status

Reduced research work has been conducted regarding the intersection between algorithmic fairness and the automated detection of mental disorders. Although, this has an increasing interest especially for social media platforms or in scenarios where users give their consent to be tracked on social media for health monitoring (schools or medical centers). In particular, (Chancellor et al., 2019) highlight the existence of methodological concerns of data collection and bias related to the application of ML methods to infer mental health status. In general, discussions of consent, validity, underlying bias from data collection techniques, or ML model selection is very limited. Moreover, the outcome of such algorithms, perpetuating unintended biases might lead to negative and discriminatory repercussions.

In this respect, (Straw and Callison-Burch, 2020) conducted a literature review over fifty-two articles that address the use of Natural Language Processing (NLP) in mental health across multiple disciplinary databases, and explore each stage of AI models' development to analyze which and how biases arise. The literature review found that no studies stratified the outputs of their NLP models by demographic features. Moreover, they performed an analysis of biases in

word embeddings that relate to mental health by evaluating GloVe and Word2Vec pretrained embeddings. They explore the relationships within word embeddings using analogy completion approaches to compare demographic labels and psychiatric terms (e.g. man is to depression, as woman is to "perinatal depression").

More recently, (Aguirre et al., 2021) explored the susceptibility to gender and racial biases of different computational methods for the automatic assessment of depression. In particular, they focused on the detection and mitigation of such demographic biases analyzing two widely-used datasets for the study of depression on social media: CLPsych (Coppersmith et al., 2015) and MULTITASK (Benton et al., 2017). They considered four demographic groups and two genders. Outcomes from their study revealed that existing datasets are not demographically representative and, without accounting for this, depression classifiers performed worse on people of color, specifically female in CLPsych and male in MULTITASK. Both groups are underrepresented in the datasets. Finally, they provided a series of recommendations on how to avoid such biases in future research using these datasets.

Prior work differs from ours as we present the analysis and characterization of gender related biases regarding a particular use case, which consists in a task addressed to the detection of Anorexia Nervosa on Twitter. For our predictive models, we take into account several features and propose approaches to address biases by applying fairness assessment approaches.

6.2.3 Bias mitigation in classification

In this section, we review the state-of-the-art bias mitigation algorithms for classification problems.

Existing methods to mitigate biases in ML models fall under three categories: (i) Pre-processing. Pre-processing methods modify the input data with the objective of reducing input data biases that might lead to performance disparities. (ii) In-processing. In-processing techniques modify the learning algorithm to incorporate fairness constraints. (iii) Post-processing. Post-processing approaches treat the ML model as a black-box and modify its outputs to achieve fairer outcomes.

In particular, in next chapter, we use as baseline a Logistic Regression model (LogReg) model, identified as the model with a better trade-off between balanced accuracy and FNR ratio. Such a baseline is compared with the effects of applying two pre-processing algorithms named Optimized Pre-processing [45] with repair level 0.85 and Reweighting [110]. Additionally, we tested a post-processing algorithm named Calibrated Equalized Odds [156] optimizing for generalized false negative rates.

6.2.4 Training calibrated classifiers

Previous work [156] analyzed the trade-off between minimizing error disparities across population groups while maintaining calibrated probability estimates. Obtaining calibrated probability estimates is considered crucial for empirical risk analysis tools [187].

Models calibration is often considered in algorithmic fairness analysis, as in the case that there is a disparity of calibration between population groups, a decision maker may be inclined to take the predictions less seriously for the group that lacks calibration [114].

When the classifier predictions are properly calibrated, its output can be directly used as a probability. It requires that for each classifier output range, the proportion of samples that actually have the true label is equivalent to the output value. For example, if a given (binary) classifier is properly calibrated, a prediction score of 0.2 for a given sample would require it to have a 20% chance of belonging to the positive class, 0.5 would require a 50% chance of belonging to the positive class and so forth.

In the task of detecting Anorexia Nervosa from online traces, certain use cases such as giving treatment priority to higher risk cases, would also require the use of a continuous output, i.e. predicting values in the range [0,1] so that those predicted with higher values can be used to prioritize treatment for those cases that are in a higher risk or have higher probability of suffering the disorder.

Additionally, comparing calibration across demographic groups can be used to adapt the decision threshold individually for each demographic group, so conditional probabilities of obtaining false negatives are equalized between them.

To obtain calibrated classifiers, we compare the performance of state-of-the-art calibrators: Isotonic and Sigmoid calibrators, which can be understood as regressors that map input values to new projected values in the same range [0,1] forming a new distribution where the obtained scores are equivalent to the actual chances of being a positive sample.

To train each calibrator, we use 5-fold cross-validation to ensure correct generalization of the obtained results. For each data split, the predictions of the trained model are used to fit an instance of each calibrator. To obtain calibrated predictions, model predictions are then transformed into the average of the 5 trained calibrators.

6.3 Methods

6.3.1 Dataset

We used the dataset collected by [163] for characterizing Anorexia Nervosa on Social Media, in particular, on Twitter. This dataset consists of publications corresponding to a one-year period between 2017-12-21 and 2018-12-21. The metadata elements and texts extracted passed through a strict transformation process in order to build and store vector representations of the features of interest at user level, guaranteeing the analysis of fully anonymized data.

Each element in the dataset (user) is independently annotated by a group of psychologists, psychiatrists, and therapists within one of the following classes: 1) Anorexia Nervosa (positive) - 177 users that manifest signs and symptoms of Anorexia Nervosa on their texts or if they explicitly state that they have been diagnosed with Anorexia Nervosa and/or are in treatment. This includes users at the precontemplation, contemplation and treatment stages according to the Transtheoretical Model [159]; 2) control - 326 users that do not make use of terms related to Anorexia Nervosa or users that use terms related to the disorder, but they do not manifest signs of anorexia. Table 6.1 reports the number of positive and control cases, split by gender. We consider just male and female, discarding those users corresponding to organizations that were also included in the original dataset.

Positive	Control	
No. samples	177	326
Female	127	157
Male	50	169

Table 6.1: Base rates for each class and gender in the dataset.

The dataset includes more than a hundred features built and inferred based on the text, images and metadata of the users' tweets. We discarded those extracted from images, as they were not present in all the users, as most of them tweeted text with no images attached. The detailed description of all the features included in the dataset can be read in the original paper [163], and they can be clustered in the following groups (see Table 6.2).

In the next sections, we describe the methodology used to detect and quantify biases on models trained on the Anorexia Nervosa dataset towards answering the RQs posed in the introduction section.

Types of features	Description
Content shared and interests	Linguistic dimensions Affective processes and emotions Personal concerns Risk factors vocabulary Anorexia related vocabulary Topics of interest Proportion of Anorexia Nervosa related tweets
Social network	Measures of interactions and engagement Analysis of followers and communities detection Analysis of interests between users and their followers
Behavioral aspects	Activity on a daily, weekly and monthly basis Sleep period tweeting ratio
Demographics	Gender Age

Table 6.2: Types of features included in the selected dataset.

6.3.2 Bias detection

In order to answer RQ1, we evaluate two scenarios: (a) the first one corresponds to the most typical case, when a unique model is trained for both genders and used to make predictions for all samples; (b) in the second scenario we train an individual model for each gender. We use this approach to evaluate whether this might have a significant impact on the final results.

To demonstrate that the observed behavior does not specifically depend on the usage of a certain category of classifiers, we compare a variety of models commonly used for the task [169] : (i) Logistic Regression, (ii) Random Forest, (iii) Support Vector Machines with different kernels, (iv) Multilayer Perceptron and (v) Ada Boost.

For testing the models, we partitioned the dataset between training and test using a cross-validation strategy based on 5-folds. For each of these data partitions, we trained a classifier using the training set and evaluated the observed performance for the test set.

The proposed methodology allows generalizing results on multiple data partitions and different models.

Biases are measured in terms of balanced accuracy (bAcc) and false negative rate ratios (FNR) between samples of different gender. FNR is related to the criteria of sufficiency [19] and requires a fair model to have similar false negative rates across demographic groups. The balanced accuracy metric is generally preferable in scenarios where data is not well-balanced, as it is the case in the collected

dataset:

- bAcc normalizes true positive rate (TPR), also known as recall, and true negative rate predictions (TNR) by the number of positive and negative samples, respectively, and divides their sum by two. True positive rate (TPR) and True negative rate (TNR) measure respectively the fraction of correctly detected positives and negatives between the total number of them:

$$\text{bAcc} = \frac{\text{TPR} + \text{TNR}}{2}$$

- FNR quantifies the fraction of false negatives (FN) between the number of positives (P):

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP}$$

Finally, we measure the following ratios (values closer to one indicate less biased predictions):

- $\text{bAcc}_{ratio} = \frac{\text{bAcc}_{female}}{\text{bAcc}_{male}}$
- $\text{FNR}_{ratio} = \frac{\text{FNR}_{female}}{\text{FNR}_{male}}$

6.3.3 Bias characterization

With the purpose of investigating the causes of the algorithmic bias when assessing AN on social media, we studied the features considered as input for the predictive models to identify which of those variables are more predictive for each gender (see Table 6.2).

As described by [163], the collected dataset was annotated by up to 5 human experts and the final label was decided based on the agreement of at least 3 annotators. For this analysis, the assigned labels were simplified to two classes: control and anorexia, with doubtful cases assigned to control.

Following the procedure described by Shing H.C et al. [183] we evaluated the performance of the individual human labelers with respect to the obtained ground truth in terms of balanced accuracy (bAcc) and false negative rate (FNR) ratios between female and male samples.

A different perspective on Bias characterization for this problem can be found in the PhD thesis on mental health and social media [164] section 5.5.3.

6.3.4 Bias Mitigation

In this section, we assess the effect of state-of-the-art bias mitigation algorithms applied to the use case studied in this work in order to answer RQ2.

Training fair classifiers

Existing methods to mitigate biases in ML models fall under three categories: (i) Pre-processing. Pre-processing methods modify the input data with the objective of reducing input data biases that might lead to performance disparities. (ii) In-processing. In-processing techniques modify the learning algorithm to incorporate fairness constraints. (iii) Post-processing. Post-processing approaches treat the ML model as a black-box and modify its outputs to achieve fairer outcomes.

In particular, we use as baseline a Logistic Regression model, identified as the model with a better trade-off between balanced accuracy and FNR ratio. Such a baseline is compared with the effects of applying two pre-processing algorithms named Optimized Pre-processing [45] with repair level 0.85 and Reweighting [110]. Additionally, we tested a post-processing algorithm named Calibrated Equalized Odds [156] optimizing for generalized false negative rates.

Training calibrated classifiers

Previous work [156] analyzed the trade-off between minimizing error disparities across population groups while maintaining calibrated probability estimates. Obtaining calibrated probability estimates is considered crucial for empirical risk analysis tools [187].

Models calibration is often considered in algorithmic fairness analysis, as in the case that there is a disparity of calibration between population groups, a decision maker may be inclined to take the predictions less seriously for the group that lacks calibration [114].

When the classifier predictions are properly calibrated, its output can be directly used as a probability. It requires that for each classifier output range, the proportion of samples that actually have the true label is equivalent to the output value. For example, if a given (binary) classifier is properly calibrated, a prediction score of 0.2 for a given sample would require it to have a 20% chance of belonging to the positive class, 0.5 would require a 50% chance of belonging to the positive class and so forth.

In the task of detecting Anorexia Nervosa from online traces, certain use cases such as giving treatment priority to higher risk cases, would also require the use of a continuous output, i.e. predicting values in the range [0,1] so that those predicted with higher values can be used to prioritize treatment for those cases that are in a higher risk or have higher probability of suffering the disorder.

Additionally, comparing calibration across demographic groups can be used to adapt the decision threshold individually for each demographic group, so conditional probabilities of obtaining false negatives are equalized between them.

To obtain calibrated classifiers, we compare the performance of state-of-the-

art calibrators¹: Isotonic and Sigmoid calibrators, which can be understood as regressors that map input values to new projected values in the same range [0,1] forming a new distribution where the obtained scores are equivalent to the actual chances of being a positive sample.

6.4 Results

In this section, we aim to answer RQ1, RQ2 and RQ3 following the methodology described in the previous section.

6.4.1 Bias Detection

In order to know if ML-based predictive models exhibit performance disparities across AN demographic groups (RQ1), we train and evaluate different estimators for assessing the risk of AN on male and female samples. Performance is measured using FNR and bAcc ratios. As stated in previous sections, we compare two different scenarios: (a) a single model trained for both genders; (b) an individual model trained for each gender separately, data is averaged with the same weight across genders.

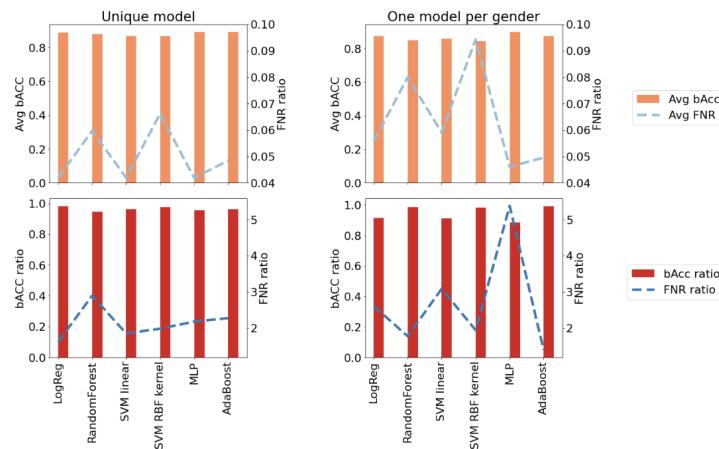


Figure 6.1: (best seen in color) average bAcc and FNR compared to bAcc ratio and FNR ratio across genders on the trained models. Figures on the left show scenario (a) -unique model- and figures on the right show scenario (b) -one model per gender-.

