


Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (33.rd cycle)

Algorithms for Social Good: A Study of Fairness and Bias in Automated Data-Driven Decision-Making Systems

Elena Beretta
* * * * * *

Supervisors
Prof. Juan Carlos De Martin, Supervisor
Dr.Bruno Lepri, Supervisor
Dr. Antonio Vetro', Advisor

Politecnico di Torino
2021

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Elena Beretta

Turin, 2021

Summary

Nowadays, it is widely recognized that algorithms risk to reproduce and amplify human bias that historically have led to discriminatory functioning, especially towards disadvantaged groups. Evidence of such discrimination has been collected and reported in several fields: credit score, allocation problems, criminal justice, advertising, job placement, etc. Solutions to mitigate the effect of biased decision systems focused on metrics to measure the degree of equity of the algorithms and different notions of fairness have been introduced. As a consequence, achieving fairness don't merely involve the process of planning and engineering algorithms that satisfy mathematical and statistical properties. These algorithms indeed should also explicitly encode specific values and equity criteria.

As a result, a significant ethical and political challenge arises for those who are responsible to decide which measures of fairness and which values an algorithm should embody. Several recent studies have drawn attention to this issue related to the implementation of machine learning systems. Evidence emerging from these studies suggests that fairness should be considered as a trade-off process whereby the system background priorities are established. In fact, since the beginning of the first studies on fairness in the field of machine learning, the main challenge has been to define what fairness means: the large number of fairness measurements appeared in the literature is due to this effort, although conciliating different metrics of fairness might be mathematically not achievable, except under constrained special cases. As a consequence, choosing a fairness metric not only involves mathematical aspects or technical requirements the model is supposed to exhibit, but also conditions belonging to moral and political philosophy, as well as issues of human perception of fairness metrics, thereby shifting the focus from purely technical requirements to a multi-facet problem.

In such a context, the primary thesis goal is to investigate the role of fairness and bias in Automated Data-Driven Decision-Making Systems (ADMs). The current work lies at the interface of science, technology and society by offering an wide-ranging interdisciplinary perspective on fairness and bias in automated systems. The discussion about fairness and bias is approached from different perspectives across different application domains. In this vein, four case-studies are provided.

The thesis initially introduces three major Research Problems that constitute

the ground on which the whole work is based. More specific Research Questions are subsequently outlined for each of the case studies.

The first case-study analyses the limitations of the mainstream definition of Artificial Intelligence (AI) as a rational agent, which currently drives the development of most AI systems. In this work, the need of a wider range of driving ethical principles for designing more socially responsible AI agents is drawn.

In the second case-study we propose a method of data annotation based on Bayesian statistical inference that aims to warn about the risk of discriminatory results of a given data set. The method aims to deepen knowledge and promote awareness about the sampling practices employed to create the training set, highlighting that the probability of success or failure conditioned to a minority membership is given by the structure of the data available.

The third case-study a decision-making model to mitigate potential discriminatory effects of ranking systems is presented. We introduce AFteRS, an Automated Fair-Distributive Ranking System, that has the objective of determining the best top-N-ranking in a set of candidates while simultaneously satisfying fairness constraints and preserving the general utility of the system.

Lastly, in the fourth case-study we propose a Decision Support System that aims to ensure long-term fairness. The methodology extends Decision Theory to automated decision-making systems by introducing a theoretical model to apply fairness to a binary partition of the target population. In the spirit of promoting fairer and more effective automated decision systems, the role of individual dynamics in automated decision-making is explored and integrated in our theoretical formalization.

Based on the context, functioning, Research Problems and Questions analyzed throughout the work, and based on the results obtained in the case studies, the thesis ultimately suggests and outlines New Research Trajectories, *Cross-Disciplinary Validation, Multi-High-Interpretability and Systematic Ground Encoding.*

Acknowledgements

First and foremost I am extremely grateful to my supervisors, Prof. Juan Carlos De Martin and Dr. Bruno Lepri for their invaluable advice, continuous support, and patience during my PhD study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research. I would also like to thank the advisor Dr. Antonio Vetró for his continuous support on my study. Finally, I would like to thank all my colleagues at the Nexa Center. It is their kind help and support that have made my study and life at the Nexa Center a wonderful time. They are (in random order): Francesco Ruggiero , Marco Conoscenti, Giovanni Garifo, Giuseppe Futia, Selina Fenoglietto, Mattia Plazio, Antonio Santangelo, Eleonora Bassi, Anita Botta, Pasquale Pellegrino.