¹We used the implementation available in Scikit-learn: <https://scikit-learn.org/stable/modules/calibration.html>

When analyzing scenario (a), the results on performance disparities show that trained models obtain a good level of balanced accuracy (all of them greater than 87

Nevertheless, we observe important performance differences when decomposing such values by gender, yielding in the worst case (MLP classifier, one model per gender) an accuracy of 0.913 for male samples and 0.844 for female samples (relative difference of 8

Additionally, most of the models show around twice the false negative rates for female samples when compared to male ones (dotted blue line in Figure 1). Differences of performance are even more dramatic when using the second scenario, where an individual model is trained for each gender, incrementing the false negative rate differences up to 500

Although applying the second scenario might imply a disparate treatment by gender, which is protected by law in multiple countries, it was an interesting exercise producing mostly counter-intuitive results. The comparison of results shown in figure also prove that including the male samples in the training set benefits the performance obtained by females.

Summarizing the results obtained, we were able to achieve high accuracy models in both scenarios, but the performance was always lower for female samples. An error analysis points out that females have higher rates of false negatives, which is extremely dangerous in this context, since a false negative could lead to a lack of detection and therefore to a denial of treatment. Disparities of performance are reduced when a unique model is trained for both genders. Conversely, using a different model per gender leads to higher disparities in performance, increasing even more the lack of performance for females.

In the remainder of this paper we use the Logistic Regression (LogReg) classifier, using a unique model for both genders as a baseline. The choice is motivated by the fact that such a classifier shows the best balance accuracy and false negative rate ratios while keeping an average accuracy greater than 87

These results motivate the rest of this article. As observed female samples obtain not only lower performance, but also when the models make a misprediction, it is almost twice more probable that it has the form of a false negative for female samples.

6.4.2 Bias Characterization

The results obtained by Ramírez-Cifuentes [164], in this section can be summarized as the selected features identified to be not equally predictive across genders. This implies a difference of separability between male and female samples, with the features being more predictive for males than for females. This provides one of the answers to RQ2.

Additionally, we analyzed the labelers' performance for both genders by considering the final labels present in the dataset as the ground truth.

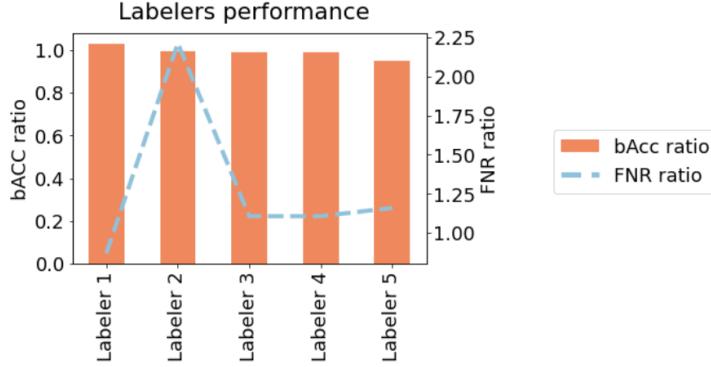


Figure 6.2: Labelers performance with respect to obtained ground truth.

As shown in Figure 6.2, the obtained results show that labelers had greater performance for male samples, with higher accuracy and lower FNR for them. The only exception between the 5 experts is Labeler 1, who showed higher accuracy for female samples and higher FNR for male ones. Additionally, we evaluated the Cohen's Kappa agreement between each pair of labelers for female and male samples separately, with an average of 0.807 versus 0.841 respectively, which could suggest that diagnosing male samples could be easier for human experts.

Interestingly, we see that most of the labelers had better performance when detecting anorexia nervosa in online traces. However, females tend to be easier to diagnose than males [75] during in-person consultations.

Understanding the performance of the annotators as an upper-bound of the performance that an autonomous system can have for this dataset, we observed that annotators have a certain level of bias (quantified in terms of performance differences across genders) that afterwards seems to be not mitigated if not amplified by the addition of the ML systems.

6.5 Bias mitigation

In this section, we test state-of-the-art techniques to obtain fair classifiers. We will use the Logistic Regression classifier, as it was the one that showed fewer disparities in performance while keeping a high level of accuracy (Figure 6.1).

In particular, we evaluate two different use cases: The first one will be described in section 6.5.1 and corresponds to the case where the model will be used to predict whether individuals might have the disorder or not. We will compare the model that has better overall accuracy and the model that obtains lower FNR

Table 6.3: Obtained performance in terms of Accuracy and FNR across genders for each bias mitigation technique.

Technique	Avg Accuracy Female	Avg Accuracy Male	FNR Female	FNR Male	F1 score Female	F1 score Male
Original classifier (LogReg)	0.812	0.916	0.082	0.005	0.844	0.947
Disparate Impact Remover	0.836	0.955	0.097	0.011	0.860	0.970
Reweighting	0.799	0.932	0.089	0.010	0.833	0.956
Calibrated Equalized Odds	0.812	0.916	0.082	0.005	0.844	0.947

inequalities by gender. The best models in terms of accuracy tend to be the ones that maximize FNR differences across genders, reducing their suitability for this use case from the point of view of algorithmic fairness. Afterwards, we evaluate a second scenario, where the predictive models will be used to sort the list of patients to be analyzed. First, it requires the prediction to be continuous, but additionally, for the individuals to be sorted, the predictions must also have a probabilistic approach, so we know that individuals predicted with a 0.8 really have an 80% of being a true positive.

6.5.1 Training fair classifiers

Results are calculated using 4-fold cross-validation, splitting the data between training, validation and test. Methods that require a validation set for adjusting parameters use the validation set, and all of them are evaluated on the test set. Results reported in Table 6 are the average of the four executions.

Table 6.3 shows the results of applying the bias mitigation algorithms, where it can be observed first that all methods lead to better or equal FNR ratios than the observed with the original classifier.

As can be seen in the table, the best bias mitigation results regarding FNR are achieved with the Reweighting pre-processing algorithm with a little improvement with respect to the performance observed for the Disparate Impact Remover algorithm (see Section 6.2.4) for a brief explanation of these algorithms). However, the improvement in terms of FNR ratio is done at the cost of incrementing the FNR for both genders. In case we aim to obtain a model that minimizes the balanced accuracy ratio, the original model and the Calibrated Equalized Odds post-processing, are reporting similar values with a bAcc ratio of 0.892.

Results showed that even when disparities can be mitigated with most of the algorithms, they can't be eliminated. Additionally, some methods failed to provide a result that was substantially better than the obtained with the original classifier without any transformation.

6.5.2 Training calibrated classifiers

In this section, we analyze and compare the calibration of the different ML models trained in our experiments and apply state-of-the-art calibrators to obtain properly calibrated models after post-processing their outputs.

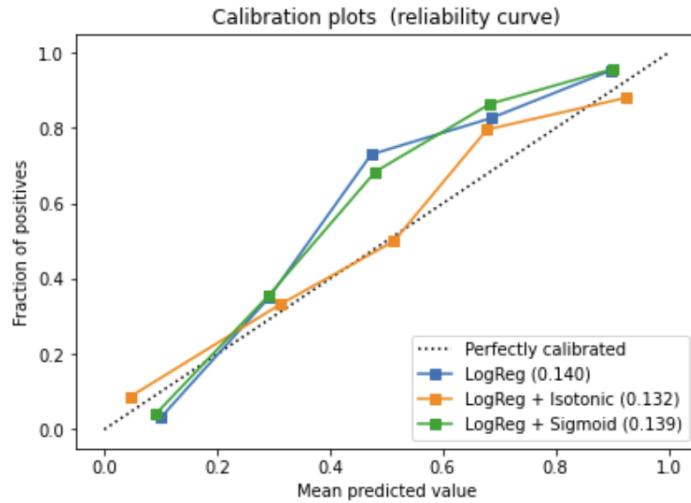


Figure 6.3: Calibration curves obtained for original model and calibrators

The calibration results depicted in Figure 6.3 show that the Isotonic calibrator obtains a calibration curve that is generally closer to the objective function (dotted line) for most of the thresholds.

The decision of which option is more suitable for production depends on the concrete application. Selecting the threshold that is more suitable for a given use case is generally done based on economic reasons, as occurs in many other public policy problems [77].

As an example of a calibration problem, suppose a scenario where all the samples predicted with a value greater than a certain threshold would be treated in a medical consultation by a practitioner. Consider a population of 100 individuals, a total budget of 1000€ and a cost of 33€ for each treated individual. This would allow up to 30 patients (30% of the population) to be considered for diagnosis in a medical consultation. In such a case, we would use the threshold 0.7, so only individuals with a prediction score greater than 0.7 are considered.

Using a perfectly calibrated model, we would expect 21 true positives (70% of the 30 individuals considered for diagnosis). In this case, given that the calibration curves depicted in Figure 7 show very similar calibration values for all the calibrators at such a threshold, there is not a strong preference for using one among the others.

Nevertheless, if we consider the same scenario but with a total budget of 1650€, we would be able to check up to 50% of the population. Therefore, we will use the 0.5 threshold, where the Logistic Regression classifier with the Isotonic calibrator shows a clearly better performance making it to be the preferred choice. Using such a classifier, we would be able to detect up to 25 true positives between the 50 individuals selected for medical consultation.

From these results we conclude that even when the LogReg with the Isotonic calibrator is better calibrated better in general, certain use cases would make other options more or equally preferable to the LogReg with Isotonic calibrator.

6.6 Discussion

Lastly, social media has been offering a new space for mental health assessment. The explosive use of such platforms, especially by young people, raised questions about the potential negative impact of these applications on the mental health of the most vulnerable ones. Particularly, users with Anorexia Nervosa (AN) represent a very tight online community that have even created their own vocabulary to be identified among themselves with the goal of exchanging very unhealthy tips. In this regard, new research has been developed on automatically detecting users at risk by analyzing their online activity using artificial intelligence. These studies usually work on datasets containing data from, mostly, female users. Thus, we would expect that these algorithms would perform better when classifying females than males, as they are provided with more female instances.

However, we have attested that ML models exhibit gender bias on assessing AN, as they produce higher False Negative Rate for females (RQ1). This might cause wrong and late diagnosis that can be extremely harmful.

We later experimented on characterizing this bias (RQ2), by analyzing the most relevant features selected by our models for assessing female and male users, separately, and comparing these features with those selected by clinicians when classifying risk of AN just based on the writings of the users. We could attest that biological processes and suicide risk factors are the key for a good precision in classifying positive AN cases in males, and age, emotions and personal concerns are more relevant for females, probably because they tend to express more their feelings on their posts. We could also confirm that automated models are not capable of identifying suicide risk factors that are described implicitly in text.

Finally, we proposed several techniques for bias mitigation (RQ3) and we could see that even when disparities can be mitigated with the new proposed algorithms, they can't be eliminated.

In conclusion, online assessment of mental health issues using automated methods needs more attention. Most of the state-of-the-art works in this regard

just measure global precision metrics without making any effort on detecting gender bias. Fairness should be considered when deploying automated systems that could affect the diagnosis of people at risk.

6.6.1 Limitations

It is important to note the methodological limitations present in the study of algorithmic fairness on social media conducted in this work.

First, we acknowledge that the findings and conclusions obtained are limited to a specific eating disorder known as Anorexia Nervosa. While other eating disorders, like bulimia nervosa or Eating Disorder Not Otherwise Specified (EDNOS), might share common symptoms with Anorexia Nervosa [7], the behavior and actions that characterize the individuals suffering from any of them is quite different. Therefore, it is expected that the manifestation through language use on social media might be different between these disorders, and hence also are the features which most accurately denote their presence and development. The exploration and comparison between the language expression on social media of these disorders is still an open problem [174].

Another limitation we observe with our study is related to the dataset. As stated by [163], demographics attributes were inferred using an automatic approach. As any computational method, this procedure is not free from error. Nonetheless, as the authors explained, the accuracy of the method was manually tested on a subsample of the dataset. A macro average accuracy of 0.84 for all the gender groups of all the classes and a macro average accuracy of 0.80 for all the age groups of all the classes was achieved. For this reason, we still consider this as a good approximation of the demographics attributes, which enables us to study the manifestation of bias and shed light on possible solutions.

Moreover, it should be noted that even though the annotation process was rigorously conducted by five domain experts (three psychologists and two psychiatrists), it might not be completely accurate. The practitioners involved made their judgments using only the textual content of the posts. In a real-life scenario, a diagnosis is elaborated based on a combination of direct and indirect assessment instruments, such as unstructured observation, specific questions regarding the manifesting symptoms, and formalized psychological tests. These elements allow practitioners to acquire a comprehensive cross-sectional characterization of the person’s mental health condition. Despite this limitation, we consider that given the background and practical experience of the professionals involved in the annotation process, labeling errors would be minimal and not influence the conclusions obtained for the whole sample of individuals analyzed in this work.

Finally, we should be aware that the conclusions drawn from the data are limited in scope to individuals who use social media, meaning it is probably a

younger and more technologically literate sample than the population as a whole [146]. Moreover, our study only included users who are active on Twitter and who choose to make their tweets publicly available. Therefore, the fairness assessment considering users from other social media platforms and even of people with Anorexia Nervosa who do not have accounts on any social platforms is out of our reach.

6.6.2 Ethical concerns

Research involving human beings concerns sensitive topics related to the ethics of the treatment of data and individual's privacy [50]. The sensitive nature of mental health research requires us to consider possible benefits of this study alongside its potential harms.

The potential immediate benefit of this study is a better understanding of gender bias in computational assessment of Anorexia Nervosa using social media data. A potential second benefit is the mitigation of the disparities observed which otherwise, as shown in this study, permeate into the assessment algorithms. In particular, we ascertained the extent to which fairer classifiers can be developed considering the trade-off with performance.

Nonetheless, we are aware of the potential harms from our work. Mental health status are sensitive personal attributes that could be used to maliciously target individuals on publicly-facing online platforms. Hence, as researchers working with social media data we have taken the necessary precautions to protect the privacy of individuals and their ethical rights to avoid any further psychological distress. We have followed the guidelines of Benton et al. [23] and Ayers et al. [13] on data use, storage, and distribution. All analysis was conducted on deidentified versions of data as all identifying metadata was either redacted or obfuscated to preserve the privacy of individuals in the dataset.

Chapter 7

UNCOVERING BIASES IN USER-MACHINE INTERACTION

7.1 Introduction

7.1.1 Motivation

Decision support systems based on artificial intelligence are increasingly being deployed in a variety of real-world scenarios [130, 31, 143, 109]. Typically, these deployments involve a moderate level of automation [54] in which a human is in charge of making the final decision based on some inputs, including a Decision Support System (DSS). Indeed, in some contexts, that a human makes the final decision when using a decision support system might increasingly become a legal mandate.¹

The process of incorporating the input from a DSS is a complex cognitive task, considering that in many domains machines are still far from achieving perfect accuracy. Users have to navigate between the traps of algorithmic aversion [64] and automation bias [54], respectively characterized by under- and over-reliance on the DSS. Several authors have studied and tested methods to increase user trust in machine predictions [15]; however, eliciting more trust might not be the best way to combine human and machine intelligence. Arguably, the ideal scenario is one in which users follow the advice of the DSS when it is correct and ignore it when it is wrong.

We study how key characteristics of a decision support system impact the decisions made.

To do this, we designed an online game, depicted in Figure 7.1. It consists of a 32x32 board representing a map having green (“forest”) and brown (“desert”)

¹See, e.g., Article 22 of the EU General Data Protection Regulation.



Figure 7.1: Experiment interface. The terrain is represented by green (forest) and brown (desert) cells. The user has drilled in the cells in black, and a recommendation from the DSS (in yellow) is shown.

cells. Players have to “drill” for oil by clicking on a cell, and obtain a score equal to the oil yield of that cell, which is revealed only after clicking, minus the “environmental cost,” which is zero for desert cells, but non-zero for forest cells. The hidden oil profile is independent of the terrain map, and is such that neighboring cells offer a similar reward. The goal of the participants is to maximize the score after a series of 25 rounds (clicks) in each of three maps having different levels of difficulty. Participants in our control group do not receive any assistance, while those in various treatment groups are assisted by decision support systems of varying levels of accuracy and bias.

7.1.2 Research question

Our main research question is *how do the characteristics of a decision support system impact human performance, time to completion, and reliance?* We address this research question experimentally.

7.1.3 Contributions

We release the design and implementation of a platform to study the stated research question. As a limitation (discussed on §7.7.1), the platform does not simulate a high-stakes scenario. However, it allows studying common elements of human interaction with a DSS, as it does not require prior knowledge from participants. In this platform, researchers can vary the problem difficulty and the accuracy of the decision support system, and introduce bias in the recommendations

in a manner that is visible to participants. The platform records all interactions, generating a wealth of data including player’s performance, timing, and reliance on the decision support system both implicitly (by observing clicks) and explicitly (by an exit survey).

We describe a series of experiments, approved by our Ethics Review Board, and involving over 400 participants recruited via crowdsourcing. These experiments uncover what we consider mostly rational behavior. For instance, participants rely on the decision support system to an extent that is well-aligned with its accuracy.

7.1.4 Chapter structure

This chapter is organized as follows: We overview related work in section 7.2 before describing the methodology in Section 7.3 and experimental design in Section 7.4. Then, we review the obtained results Section 7.5 and discuss the general findings Section 7.6. Finally, we discuss the limitations of our approach in Section 7.7.

7.2 Related work

The literature on decision support is vast; in this section, we overview research on Decision Support Systems (DSS) that provides context for our work (subsection 7.2.1), particularly research on trust and reliance on DSS (subsection 7.2.2), and on the influence of DSS accuracy on reliance (subsection 7.2.3).

7.2.1 Decision Support Systems (DSS)

Decision-making is an essential activity that involves facing choices, often in the presence of uncertainty [38]. It is also a complex cognitive process that depends on interpreting large amounts of information, evaluating the possible consequences of the decision to be made [54]. The goals of deploying a DSS to assist human decision-making typically include improving the quality or accuracy of decisions, reducing subjectivity, reducing costs, and increasing the efficiency of a decision-making process [89, 125].

An important characteristic of a DSS is its degree of autonomy with respect to human decision makers [54]. A high level of automation, e.g., a DSS that can automatically implement its recommendations, can be useful in scenarios where the workload is high, and the DSS can correctly make decisions in a reliable manner, helping to reduce workload [15]. A lower level of automation, e.g., a DSS that only recommends a choice but does not act upon it, might be less useful in some

scenarios, but can also help human decision-makers detect failures or errors in the DSS [195].

One of the most challenging aspects of implementing a DSS is that humans embed values, biases, and assumptions in their decision-making without acknowledging the ambiguity, incompleteness, and uncertainty that are part of this process. Hence, there can be misalignment or complete lack of alignment between the recommendations of a DSS and what humans would choose [30, 198].

7.2.2 Trust and Reliance on Decision Support

Multiple studies from different disciplines have been carried out to understand which factors affect human trust and reliance on a DSS. Some controlled user studies indicate that participants tend to exhibit *automation bias*, i.e., a tendency to follow the DSS even in cases in which they could have made better decisions by ignoring it [66, 16, 193]. Other controlled user studies have uncovered *algorithmic aversion*, i.e., a tendency to distrust the recommendations of a DSS, or a rapid drop in confidence on an algorithmically-supported DSS after seeing it make a mistake, which would not have been of the same magnitude if the decision support were offered by a human [64].

In general, the trust that users place in an automated system is affected by contextual, cultural, and societal factors [126]. *Trust* in this case is a complex construct that depends on the interplay of the users' disposition to the system, the situation in which the interaction happens, and what the users can learn about the system. *Reliance*, on the other hand, is more narrowly defined as compliance with an automation's recommendation [98]. The extent to which trust determines reliance on a system is also subject to influences such as the complexity of the situation, its novelty, the degree of decisional freedom of the user, and whether s/he can compare the performance of automated and manual decisions [98].

In general, decision-makers consider the guidance of a DSS relative to the information context in which it is provided. Hence, they may deviate from the DSS suggestions for different reasons, including their own biases, preferences, and deviating objectives [106, 85, 133, 189]. For instance, some user studies suggest that experienced decision-makers in a given domain are less inclined to follow algorithmic suggestions, and rely more on their own cognitive processes [88, 157].

7.2.3 Effects of DSS Accuracy

While a DSS does not have to be perfectly accurate for it to be useful [3], better decisions can be made if decision-makers can rely more on a more accurate DSS than on a less accurate DSS. Some user studies indeed have shown that humans

rely on machine predictions more when they are correct than when they are incorrect [121]. However, in other user studies, participants have consistently followed incorrect recommendations even for tasks they perform well [193]; or have failed to correctly evaluate the accuracy of the DSS and their own accuracy, and hence have not been able to adapt their reliance on the DSS to its performance [87, 86]. In general, perceiving the accuracy of a DSS is easier when DSS errors are consistent and deterministic, and when there is a simple boundary separating cases in which the DSS is correct from cases in which the DSS is incorrect [17].

Inferences about the accuracy of a DSS might be influenced not only by the correctness of the recommendations, in cases where humans can to some extent directly observe correctness, but also by the information provided by the DSS [121]. For instance, simply stating that a DSS is accurate can increase reliance on it, up to a point; however, the effect of these statements is weaker than direct observation of correct recommendations by users [213]. Similarly, displaying a confidence score accompanying each recommendation or prediction of a DSS has been shown to increase willingness to rely on these recommendations, when the confidence score is high [220]; however, in other user studies even when informing users that the DSS has low confidence in a recommendation, users have followed it [193].

We build upon previous work by studying a common interaction sequence with a DSS: (1) the environment provides an input, (2) a DSS recommends an action, (3) the human makes a decision, and (4) the environment returns a reward [16]. We provide a simple yet expressive scenario for which no prior experience from the participants is required. With few exceptions (e.g., [193]) studies on the interaction of AI with human decision-making do not experiment with task difficulty, and in contrast with most previous work (e.g., [86, 128, 16, 66, 121, 213]), the systems we consider are less accurate than humans acting alone. To the best of our knowledge, ours is the first study that considers different levels of bias in the decision support.

7.3 Methodology

In this section, we describe the details of our methodology. We first present an overview of our methodology (Section 7.4) and the platform we designed to perform our experiments (Section 7.4.1); then we describe the independent and dependent variables we take into consideration in our experiments (Section 7.4.2).

7.4 Overview

Our methodology is experimental and based on a simple yet expressive game, depicted in Figure 7.1. The game, inspired by “Wildcat Wells” [136] and described in Section 7.4.1, has several characteristics that make it appropriate for this research, including that (i) it is simple, (ii) does not require training or prior experience, (iii) uses a random game generator, (iv) provides fine-grained control over game difficulty, and (v) naturally lends itself to decision support.

We first perform a simple experiment without any decision support system, in which we randomly generate a set of maps and select three maps that (according to the scores users obtain) are labeled respectively as easy, medium, and hard. Then, we experiment with these three maps by providing machine assistance in the form of a recommendation on where to click next, varying parameters such as the accuracy of the decision support or the amount of observable bias it might have. We also include an exit survey in which we ask questions related to algorithmic reliance.

Participants are recruited through a crowdsourcing platform specialized in research² and paid above the platform-recommended fee of 7.5 GBP per hour of work.

7.4.1 Platform

This section describes the platform we designed and implemented to conduct experiments.³

User interface. The interface is a web application composed of five screens: (i) informed consent, (ii) demographic questions, (iii) tutorial, (iv) game, and (v) survey. The informed consent form explains the purpose, duration, risks and benefits of the experiment, and asks for explicit consent to participate. Then, users are asked optional socio-demographic questions: (i) gender including male, female, and other; (ii) age bracket in five years increments; (iii) country of residence; (iv) level of education; and (v) professional background. Then, a short tutorial is shown to explain the gameplay.

The core portion of the experiment is the game, which consists of three maps shown in random ordering. Each of the three maps has a different level of difficulty: “easy”, “medium”, or “hard”. These maps were selected from a collection of candidate maps generated using Perlin-noise [155] random generators for the terrain and oil profile. The selection was performed through a preliminary exper-

²Prolific (<https://prolific.co/>)

³All the code of the experimental platform will be released as free/open-source software with the camera-ready version of this paper.

iment in which participants were asked to play the game without any assistance, as we explain in Section 7.5.

In each 32x32 map, the game proceeds in 25 rounds, i.e., allowing the user to click on 25 of the 1,024 cells. The total score of a user is income minus costs. The income is the combined “oil” yield of the 25 selected cells, which remains hidden until the cell is clicked. The cost is a fixed “environmental cost” multiplied by the number of “forest” cells that are drilled; “desert” cells have zero cost.

During the game, the platform records the time required to complete the tutorial and each game, as well as a timestamped record of all user interactions, including the recommendations that are generated and the cells that are clicked.

Decision support system.

To assist users when making the decision on where to click, we provide a Machine Learning (ML) based Decision Support System. The DSS is trained by performing a number of random “test drills” to try to reconstruct the hidden oil profile. The DSS uses an ML model based on a Lasso model fit with Least Angle Regression (LARS) [68]; it corresponds to a Linear Model trained with an L1 prior as regularizer. Two dimensions of the DSS that we control during the experiment are *accuracy* and *bias*.

Accuracy. We use a setting where three versions of the DSS are generated with respectively high, medium, and low accuracy. The high accuracy DSS is based on an ML model trained with 20 randomly selected points, and recommends a randomly chosen cell among those predicted to have a revenue in the top 20%. The medium and low accuracy DSS simply generate a high accuracy recommendation and add to it circular two-dimensional Gaussian noise with $\mu = 0$ and either $\sigma^2 = 3$ (small variance) or $\sigma^2 = 20$ (large variance). The units for σ are cells; remember each side of the map measures 32 cells. The medium accuracy DSS adds the small variance noise with probability 80% and the large variance noise with probability 20%. The low accuracy DSS adds the small variance noise with probability 20% and the large variance noise with probability 80%. These parameters are set experimentally through preliminary tests to induce a situation in which the performance of the model is not immediately obvious. Nevertheless, participants react to the accuracy of the DSS, as we describe in Section 7.6.

Bias. We experiment with two versions of the DSS, one providing biased predictions, and another one providing unbiased ones. To create the biased predictions in a manner that was visible by participants, we train the biased DSS to optimize only for income, i.e., ignoring the “environmental cost.” We communicate bias to users, when present, by stating that the recommendation takes into account only the oil yield, but disregards the costs. In the unbiased scenario, the DSS is trained to optimize income minus cost. In the low-bias scenario, the DSS

ignores (does not consider it during the learning phase) an environmental cost of 20% of the maximum oil yield. In the high-bias scenario, the DSS ignores a cost that is 40% of the maximum oil yield.

A sequence of recommendations is pre-computed so that users get similar recommendations during their games. The only difference they might experience is that recommendations do not suggest cells that have already been clicked.

7.4.2 Experiment variables

In this section, we describe the set of *independent* and *dependent* variables that we identified for our experiment. The first type of variables are controlled and changed during our experimentation in order to observe their effect on the dependent variables to be measured (e.g., the user’s performance). In general, any change in the independent variables may cause a change in the dependent variables.

Independent Variables. As independent variables we take into consideration three levels of map difficulty (easy, medium, or hard), and conditions with and without decision support. When decision support is present, we also consider its accuracy (low, medium, or high) and bias (absent, low, or high).

Dependent Variables. The key dependent variable is each participant’s *score*, which is computed by adding the income minus cost across the three maps. Additionally, we measure the *time to complete* the three maps, and we ask participants which of the three maps they perceive as the most difficult.

We measure *reliance* in two ways: implicit and explicit. Implicitly, we measure the distance between the selected cell and the provided recommendation for each play; we interpret a short distance as more reliance. Explicitly, we use a technology acceptance survey proposed by Hoffman et al. [99].

Technology Acceptance Survey Questions

The technology acceptance survey [99] contains the following questions:

1. *I am confident in the algorithm (DSS). I feel it works well.*
2. *The outputs of the algorithm are very predictable.*
3. *The tool is very reliable. I can count on it to be correct all the time.*
4. *I feel safe that when I rely on the algorithm I will get the right answers.*
5. *The algorithm is efficient in that it works very quickly.*
6. *I am wary (suspicious/distrustful) of the algorithm.*
7. *The app can perform the task better than a novice human user.*

8. *I like using the system for decision making.*

This survey has 8 questions that are answered on a Likert scale (1-5). The questions, which can be found in 7.4.2, address the confidence, trust, predictability, reliability, and safety of the DSS. A score close to the maximum (40 points) indicates a high level of acceptance of the DSS, while a score close to the minimum (0 points) indicates low acceptance.

7.5 Experimental Design

In this section we first describe the preliminary experiment used to select the three maps with different difficulty levels (Section 7.5); then, we provide the details on the setting of our main experiments (Section 7.5.1).

Map Selection Experiment

We used a crowdsourcing-based experiment to select three maps with different levels of difficulty. In particular, for this experiment, we generated 10 candidate maps composed of a terrain profile plus an oil yield profile. All maps were generated with equal parameters for the Perlin-noise generators [155]. In particular, we used the following parameters for the generator of each of the profiles: The terrain profile generator uses *octaves*= 9, *persistence*= 0.5 and *lacunarity*= 20, whereas the oil yield generator was parametrized with: *octaves*= 1, *persistence*= 1 and *lacunarity*= 1. This selection of parameters yields higher surface roughness for the terrain profiles and less surface roughness for the oil profiles.⁴

A total of 120 crowdsourcing workers, 15 per map, participated in this phase. While examining the gameplay traces (score per round), we noticed in some maps a high percentage of users that started with a high score. We called this class of participants “luckers” and used the proportion of these to decide which map to select; maps with a higher proportion of “luckers” correspond to overly simple maps.

We first selected the two maps with a lower percentage of “luckers”. The map with the lowest average score was considered ‘hard’ and the other “medium” in terms of difficulty. Among the rest of the maps, we selected an “easy” map at random among the ones in which users had the highest scores; this map has a single global optimum in the oil yield profile, and this optimum is located in a deserted area, which means users do not need to consider costs when “drilling”. Figure 7.2 depicts the terrain and oil yield of the three selected maps.

⁴Examples of maps generated with different ranges of parameters for the Perlin-noise generator will be available in our code release with the camera-ready version of this paper.

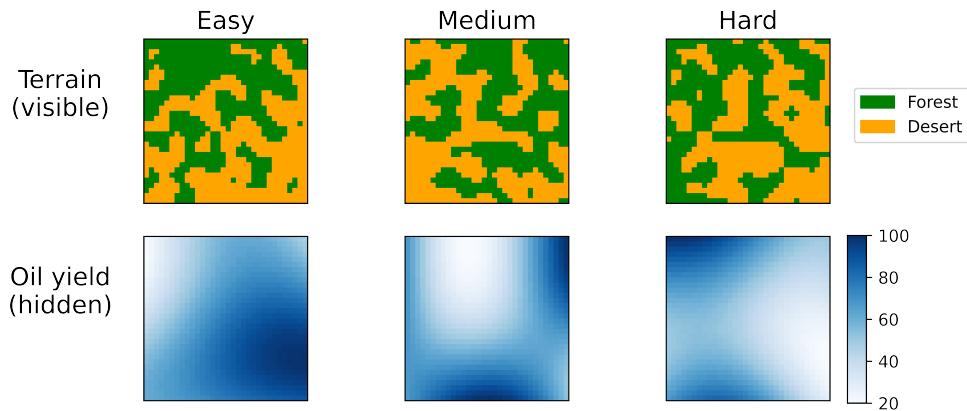


Figure 7.2: Selected easy, medium, and hard maps, displaying terrain (top) and oil yield distribution (bottom). The terrain is visible to participants; green cells represent forest, and yellow cells represent desert. The oil yield is hidden; darker shades indicate higher yield.

The distribution of scores that we observed was aligned with map difficulty, as we show in Figure 7.3 (Section 7.6.1). User perceptions of difficulty were also aligned with these choices, as in the exit survey when asked which one they considered to be the easiest map, the “easy” map was selected by 59% of participants, the “medium” by 23% of them, and the “hard” by the remaining 18%.

7.5.1 Main Experiments

Control group. The control group received no machine assistance. Users played the three maps in random order. In total, we gathered data from 27 control group participants, testing at least three times each of the six possible map orderings.

Treatment groups. All the treatment groups received help from a DSS. Here, we consider conditions combining a level of accuracy of the DSS (high, medium, or low) as described in Section 7.4.1, with the presence or absence of bias and the amount of potential bias, as described in Table 7.1. This produced 36 (6 possible map orderings \times 6 possible orderings of the DSS by levels of accuracy) possible experimental units; each one was completed by at least three different participants. In total, 435 participants played 1,305 games. Unique participants were recruited for each experiment separately.

Environmental cost	Biased DSS	Unbiased DSS
20 (low cost)	LB	LU
40 (high cost)	HB	HU

Table 7.1: Experimental conditions, with “L” representing low cost of drilling a forest cell, and “H” representing high cost. LB and HB correspond to cost-unaware decision support, which yields biased suggestions. LU and HU correspond to cost-aware decision support.

7.6 Results

In this section, we analyze under different experimental conditions the obtained score (Section 7.6.1), the time to complete the task (Section 7.6.2), and the reliance of participants on the DSS (Section 7.6.2).

7.6.1 Score

In this section, we observe how the score obtained by participants changes under various conditions.

Map Difficulty and Decision Support.

We compare the score obtained by participants with machine assistance with respect to the control group, to understand whether the presence of a DSS improves performance or not.

Figure 7.3 shows the score distribution in each of the three maps, with and without machine assistance. In this and the following figures, scores are presented per play (click), and the maximum score is 100, which is the maximum oil yield. Median scores per play obtained using the DSS in the easy, medium, and hard maps are respectively 88, 71, and 62 points. Without using the DSS, these scores are respectively 80, 66, and 61 points. Using a t-test we observe that the increase in score due to machine assistance is statistically significant at $p < 0.0001$ for the easy and medium map, and at $p < 0.05$ for the hard map.

Decision Support Quality and Bias.

We evaluate the performance obtained with each level of accuracy and condition of bias of the DSS, to understand the impact that these model characteristics might have on the obtained score.

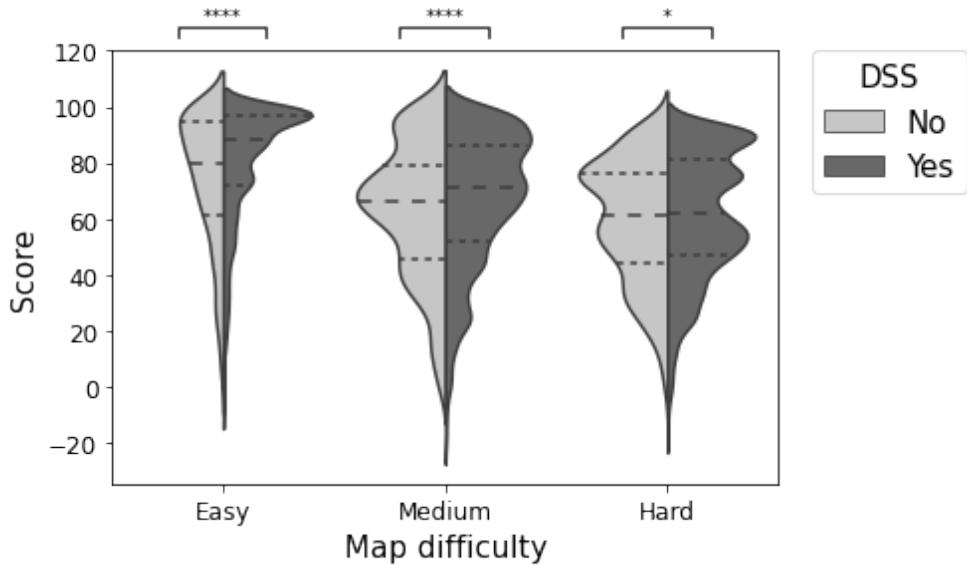


Figure 7.3: Distribution of scores in the three maps (easy, medium, hard), without machine assistance (left) and with machine assistance (right). In statistical significance tests, “ns” stands for no significance, and asterisks significance at: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$).

Figure 7.4 compares the distribution of scores that the DSS systems would obtain on their own (remember the DSS recommendations are randomized), against the scores obtained by participants with the help of the DSS. From the figure, it is evident that the accuracy of the decision support system impacts the score, with more accurate systems inducing a better score. Median scores per play obtained by participants across all three maps using the DSS are 70, 75, and 77 points using the low, medium, and high accuracy DSS respectively. These scores are higher than what this DSS would obtain on its own: 33, 48, and 74 points respectively. Differences are statistically significant at $p < 0.0001$ for the low and medium accuracy case, and at $p < 0.001$ for the high accuracy case.

Furthermore, participants on their own outperform the DSS in every map, as we mentioned they obtain 80, 66, and 61 points respectively for the easy, medium, and hard maps; in contrast, the median scores obtained by even the high-accuracy DSS are 69, 65, and 60 points respectively (figure omitted for brevity). The median scores obtained by the medium-accuracy DSS and low-accuracy DSS are even lower.

Next, we consider differences in performance due to *bias*. We introduced a bias that is visible to participants by providing a biased DSS that considers only the reward but not the cost. In contrast, the unbiased DSS considers the costs. We

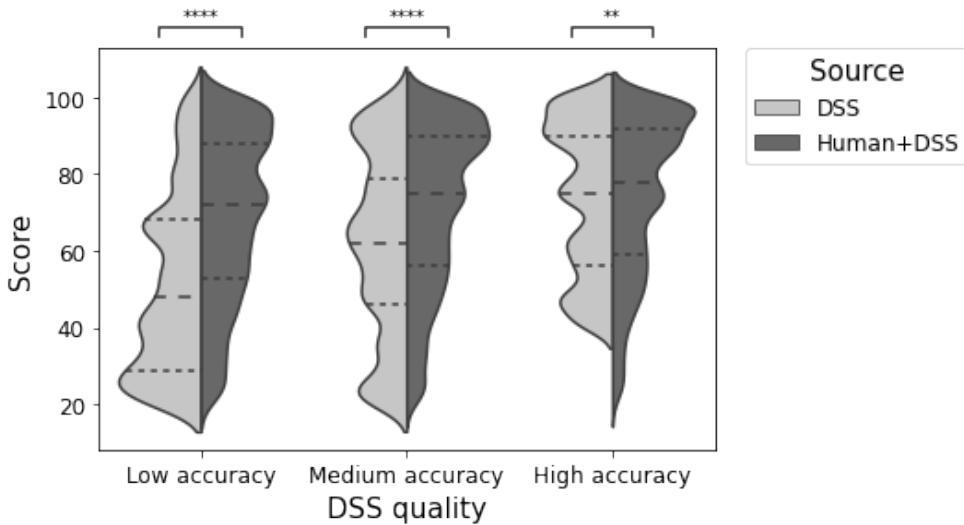


Figure 7.4: Score distribution by DSS quality. We compare the score obtained by human participants with machine assistance and the score that the machine would obtain.

considered two conditions of high terrain cost and low terrain cost (see Table 7.1).

Figure 7.5 shows the score distributions under the four studied conditions. These results suggest there are some variations in score distributions, but there is no consistent increase or decrease in the median score. Under the low-cost condition, the unbiased DSS leads to an average score per play of 75 points, while the biased DSS leads to a score of 74 points. Under the high-cost condition, the unbiased DSS leads to a median score of 74 points, while the biased DSS to 71 points. The difference is statistically significant in the low cost condition ($p < 0.0001$), but not in the high cost condition ($p > 0.05$).

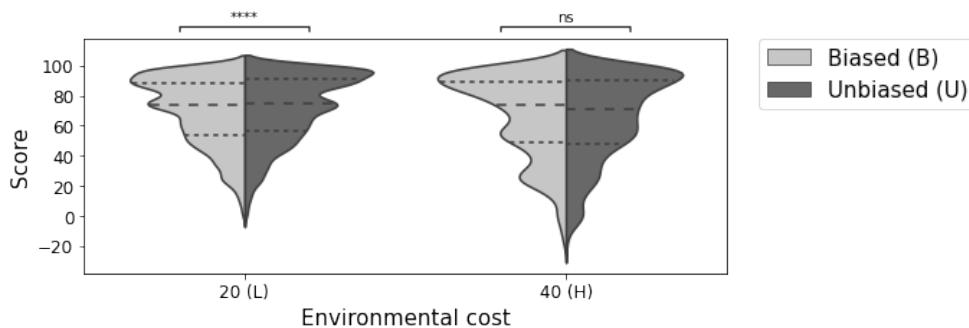


Figure 7.5: Score distribution using a biased DSS or an unbiased one, for the low (LB vs LU) and high cost (HB vs HU) conditions.

In Section 7.6.1 we perform an analysis per map and per cost condition, observing that the unbiased DSS leads to higher scores in almost all cases, with the exception of the medium-difficulty map in the high-condition cost.

Score distribution by bias condition and environment cost

We compare here the biased and unbiased DSS per map and under different cost conditions: high cost (HB vs HC, Figure 7.7) and low cost (LB vs LC, Figure 7.6).

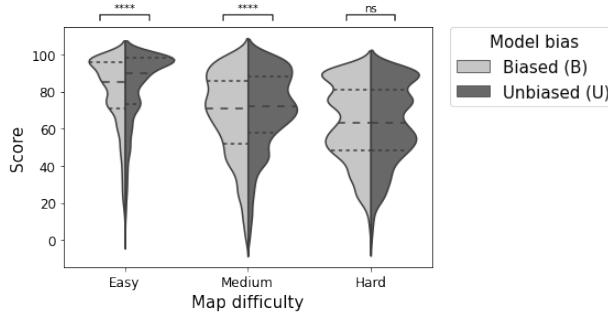


Figure 7.6: Score comparison of biased (LB) and unbiased (LU) DSS under a low cost condition.

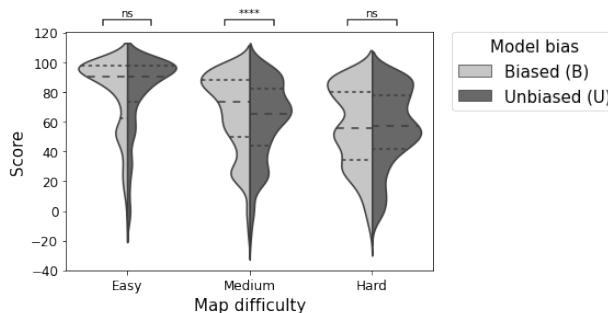


Figure 7.7: Score comparison of biased (HB) and unbiased (HU) DSS under a high cost condition.

Probability of Bad Plays.

We now examine the extent to which the DSS might prevent users from clicking on low-scoring cells, to understand whether the DSS, under various experimental conditions, helps the user avoid these clicks.

The distributions shown in Figure 7.3 suggest that the increase in performance due to the DSS is, at least in part, due to a decrease in the probability of obtaining

a low score. To study this hypothesis, we define a “bad play” as a click on a cell with a score lower than the median in a map. Figure 7.8 shows that the DSS reduces the probability of bad plays, and that the reduction is in general larger when the DSS is more accurate, particularly in the medium and hard maps.

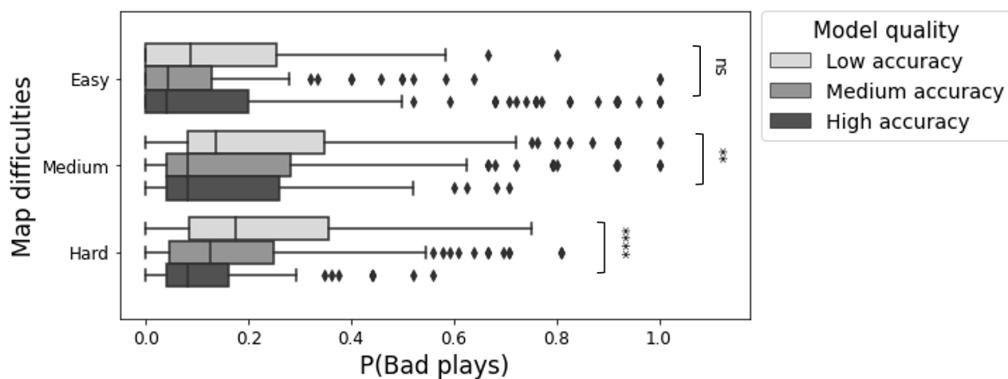


Figure 7.8: Probability of “bad plays” (below median score) under different DSS accuracy.

Figure 7.9 show that unbiased DSS (LU and HU conditions) reduce more the chances of a bad play than the biased DSS (LB and HB).

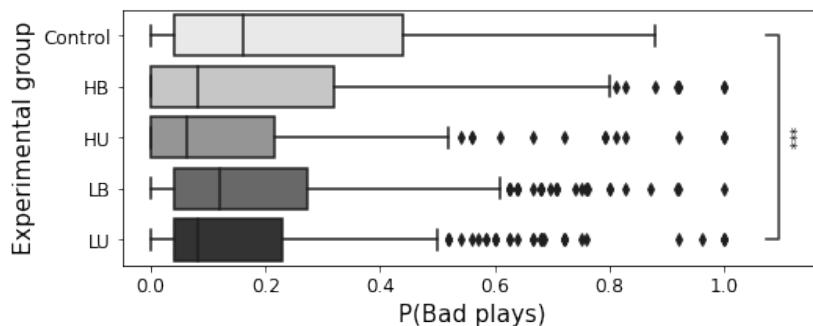


Figure 7.9: Probability of “bad plays” (below median score) under different conditions of DSS bias. HB and HU are biased and unbiased DSS, respectively, in the high cost condition. LB and LU correspond to the low cost condition.

Map Ordering and Learning Effects.

We study how the score per click obtained by participants changes as they play the game, and we observe it increases in general, which we interpret as a learning effect. We do this first across maps/games and then within a map/game.

The experiments are designed so that players play three games in a row, with the three maps shown in random order. Players, in general, improve their performance as they gain experience. In all maps, the scores participants obtain when the map is in the third position are higher than the scores they obtain when the map is in the first position (figure omitted for brevity). On average, in maps played in the first position participants obtain an average median score of 69 points whereas, in maps played in the third position, participants obtain an average median score of 79 points. We use a Kolmogorov-Smirnov (KS) two-sample test to compare them, finding that obtained differences between score distributions are statistically significant ($p < 0.0001$) in each map.

We can also compute a *learning curve* for each user by concatenating the scores they obtain in each play and in each map, i.e., games are characterized as a time series composed of the timestamps and scores obtained for each of the 25 plays from each game. To understand these learning curves, we follow a clustering-based approach that has been shown to be useful for examining learning behavior [152]. This approach uses Dynamic Time Warping (DTW) to compute distances and construct a similarity matrix, where values express the level of similarity between each pair of games. Using a Relaxed Minimum Spanning Tree (RMST), we prune the weakest similarities and then use Markov Stability [58, 123] to obtain clusters of games with similar temporal behaviors. Between the produced multi-scale clustering, we use the one that yields a lower number of clusters.

Figure 7.10 describes the centroids of the obtained clustering, composed of 4 clusters. This figure suggests that all users tend to improve their score as their progress, and that this improvement is to a large extent dependent on their initial scores.

Exploration and exploitation behavior.

Finally, we consider the extent to which exploration/exploitation behavior may be affected by experimental conditions.

This game requires participants to balance *exploration*, i.e., seeking new high-yield areas, and *exploitation*, i.e., reaping the rewards from high-yield areas already found. This behavior is, to some extent, observable. Two consecutive clicks near each other can be interpreted as exploitation, while consecutive clicks far from each other can be interpreted as exploration. We are particularly interested in the extent to which these happen in the presence of a DSS, and on whether they are fruitful in the sense of leading to high-score plays. Figure 7.11 compares the euclidean distance between two consecutive clicks (in the X-axis) and the obtained score (in the Y-axis). For clarity, we group distances and scores. We observe that exploration (third column, “far”) often leads to low scores, while

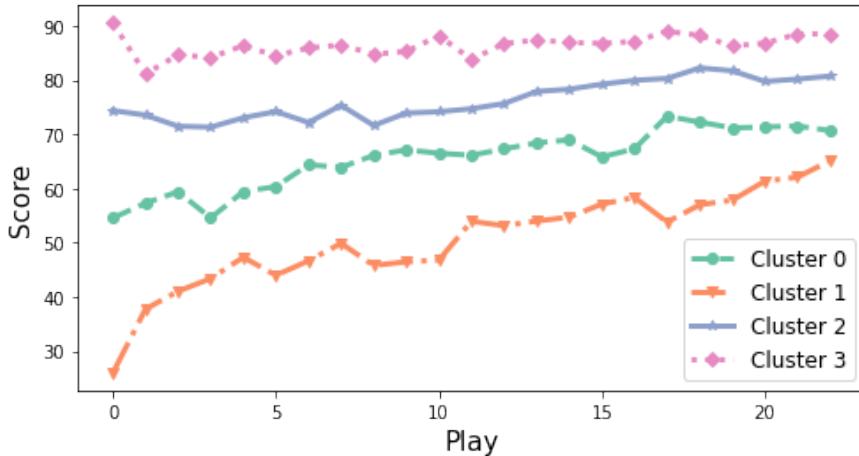


Figure 7.10: Centroids of clusters illustrating how the average score per play increases as players learn during each game.

exploitation (first column, “near”) often leads to high scores.

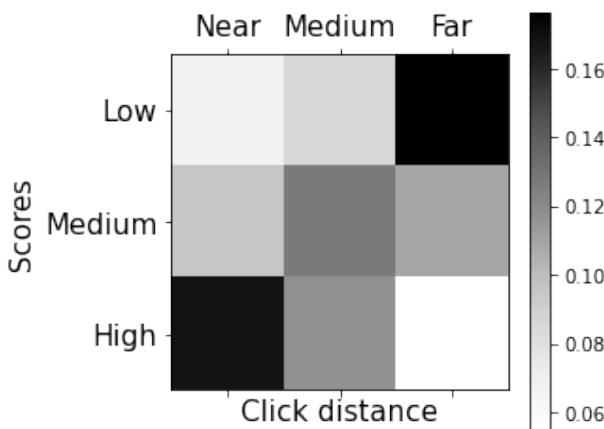


Figure 7.11: Exploration/exploitation behavior and performance. We associate clicks near each other (first column) with exploitation, and clicks far from each other (third column) with exploration. Values indicate percentage of plays.

Examining the score profile of several games, we observe that many participants spend the initial plays locating high-yield areas, and then switch to exploiting those. Comparing the proportion of exploration/exploitation behaviour described in Figure 7.11 with that obtained without the DSS (figure omitted for brevity), we do not find significant differences. Both matrices differ by less than 0.02 in terms of RMSE (Root-Mean-Square Error). Furthermore, we do not observe statistically significant differences in the performance of a given type of

play (near/medium/far click). Binomial tests focused on determining whether the probability of obtaining a high score changes for a given click distance, between the no-DSS and the DSS conditions, yield p-values larger than 0.05. This suggests that the DSS does not lead to a change in strategy by making participants more willing to explore or more willing to exploit, but instead makes both exploration and exploitation more efficient, proportionally.

7.6.2 Time

In this section, we evaluate the completion time across games.

For games played without a DSS, we measured a completion time of 37 ± 18 seconds (average and standard deviation) per map. Given there are 25 rounds, this means users click on a cell roughly once every 1.5 seconds. In games with the DSS, the completion time was 55 ± 40 seconds per map. This corresponds to about one cell clicked every 2.2 seconds. It is clear that the DSS induces a longer completion time, about 50% longer. The distribution of completion times is shown in Figure 7.12, where we also observe that completion time is in general correlated with map difficulty, and harder maps take longer to complete.

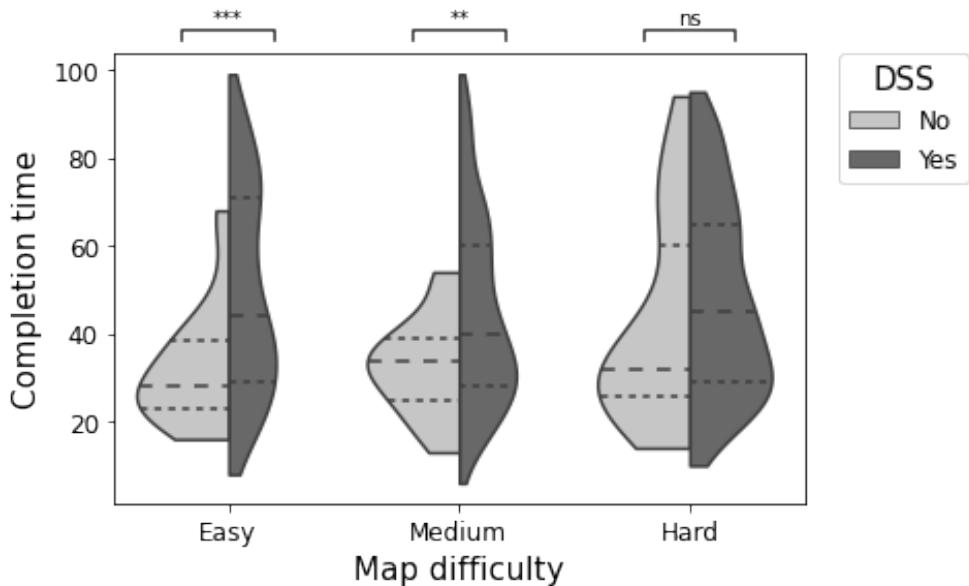


Figure 7.12: Completion time distribution by map. For representation purposes, outlier games taking more than 100 seconds have been removed.

Reliance on Machine Assistance

In this section, we use two approaches to measure the extent to which participants are willing to rely on the DSS.

In the *implicit* approach, we measure the distance between the recommended point and the cell selected by the user. This is depicted in Figure 7.13. Users seem to correctly account for the accuracy of the system and rely less on the low accuracy DSS than on the medium or high accuracy DSS. Indeed, in the low accuracy condition, after the initial play users basically ignore the DSS – note that the expected distance between two randomly-chosen points in a 32×32 square⁵ is approximately $0.52 \cdot 32 = 16.6$, which is close to what we observe in this condition. In contrast, in the medium accuracy condition and particularly in the high accuracy condition, they tend to click closer to the recommended point.

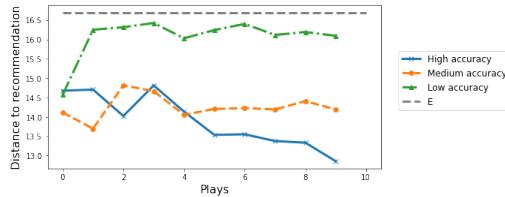


Figure 7.13: Average distance between recommendation and selected cell by model quality. The horizontal line (E) is the expected distance between two random points in a 32×32 grid

We also ask users to tell us *explicitly* their acceptance of the DSS by using the Technology Acceptance Survey described in Section 7.4.2, which yields a score between 0 (complete rejection) and 40 (complete acceptance). Results, shown in

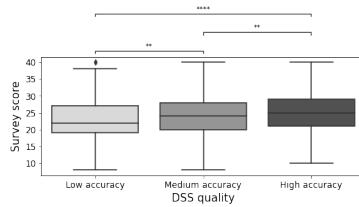


Figure 7.14: Results of the exit survey on technology acceptance, by DSS accuracy.

Figure 7.14, indicate that most users have an intermediate level of acceptance (20–30 points out of 40) and median acceptance differs by less than 5 points across accuracy conditions. However, high acceptance is more likely in the high and

⁵Analytically, in a unit square this is $(2 + \sqrt{2} + 5 \ln(\sqrt{2} + 1)) / 15 \approx 0.52140 \dots$

medium accuracy conditions than in the low accuracy condition, and differences in the distribution of acceptance are statistically significant at $p < 0.01$.

7.7 Discussion

Our research question was “*How do the characteristics of a decision support system impact human performance, time to completion, and reliance?*”

First, we observe that the platform and the experiments we designed allow experimentation on the stated variables, as participants respond in an observable manner to aspects such as the accuracy and bias of the received support. We also observe that completing a game does not take much time, and that all participants seem to learn how to increase their score after a few interactions. The decision support we provide, when operating on its own, has lower performance than an average human participant; however, the combination of human and machine intelligence in this game can outperform both the machine and the human acting on their own.

Second, our experimental results suggest that, in this context, to a large extent, participants behave rationally with respect to the accuracy of the DSS. Despite us intentionally adding noise to the DSS recommendations, participants respond to the average accuracy of the DSS, following more closely the high accuracy DSS than the medium accuracy DSS, and to a large extent ignoring the low accuracy DSS. This result is aligned with previous work in which participants correctly calibrated their reliance according to DSS accuracy (e.g., [213, 214]).

Third, participants’ response to the presence of bias in this DSS is found to be small and inconsistent, as they obtain slightly higher scores with the unbiased DSS in the presence of low cost but slightly lower scores with the unbiased DSS when the cost is higher. It might be possible that a larger bias might lead to a more directly observable effect, but we remark that the high cost, biased condition introduces a bias of 40 points per play (out of 100), which is fairly substantial.

Fourth, we observe that the presence of a DSS leads users to take about 50% longer to complete each task. Previous work in decision-making for other tasks also observed an increased completion time (e.g., [157, 188, 124]).

We observed no statistically significant differences among participants of different sociodemographics (gender, age, country, education, professional background) in terms of any of the dependent variables studied (score, time to completion, or reliance).

A result that is somewhat unexpected is that participants, despite apparently ignoring the low accuracy DSS, in many cases express a moderate acceptance of it in the exit survey. Previous work has found that a wrong recommendation is less penalized if the final task performance is not harmed [212], but in our case

participants indeed obtain a lower score with the low-accuracy DSS. This suggests that, in practice, a malfunctioning DSS might not be detectable by users.⁶ This can have severe consequences when deploying a DSS without taking into account that its users might not be able to evaluate correctly the quality of the recommendations or predictions.

7.7.1 Limitations

Decision-making processes in professional usage and/or in high-stakes scenarios are different from those of inconsequential decisions, such as the game we have designed. First, feedback is rarely available immediately, and indeed the process of acquiring expertise in some domains, such as criminal justice, involves to some extent observing the consequences of decisions made years ago [100]. Second, while we compensate economically participants for executing their task, this is to encourage attentiveness, and not to simulate a high-stakes situation. Third, different professional contexts in which DSS are used (such as healthcare, human resources, criminal justice) may encourage different practices with respect to the DSS; they may also involve people with different backgrounds, including varying degrees of numeracy and different levels of previous experience with algorithmic support. Considering this, we believe DSS studies leading to domain-specific designs require domain-specific experimentation. What we provide, in contrast, is a platform to explore quantitatively user response to key aspects of a DSS at a scale.

A further limitation is that while we release a platform in which parameters can be varied by researchers, we do not provide mechanisms to, for instance, predict the difficulty of a game. We provide three test maps, and notice that varying some parameters of the map generator and of the DSS probably requires experimentally fine-tuning other parameters. We recommend doing this empirically – as we have done in this paper – with a group of participants.

Ethics and data protection.

Our research was approved by our university’s Ethics Review Board, including a data protection assessment.

Reproducibility.

Our platform will be available as free software with the camera-ready version of this paper.

⁶This was humorously dubbed the “Functional Indeterminacy Theorem” by John Gall [79].

Chapter 8

CONCLUSIONS

8.1 Main contributions

This thesis addresses the problem of Algorithmic Fairness at different stages of the ML life-cycle.

As a starting point, we showcased the feasibility of Poisoning Attacks to induce Algorithmic Unfairness which corresponds to the first investigation done in the topic. Our results prove that state-of-the-art attacks can be adapted to increase existing biases in input data, leading to unfair outcomes on the predictive model with a very limited amount of carefully crafted samples. This novel type of attacks can be used to both introduce algorithmic unfairness, as well as for increasing it where it already exists. This can be done even without access to the specific model being used, as a surrogate model can be used to mount a black-box transfer attack.

Furthermore, we evaluated the professional biases across AI stakeholders by using the theory of Economics of Conventions. Our results suggest important differences in motifs and moral orders across developers, researchers and general-public involved in the field of AI.

After, we evaluated model biases in two settings: first, a multi-sided market scenario, assessing how disparities in predictive performance affect both sides of the market, leading to disparities in user satisfaction in both sides of the market. Second, we detected predictive biases when building a predictive model for detecting Anorexia Nervosa using online posts. In this second setting, we assessed state-of-the-art techniques to reduce predictive biases, discussing the trade-offs between the distinct options.

Finally, we developed an experimental platform based on an online game to study how user behavior varies when making use of a DSS, depending on its characteristics. The results show how users in general benefit from the presence of automatic help, and are able to effectively calibrate their reliance according

to the capacities of the system. Nevertheless, we found many cases where users expressed a moderate acceptance of low accuracy DSS while answering the opinion survey. This result is interesting given the fact that users look to ignore the suggestions of the low-accuracy model in practice.

Discussing more general findings, we acknowledge that existing biases can be often mitigated, but there is a limit to this mitigation if the source of bias is not known. Biases can be created in an unintended manner as seen in Chapter 6, but the results of Chapter 3 prove that biases can also be created intentionally, what would require recurrent monitoring in cases of continuous learning, where the ML keeps learning while deployed.

The results on the analysis of professional bias (Chapter 4) and user-machine interaction (Chapter 7) suggest that the subjectivity and preconceptions the user might have before interacting with the system could play a role in the manner she evaluates its the quality and capacities. As an example, a user guided by the *Industrial* convention might accept any system that yields high-enough accuracy, while another user that is more conditioned by the *Civic* convention would require the system to be fair before accepting it. In this example, the first could easily incur in overreliance in the machine capacities [87] while the latter could develop Algorithmic Aversion [65]

8.2 Recommendations for practitioners

Bias in the ML life-cycle occurs at different stages. Removing all possible sources of inequalities becomes a nearly impossible task for practitioners, hence, mitigating all possible bias must be an objective of any data-oriented project. The results of this thesis and the background and related work highlight the necessity of not only accounting for predictive performance overall, but comparing it across demographic groups to detect potential discrimination before deploying ML systems in production.

8.2.1 Poisoning attacks

The shown feasibility of poisoning attacks on Algorithmic Fairness, there is a need for accounting for such cases in at least two possible scenarios. In one hand, there is a need for approaches that are robust against such a type of attacks. Although the situations where an attacker can have enough knowledge about the model and access to the training data at the same time are not so common, the strong effect that can be caused by adding a small percentage of perturbed samples requires special attention. In the other hand, our results show how the worst-case scenario

in which errors committed during the data labeling phase might lead. In this case, intention to harm is not required, but still might lead to strongly biased predictions.

8.2.2 Professional biases

Although our work on professional biases is just an initial effort on the topic, it clearly highlights that different stakeholders give value to AI in very different forms. An initial step towards addressing this issue requires first to acknowledge it. Fostering the inclusion and diversity of the teams involved in the research and development of AI systems would help into accounting for more diverse points of view. Involving all types of stakeholders during the creating of AI systems can help into accounting for all values, moral principles and orders of worth of the people affected by them.

8.2.3 Algorithmic assessments

The results obtained during the Algorithmic Assessment detailed in Chapter 5 show how the incorporation of ML techniques might not benefit equally all demographic groups. There is a need for accounting for this, while at the same time considering if the benefits of including the system exceed the harms and inequalities it might cause. Additionally, our results show how the presence of a trade-off between accuracy and fairness is very common. Whereas maximizing the first has been the main focus of most of the advancements in the ML field, mitigating unfairness typically requires lowering the overall performance of the system. Although it might look undesirable, there are several use cases where the risk of using biased ML models exceeds the benefits of slightly higher accurate ones.

8.2.4 Algorithmic unfairness mitigation

The work done with the Anorexia Nervosa dataset shows the limitations of state-of-the-art techniques for bias mitigation. As shown in our analysis, the preference for one solution over the rest might depend on the context of the application, as little improvements in one fairness criteria often lead to worsen results of others.

Also, the analyzed case corresponds to a particular example, where the demographic group with higher prevalence in the dataset (females) was obtaining lower predictive performance. Nevertheless, previous work refers to Sample Size Bias [40] as the main root-cause of the subsequent performance disparities yielded by ML systems. Our observations suggest that accounting for a number of samples might not always be enough to equalize predictive performance. With the cause of this effect being that the selected features were not equally predictive across genders, it suggests that considering data separability might be a need.

8.2.5 Interaction bias

Our results suggest that in the low-stake scenario where the experimentation is conducted, user's performance benefits from the presence of a DSS system, even in the case that its predictive performance is low. The low-complexity nature of the analyzed scenario facilitates that users adapt their reliance to the quality of the received predictions, what needs to be accounted to avoid Algorithmic aversion [65] and Overreliance [87] in further versions of the same system. Additionally, we also observed that users require more time to take the decision with the presence of the DSS, what might be relevant in scenarios where the time to take the decision is limited as the cases described by Cummings et al. [54].

8.3 Future work

There are several future directions that could potentially be motivated by this thesis. Below, we list several research opportunities and briefly discuss their applicability.

8.3.1 Poisoning attacks

Studying adversarial attacks on algorithmic fairness can help to make machine learning systems more robust. Additional type of models such neural nets and/or other data sets can be considered in the future to extend the work proposed here. Although experiments in this paper are done using a specific technique based on a poisoning attack, other techniques can be certainly considered. Other approaches such as causality-based techniques could be explored as future work.

8.3.2 Professional biases

The approach presented in the corresponding chapter is the first contribution towards building an automatic text classifier of EC. The use of automatic models to perform the analysis enables the possibility of considering large amounts of information when accounting the conventions in a given dataset. This approach could be used in future analysis to extract conclusions in a variety of domains where prevalence of the EC needs to be studied. To facilitate the re-usage of this work, a repository containing the implemented code and the collected data has been published.

This work focused on three data sources which we considered relevant to reflect different perceptions about AI, i.e. the perspective of researchers, developers

and the general public. In the future, it would be interesting to study other types of interactions in data sources such as newspapers, online videos and chats.

In further steps, one focus will aim to detect and analyze common conflicts in software development and their underlying (assumable conflicting) conventions, beyond the already obvious problems of coordination between open source- and profit oriented AI development. With this, we hope to contribute to a more plural understanding of AI research and development, considering underlying moral registers which influence the motivations, objectives, processes and values of these projects.

8.3.3 Algorithmic assessments and bias mitigation

The combination of knowledge extracted from the use cases described in this document and the advancements in the formalization and standardization of algorithmic assessments and audits enable to facilitate the analysis of future use cases in novel domains. Additionally, in case new instances of algorithmic bias or unfair performance disparities were found in new cases, existing or novel bias mitigation procedures could be applied to these use cases.

8.3.4 Interaction bias

We offer an expressive platform that is useful for various types of research; indeed, we encourage researchers to use this platform, and we make freely available its source code. As future work, we would like to consider situations that induce over-reliance or under-reliance in the DSS. We would also like to study whether communicating the accuracy and confidence of the DSS, or using other mechanisms for transparency or explainability can prevent these situations, or lead to increased user performance. Another line of research we would like to explore is the response of participants to DSS failures, such as a sudden drop in accuracy, both in terms of how they perform and how they perceive the system.

8.4 Reproducibility

The code developed within this thesis uses the following Python libraries:

1. *Pandas, NumPy and Spark* for data management
2. *Scikit-learn and PyTorch* for training machine learning models
3. *IBM AIF360* for quantifying and mitigating unfairness in input data and models

Open-source code, data and models produced during the development of this thesis are available at: <http://www.github.com/dsolanno/phdthesis2022>

Bibliography

- [1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and et al. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI’16, page 265–283, USA, 2016. USENIX Association.
- [2] H. Abdollahpouri and R. Burke. Multi-stakeholder recommendation and its connection to multi-sided fairness, 2019.
- [3] A. Acharya, A. Howes, C. Baber, and T. Marshall. Automation reliability and decision strategy: A sequential decision making model for automation interaction. *Proceedings of the Human Factors and Ergonomics Society*, 1:144–148, 2018.
- [4] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification, 2018.
- [5] A. Agarwal, M. Dudík, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms, 2019.
- [6] O. Alvarado and A. Waern. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI ’18, New York, NY, USA, 2018. Association for Computing Machinery.
- [7] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR)*. American Psychiatric Association, 2013.
- [8] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. Peters, J. Power, S. Skjonsberg, L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, and O. Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pages 84–91, New Orleans - Louisiana, June 2018. Association for Computational Linguistics.

- [9] M. M. Archibald. Investigator triangulation: A collaborative strategy with potential for mixed methods research. *Journal of Mixed Methods Research*, 10(3):228–250, 2016.
- [10] C. Aschwanden. Science isn’t broken. it’s just a hell of a lot harder than we give it credit for., 08 2015.
- [11] J. Asplund, M. Eslami, H. Sundaram, C. Sandvig, and K. Karahalios. Auditing race and gender discrimination in online housing markets. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):24–35, May 2020.
- [12] I. Augenstein, M. Das, S. Riedel, L. Vikraman, and A. McCallum. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada, Aug. 2017. Association for Computational Linguistics.
- [13] J. Ayers, T. Caputi, C. Nebeker, and M. Dredze. Don’t quote me: reverse identification of research participants in social media studies. *npj Digital Medicine*, 1, 12 2018.
- [14] R. Baeza-Yates. Bias on the web. *Commun. ACM*, 61(6):54–61, may 2018.
- [15] N. Balfe, S. Sharples, and J. R. Wilson. Understanding is key: An analysis of factors pertaining to trust in a real-world automation system. *Human Factors*, 60:477 – 495, 2018.
- [16] K. Bansak. Can nonexperts really emulate statistical learning methods? a comment on the accuracy, fairness, and limits of predicting recidivism. *Political Analysis*, pages 370–380, 2019.
- [17] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):19, 2019.
- [18] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, D. Wadsworth, and H. Wallach. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs, 2021.

- [19] S. Barocas and M. Hardt. Fairness in machine learning nips 2017 tutorial, 2017.
- [20] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [21] P. Batifoulier, N. Da Silva, and J.-P. Domin. *Economie de la santé*. Armand Colin, 2018.
- [22] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676, 2019.
- [23] A. Benton, G. Coppersmith, and M. Dredze. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain, Apr. 2017. Association for Computational Linguistics.
- [24] K. Bergstrom. "don't feed the troll": Shutting down debate about community expectations on reddit.com. *First Monday*, 16, 07 2011.
- [25] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression, 2017.
- [26] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons, 2019.
- [27] A. J. Biega, K. P. Gummadi, and G. Weikum. Equity of attention. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR '18*, 2018.
- [28] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In J. Langford and J. Pineau, editors, *29th Int'l Conf. on Machine Learning*, pages 1807–1814. Omnipress, 2012.
- [29] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.
- [30] A. Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, feb 2021.
- [31] M. Bogen and A. Rieke. Help wanted: an examination of hiring algorithms, equity, and bias, 2018.

- [32] P. Boisard. *Camembert: A national myth*, volume 4. University of California Press, 2003.
- [33] L. Boltanski and E. Chiapello. The new spirit of capitalism. *International journal of politics, culture, and society*, 18(3-4):161–188, 2005.
- [34] L. Boltanski and L. Thévenot. *On justification: Economies of worth*, volume 27. Princeton University Press, 2006.
- [35] N. Bosch, S. K. D’Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao. Detecting student emotions in computer-enabled classrooms. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI’16, page 4125–4129. AAAI Press, 2016.
- [36] A. Bosu, A. Iqbal, R. Shahriyar, and P. Chakraborty. Understanding the motivations, challenges and needs of blockchain software developers: A survey. *Empirical Software Engineering*, 24(4):2636–2673, 2019.
- [37] T. Bucher. The algorithmic imaginary: exploring the ordinary affects of facebook algorithms. *Information, Communication & Society*, 20:1–15, 02 2016.
- [38] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, volume 20, pages 454–464, 2020.
- [39] C. Buntain and J. Golbeck. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW ’14 Companion, page 615–620, New York, NY, USA, 2014. Association for Computing Machinery.
- [40] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.
- [41] R. Burke. Multisided fairness for recommendation, 2017.
- [42] R. Burke. Multisided fairness for recommendation, 2017.
- [43] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21:277–292, 09 2010.

- [44] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, apr 2017.
- [45] F. P. Calmon, D. Wei, K. N. Ramamurthy, and K. R. Varshney. Optimized data pre-processing for discrimination prevention, 2017.
- [46] M. Castells. *The Internet Galaxy: Reflections on the Internet, Business, and Society*. Oxford University Press, Inc., USA, 2001.
- [47] C. Castillo. Fairness and transparency in ranking. *SIGIR Forum*, 52(2):64–71, Jan. 2019.
- [48] L. E. Celis, A. Mehrotra, and N. K. Vishnoi. Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 369–380, New York, NY, USA, 2020. Association for Computing Machinery.
- [49] R. Chambers and G. Conway. Sustainable rural livelihoods: practical concepts for the 21st century. *IDS Discussion Paper*, 296, 01 1992.
- [50] S. Chancellor, M. L. Birnbaum, E. D. Caine, V. M. B. Silenzio, and M. De Choudhury. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, page 79–88, New York, NY, USA, 2019. Association for Computing Machinery.
- [51] D. Chen, K. T. Stolee, and T. Menzies. Replicating and scaling up qualitative analysis using crowdsourcing: A github-based case study. arXiv preprint arXiv:1702.08571, 2017.
- [52] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.
- [53] H. Cramer, K. Holsteina, J. W. Vaughan, H. D. III, M. Dudík, H. Wallach, S. Reddy, and J. Garcia-Gathright. Tutorial: Challenges of incorporating algorithmic fairness into industry practice, 2019.
- [54] M. L. Cummings. Automation bias in intelligent time critical decision support systems. In *AIAA 3rd Intelligent Systems Conference*, pages 2004–6313. AIAA, 2004.

- [55] T. O. Cunha, I. Weber, H. Haddadi, and G. L. Pappa. The effect of social feedback in a reddit weight loss community. In *Proceedings of the 6th International Conference on Digital Health Conference*, DH '16, page 99–103, New York, NY, USA, 2016. Association for Computing Machinery.
- [56] N. Da Silva. L'industrialisation de la médecine libérale: une approche par l'économie des conventions. *Management Avenir Sante*, 1(1):13–30, 2018.
- [57] S. Datta and E. Adar. Extracting inter-community conflicts in reddit. *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):146–157, Jul. 2019.
- [58] J. C. Delvenne, S. N. Yaliraki, and M. Barahona. Stability of graph communities across time scales, 2009.
- [59] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, 2019.
- [60] J.-L. Denis, A. Langley, and L. Rouleau. Strategizing in pluralistic contexts: Rethinking theoretical frames. *Human Relations*, 60(1):179–215, 2007.
- [61] R. Diaz-Bone. *Die “Economie des conventions”*. *Grundlagen und Entwicklungen der neuen französischen Wirtschaftssoziologie*. Springer, 2018.
- [62] R. Diaz-Bone. *Valuation an den Grenzen von Datenwelten*, pages 71–95. Springer Fachmedien Wiesbaden, Wiesbaden, 2019.
- [63] R. S. Diaz-Bone. Economics of convention and the history of economies: towards a transdisciplinary approach in economic history. *Historical Social Research*, 36:7–39, 2011.
- [64] B. Dietvorst, J. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General*, 144, 11 2014.
- [65] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
- [66] J. Dressel and H. Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):1–6, 2018.

- [67] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.
- [68] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), Apr 2004.
- [69] M. Eslami, K. Vaccaro, K. Karahalios, and K. Hamilton. "be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *ICWSM*, 2017.
- [70] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 01 2017.
- [71] E. Fast and E. Horvitz. Long-term trends in the public perception of artificial intelligence. arXiv preprint arXiv:1609.04904, 2016.
- [72] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact, 2015.
- [73] S. Ferryman and M. Pitcan. Fairness in precision medicine, 2018.
- [74] A. Fitó, X. Espinach, R. Gras, and J. Ramos. La clau pot ser un nom: Detecció d'evidències de discriminació en l'accés al mercat de lloguer d'habitatge a Barcelona., 2020.
- [75] A. C. Freeman. Eating disorders in males : a review. *African Journal of Psychiatry*, 8, 2005.
- [76] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning, 2018.
- [77] M. G. Dealing with fairness in public policy analysis: A methodological framework. Scientific analysis or review KJ-NA-28751-EN-N, Publications Office of the European Union, Luxembourg (Luxembourg), 2017.
- [78] G. Galdon Clavell, M. Martín Zamorano, C. Castillo, O. Smith, and A. Matic. Auditing algorithms: On lessons learned and the risks of data minimization, 2020.
- [79] J. Gall. *Systemantics: the underground text of systems lore: how systems really work and especially how they fail*. General Systemantics Press, 1986.

- [80] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, and K. Crawford. Datasheets for datasets, 2018.
- [81] D. Gershgorn. If ai is going to be the world’s doctor, it needs better textbooks, 09 2018.
- [82] E. Gkerekakis. The constitutive role of conventions in accomplishing coordination: Insights from a complex contract award project. *Organization Studies*, 35(10):1473–1505, 2014.
- [83] M. Glenski and T. Weninger. Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ASONAM ’17, page 609–612, New York, NY, USA, 2017. Association for Computing Machinery.
- [84] N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria, 12 2018.
- [85] B. Green. The false promise of risk assessments: Epistemic reform and the limits of fairness, 2020.
- [86] B. Green and Y. Chen. Disparate Interactions. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99, New York, NY, USA, jan 2019. ACM.
- [87] B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019.
- [88] B. Green and Y. Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts, 2020.
- [89] N. Grgic-Hlaca, C. Engel, and K. Gummadi. Human decision making with machine assistance: An experiment on bailing and jailing, 10 2019.
- [90] S. Gupta and C. Manning. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1–9, Chiang Mai, Thailand, Nov. 2011. Asian Federation of Natural Language Processing.
- [91] S. Hajian, F. Bonchi, and C. Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126, 2016.

- [92] G. Haralabopoulos and I. Anagnostopoulos. Lifespan and propagation of information in on-line social networks: A case study based on reddit. *Journal of Network and Computer Applications*, 56, 03 2014.
- [93] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning, 2016.
- [94] F. Hassan and X. Wang. Mining readme files to support automatic building of java projects in software repositories: Poster. In *Proceedings of the 39th International Conference on Software Engineering Companion*, ICSE-C '17, pages 277–279, Piscataway, NJ, USA, 2017. IEEE Press.
- [95] L. He, K. Lee, O. Levy, and L. Zettlemoyer. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [96] G. Hertel, S. Niedner, and S. Herrmann. Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, 32(7):1159–1177, 2003.
- [97] Q. H.L., H. A.S., D. R., B. U., and V. S. Body dissatisfaction, importance of appearance, and body appreciation in men and women over the lifespan. *Front Psychiatry*, 10, 2019.
- [98] K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.
- [99] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects, 2019.
- [100] A. M. Holsinger, C. T. Lowenkamp, E. Latessa, R. Serin, T. H. Cohen, C. R. Robinson, A. W. Flores, and S. W. VanBenschoten. A rejoinder to dressel and farid: New study finds computer algorithm is more accurate than humans at predicting arrest and as good as a group of 20 lay experts. *Fed. Probation*, 82:50, 2018.
- [101] K. Holstein, B. McLaren, and V. Aleven. Student learning benefits of a mixed-reality teacher awareness tool in ai-enhanced classrooms, 06 2018.
- [102] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach. Improving fairness in machine learning systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 2019.

- [103] T. Hurni, T. Huber, and J. Dibbern. Coordinating platform-based multi-sourcing: introducing the theory of conventions. In *36th International Conference on Information Systems*, 2015.
- [104] J. A. Hutson, J. G. Taft, S. Barocas, and K. Levy. Debiasing desire. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, Nov 2018.
- [105] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning, 2018.
- [106] F. Jahanbakhsh, J. Cranshaw, S. Counts, W. S. Lasecki, and K. Inkpen. An experimental study of bias in platform worker ratings: The role of performance quality and gender. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.
- [107] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446, 2002.
- [108] A. Javaheri, N. Moghadamnejad, H. Keshavarz, E. Javaheri, C. Dobbins, E. Momeni, and R. Rawassizadeh. Public vs media opinion on robots. *ArXiv*, abs/1905.01615, 2019.
- [109] S. M. Julia Angwin, Jeff Larson and P. Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks., 05 2016.
- [110] F. Kamiran and T. Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011.
- [111] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer, 09 2012.
- [112] K. Kersting, J. Peters, and C. Rothkopf. Was ist eine professur für künstliche intelligenz? arXiv preprint arXiv:1903.09516, 2019.
- [113] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.
- [114] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016.

- [115] R. Kohavi and R. Longbotham. Online controlled experiments and a/b testing, 01 2017.
- [116] A. Kozica, S. Kaiser, and M. Friesl. Organizational routines: Conventions as a source of change and stability. *Schmalenbach Business Review*, 66(3):334–356, 2014.
- [117] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompat-siaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*, WWW ’18, page 853–862, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
- [118] B. Kulynych, R. Overdorf, C. Troncoso, and S. Gürses. Pots. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan 2020.
- [119] C. Lafaye and L. Thévenot. Une justification écologique?: Conflits dans l’aménagement de la nature. *Revue française de sociologie*, pages 495–524, 1993.
- [120] P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making, 2018.
- [121] V. Lai and C. Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.
- [122] K. R. Lakhani and R. G. Wolf. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *MIT Sloan Working Paper No. 4425-03*, 2003.
- [123] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Random walks, markov processes and the multiscale modular organization of complex networks. *IEEE Transactions on Network Science and Engineering*, 1(2):76–90, 2014.
- [124] M. Langer, C. J. König, and V. Busch. Changing the means of managerial work: effects of automated decision support systems on personnel selection tasks. *Journal of business and psychology*, 36(5):751–769, 2021.
- [125] J. Larus, C. Hankin, S. G. Carson, M. Christen, S. Crafa, O. Grau, C. Kirchner, B. Knowles, A. McGettrick, D. A. Tamburri, et al. When computers

decide: European recommendations on machine-learned automated decision making, 2018.

- [126] J. D. Lee and K. A. See. Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, jan 2004.
- [127] K. Lehavot, J. Katon, J. Chen, J. Fortney, and T. Simpson. Post-traumatic stress disorder by gender and veteran status. *American journal of preventive medicine*, 54:e1–e9, 01 2018.
- [128] Z. J. Lin, J. Jung, S. Goel, and J. Skeem. The limits of human predictions of recidivism. *Science Advances*, 6(7):1–8, feb 2020.
- [129] D. E. Losada, F. Crestani, and J. Parapar. Overview of erisk 2019 early risk prediction on the internet. In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 340–357, Cham, 2019. Springer International Publishing.
- [130] S. N. . G. A. Lotlikar V. S. Brain tumor detection using machine learning and deep learning: A review. *Current medical imaging*, 2021.
- [131] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, Oct. 2018. Association for Computational Linguistics.
- [132] K. Lum and W. Isaac. To predict and serve? *Significance*, 13:14–19, 10 2016.
- [133] K. Mallari, K. Inkpen, P. Johns, S. Tan, D. Ramesh, and E. Kamar. Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, New York, NY, USA, apr 2020. ACM.
- [134] L. Manikonda, A. Deotale, and S. Kambhampati. What’s up with privacy?: User preferences and privacy concerns in intelligent personal assistants. *ArXiv*, abs/1711.07543, 2017.
- [135] L. Manikonda, C. Dudley, and S. Kambhampati. Tweeting ai: Perceptions of ai-tweeters (ait) vs expert ai-tweeters (eait). *arXiv preprint arXiv:1704.08389*, 2017.

- [136] W. Mason and D. J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, 2012.
- [137] P. Mayring. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Social Science Open Access Repository, Klagenfurt, Germany, 2014.
- [138] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning, 2019.
- [139] R. Mehrotra, A. Anderson, F. Diaz, A. Sharma, H. Wallach, and E. Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW ’17 Companion, page 626–633, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.
- [140] S. Mei and X. Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *29th AAAI Conf. Artificial Intelligence (AAAI ’15)*, 2015.
- [141] A. K. Menon and R. C. Williamson. The cost of fairness in classification, 2017.
- [142] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrasamee, E. C. Lupu, and F. Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. In B. M. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, and A. Sinha, editors, *10th ACM Workshop on Artificial Intelligence and Security*, AISec ’17, pages 27–38, New York, NY, USA, 2017. ACM.
- [143] A. Mukerjee, R. Biswas, Y. Kalyanmoy, D. Amrit, and P. Mathur. Multi-objective evolutionary algorithms for the risk-return trade-off in bank loan management. *International Transactions in Operational Research*, 9, 03 2002.
- [144] A. Narayanan. Tutorial: 21 fairness definitions and their politics, 2018.
- [145] C. Okoli and W. Oh. Investigating recognition-based performance in an open content community: A social capital perspective. *Information & Management*, 44(3):240–252, 2007.
- [146] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

- [147] H.-L. E. G. on Artificial Intelligence. Ethics guidelines for trustworthy ai, 09 2019.
- [148] O. Papakyriakopoulos, S. Hegelich, J. C. M. Serrano, and F. Marco. Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA, 2020. Association for Computing Machinery.
- [149] N. Papernot, P. D. McDaniel, and I. J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *ArXiv e-prints*, abs/1605.07277, 2016.
- [150] F. Pasquale. *The Black Box society*. Cambridge: Harvard University Press, 2015.
- [151] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, and A. Chakraborty. FairRec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of The Web Conference 2020*. ACM, apr 2020.
- [152] R. L. Peach, S. N. Yaliraki, D. Lefevre, and M. Barahona. Data-driven unsupervised clustering of online learner behaviour, 2019.
- [153] D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 126–131, New York, NY, USA, 2012. ACM.
- [154] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.
- [155] K. Perlin. Improving noise. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, page 681–682, New York, NY, USA, 2002. Association for Computing Machinery.
- [156] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration, 2017.
- [157] M. Portela, C. Castillo, S. Tolan, M. Karimi-Haghghi, and A. A. Pueyo. A comparative user study of human predictions in algorithm-supported recidivism risk assessment, 2022.

- [158] G. A. A. Prana, C. Treude, F. Thung, T. Atapattu, and D. Lo. Categorizing the content of github readme files. *Empirical Software Engineering*, 24(3):1296–1327, Oct 2018.
- [159] J. O. Prochaska and W. F. Velicer. The transtheoretical model of health behavior change. *American Journal of Health Promotion*, 12(1):38–48, 1997. PMID: 10170434.
- [160] G. Quattrone, D. Proserpio, D. Quercia, L. Capra, and M. Musolesi. Who benefits from the "sharing" economy of airbnb? In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 1385–1394, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.
- [161] E. Rader and R. Gray. Understanding user beliefs about algorithmic curation in the facebook news feed, 04 2015.
- [162] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 469–481, New York, NY, USA, 2020. Association for Computing Machinery.
- [163] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus, and J. González. Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis. *J Med Internet Res*, 22(7):e17758, Jul 2020.
- [164] D. Ramírez-Cifuentes. *A Data Driven Framework for Mental Health States Assessment in Social Platforms*. PhD thesis, Universitat Pompeu Fabra, 2022.
- [165] B. Rastegarpanah, K. P. Gummadi, and M. Crovella. Fighting fire with fire. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19*, 2019.
- [166] S. Rendle. Factorization machines. In *2010 IEEE International Conference on Data Mining*, pages 995–1000, Dec 2010.
- [167] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 521–530, New York, NY, USA, 2007. Association for Computing Machinery.

- [168] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, Dec. 2004.
- [169] E. A. Ríssola, D. E. Losada, and F. Crestani. A survey of computational methods for online mental state assessment on social media. *ACM Trans. Comput. Healthcare*, 2(2), mar 2021.
- [170] J. A. Roberts, I.-H. Hann, and S. A. Slaughter. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Management Science*, 52(7):984–999, 2006.
- [171] J.-C. Rochet and J. Tirole. Two-sided markets: An overview. *Institut d'Economie Industrielle working paper*, 2004.
- [172] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2):1–40, 2010.
- [173] R. M. Ryan and E. L. Deci. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, 25(1):54–67, 2000.
- [174] C. F. Ríssola EA, Aliannejadi M. Beyond modelling: Understanding mental disorders in online social media. *Advances in Information Retrieval*, 12035:296–310, 2020.
- [175] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.
- [176] N. J. Salkind. *Between-subjects design*. In *Encyclopedia of research design (Vol. 1)*, pages 82–84. Thousand Oaks, CA: SAGE Publications, Inc., 2017.
- [177] C. Sammut and G. I. Webb, editors. *TF-IDF*, pages 986–987. Springer US, Boston, MA, 2010.
- [178] J. Sánchez-Monedero, L. Dencik, and L. Edwards. What does it mean to “solve” the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 458–468, New York, NY, USA, 2020. Association for Computing Machinery.

- [179] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms : Research methods for detecting discrimination on internet platforms, 2014.
- [180] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.
- [181] A. Sharma, F. Thung, P. S. Kochhar, A. Sulistya, and D. Lo. Cataloging github repositories. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, EASE’17, pages 314–319, New York, NY, USA, 2017. ACM.
- [182] T. Sharon. When digital health meets digital capitalism, how many common goods are at stake? *Big Data & Society*, 5(2):2053951718819032, 2018.
- [183] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA, June 2018. Association for Computational Linguistics.
- [184] J. Simmons, L. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 20:1–8, 01 2011.
- [185] A. Singh and T. Joachims. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2219–2228. ACM, 2018.
- [186] A. Singh and T. Joachims. Policy learning for fairness in ranking, 2019.
- [187] J. Skeem, J. Monahan, and C. Lowenkamp. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5):580–593, 2016.
- [188] P. Smeros, C. Castillo, and K. Aberer. Sciclops. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Oct 2021.
- [189] M. T. Stevenson and J. L. Doleac. Algorithmic risk assessment in the hands of humans, 2021.

- [190] K. J. Stewart and S. Gosain. The impact of ideology on effectiveness in open source software development teams. *MIS Quarterly*, pages 291–314, 2006.
- [191] M. Storper and R. Salais. *Worlds of production: The action frameworks of the economy*. Harvard University Press, 1997.
- [192] T. Sühr, A. J. Biega, M. Zehlike, K. P. Gummadi, and A. Chakraborty. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19, page 3082–3092, New York, NY, USA, 2019. Association for Computing Machinery.
- [193] H. Suresh, N. Lao, and I. Liccardi. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making, 2020.
- [194] C. Sweeney and M. Najafian. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 359–368, New York, NY, USA, 2020. Association for Computing Machinery.
- [195] M. Tatasciore, V. K. Bowden, T. A. W. Visser, and S. Loft. Should we just let the machines do it? the benefit and cost of action recommendation and action implementation automation. *Human Factors*, 2019.
- [196] T. C. Teifion Davies. *The ABC of mental health*, volume 2nd edition. BMJ Books, 06 2009.
- [197] L. Thévenot. Organized complexity: conventions of coordination and the composition of economic arrangements. *European journal of social theory*, 4(4):405–425, 2001.
- [198] S. Tolan. JRC Digital Economy Working Paper 2018-10 Fair and Unbiased Algorithmic Decision Making : Current State and Future Challenges, 2018.
- [199] M. Turner, S. L. Ross, G. Galster, J. Yinger, E. Godfrey, B. A. Bednarz, C. Herbig, S. Lee, and B. Zhao. Discrimination in metropolitan housing markets: National results from phase i hds 2000, 2002.
- [200] G. A. United Nations. Universal declaration of human rights, 1948.
- [201] U.S. Senate. United States Civil Rights Act, 1964.

- [202] B. Ustun, Y. Liu, and D. Parkes. Fairness without harm: Decoupled classifiers with preference guarantees, 01 2019.
- [203] P. van den Besselaar and U. Sandström. Vicious circles of gender bias, lower positions, and lower performance: Gender differences in scholarly productivity and impact. *PLOS ONE*, 12(8):1–16, 08 2017.
- [204] M. Veale, M. Van Kleek, and R. Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 2018.
- [205] A. Vogel and D. Jurafsky. He Said, She Said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea, July 2012. Association for Computational Linguistics.
- [206] M. Volkovs, G. W. Yu, and T. Poutanen. Content-based neighbor models for cold start in recommender systems. In *Proceedings of the Recommender Systems Challenge 2017*, RecSys Challenge '17, New York, NY, USA, 2017. Association for Computing Machinery.
- [207] G. von Krogh, S. Haefliger, S. Spaeth, and M. W. Wallin. Carrots and rainbows: Motivation and social practice in open source software development. *MIS quarterly*, 36(2):649–676, 2012.
- [208] S. M. Wachter and I. F. Megbolugbe. Racial and ethnic disparities in home-ownership, 1992.
- [209] S. Wachter-Boettcher. Why you can't trust ai to make unbiased hiring decisions, 10 2017.
- [210] C. G. Walsh, B. Chaudhry, P. Dua, K. W. Goodman, B. Kaplan, R. Kavuluru, A. Solomonides, and V. Subbian. Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence. *JAMIA Open*, 3(1):9–15, 01 2020.
- [211] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli. Is feature selection secure against training data poisoning? In F. Bach and D. Blei, editors, *JMLR W&CP - Proc. 32nd Int'l Conf. Mach. Learning (ICML)*, volume 37, pages 1689–1698, 2015.
- [212] X. J. Yang, C. D. Wickens, and K. Hölttä-Otto. How users adjust trust in automation: Contrast effect and hindsight bias. *Proceedings of the Human Factors and Ergonomics Society*, pages 196–200, 2016.

- [213] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.
- [214] K. Yu, S. Berkovsky, D. Conway, R. Taib, J. Zhou, and F. Chen. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP ’16, page 223–227, New York, NY, USA, 2016. Association for Computing Machinery.
- [215] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web - WWW ’17*, 2017.
- [216] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239, 2017.
- [217] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification, 2015.
- [218] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. Fa*ir. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM ’17*, 2017.
- [219] B. Zhang. An explorative study of github repositories of ai papers. arXiv preprint arXiv:1903.01555, 2019.
- [220] Y. Zhang, Q. V. Liao, R. K. E. Bellamy, Q. Vera Liao, and R. K. E. Bellamy. Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, jan 2020.
- [221] A. Zirikly, P. Resnik, Ö. Uzuner, and K. Hollingshead. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.