To all people who cannot
access higher education.
To my loving family.
To the family I have chosen.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the last few decades, we have witnessed a growing diffusion of automated software and increasingly sophisticated predictive models for decision-making, which exploit an ever-growing amount of personal and proxy data, suggesting how the data itself should be interpreted and what actions should be pursued, as a consequence of such analysis. As a results, the availability of large-scale data, often regarding human behavior, is profoundly changing the world in which we live. The automated flow and analysis of this type of data offers an unprecedented opportunity for actors in both public and private sectors to observe human behaviors for a large variety of purposes: to provide insights to policy-makers; to build personalized services like automated recommendations on online purchases; to optimize business value chains; to automate decisions; etc. Automated decision systems are thus very popular nowadays. The algorithms on which these systems are based, are involved in a variety of domains. For instance, they are involved in deciding whether we are reliable enough to receive a mortgage or a loan [106]; they suggest if we are inclined to re-offend [20]; they suggest what our future purchases should be [191, 145], what music we should listen to [168] or what movie we should watch [92]; they suggest which are the best candidates for a job [181, 142] or if we are good enough to attend a particular university [100], searching for characteristics that historically lead to success. These decisions are the result of a massive and increasingly powerful profiling and classification mechanism used in Machine Learning and, more generally, in Artificial Intelligence algorithms. At the heart of this kind of AI is the ability to turn intelligent robots and process-driven automation into superheroes who represent their own unique abilities. This kind of transformation is mostly due to the increase of data storage and computational power, that is giving to AI the potential to enable smarter decisions in a variety of areas, including intelligence, analytics, and data management. Big data analysis associated with AI can influence intelligence analysis by sifting through vast amounts of data, providing technicians with unprecedented levels of intelligence and analytical productivity [201, 166]. In such a context, the Automated Decision-Making (ADM) has been hailed as a silver

bullet that promises to replace human subjectivity with objective, infallible decisions [175]. This means that it is increasingly playing a central role in public life, replacing human decision-making processes.

However, automated decision-making systems, while having the potential to bring about greater efficiency and consistency, also open up new forms of discrimination that may be more difficult to identify and combat [16]. Furthermore, the way data are collected, tested and analyzed poses a number of risks and questions related to the context of use [53]. Many researchers, in fact, identified a number of ethical and legal issues where the application of software automated techniques in decision-making processes has led to intended and unintended negative consequences, and especially disproportionate adverse outcomes for disadvantaged groups [13, 116]. Recent scandals such as the one involving Cambridge Analytica and Facebook [27], the study conducted by ProPublica of the COMPAS Recidivism Algorithm [84] or the recent experiment published by AlgorithmWatch on Google Vision AI racism [104], are three well-known examples of the relevance of these issues for our societies. As a consequence, the states legal systems are starting to require a review of the risks that automated decision-making systems pose to the privacy and security of consumers' personal information, as well as any systems that may lead to or contribute to inaccurate, unfair, biased or discriminatory decisions [163, 135]. In this spirit, in the last years both public and academic institutions have provided some technical guidelines [93, 94, 133, 170, 186, 91, 34, 52] by stressing the need to proactively address the ethical and legal risks of decision-making systems such as artificial intelligence (AI) and machine learning, and bringing to light a lack of transparency in the development of these new systems [196, 144].

In such a context, the failure to recognize algorithms value led to the emergence of a new perspective. The role of algorithms in decision-making could have strong implications for science, as ethical decision-making provides lessons for algorithms [101]. This begs the question: if we base our solutions on principles and explain and examine what happens in cases of bias against people who cannot be proven fair by technology, what do we do about it? In the United States, opaque decision-making systems make decisions and serve credit advertisements, without the target ever knowing exactly why, or even if they are wrong [50]. In the case of automated decision-making, these decisions are essentially probabilistic, and the model is based on features similar to those in historical cases [142]. However, this data may contain hidden prejudices, due to the way in which certain races and ethnic groups have been treated in the past. Automated decision-making systems can have disproportionately negative effects on minorities by encrypting and perpetuating social prejudices. One problem is that the data and evidence for the decisions made may be biased if the people who write the algorithms allow their own prejudices to creep into the system. Algorithms can also reproduce or exacerbated social prejudice and discrimination by using training data that reflect existing prejudices in society or present a distorted representation.

However, it is not easy to define what constitutes bias in algorithmic decision-making, although most would probably agree that bias in the examples above is harmful and unfair. Even assuming that AI decision-making is generally reliable, total blindness to these differences can lead to injustice. The idea of a moral or ethical machine remains abstract and unsuitable for real contexts. For instance, it has been shown that ethical codes do not influence decision-making in software development [130]. In the last decade, a number of scholar have started to develop bias detection algorithms to mitigate the bias present in the data collected and in the decision-making process (e.g., [55, 83, 106, 14]). The aim is to monitor the different impacts that arise from the use of AI in a variety of contexts, such as human-machine interaction, data collection and analysis, and decision-making. The main objectives of algorithmic accountability are to increase transparency in automated decision-making, to raise awareness against prejudice and to introduce reasonable controls in the data processing practices of ADM systems.

1.1 Motivation and Challenges

It is now widely accepted that algorithms reproduce and reinforce human prejudice, especially against disadvantaged groups. Evidence for that is found in many fields of application, as credit scoring, recidivism assessment or job recruitment. For this reason, in recent years a variety of solutions have been proposed to mitigate bias and introduce different notions of fairness. Some of these solutions focus in particular on proposing metrics to measure the degree of fairness of an algorithm. In this way, the algorithm should also explicitly encode certain values or equity criteria. Recent evidence suggests that fairness should be seen as a trade process that sets the system's priorities. As the focus shifts from purely technical requirements to a complex problem, choosing a fairness metric is not only about whether a model should have a certain degree of fairness, but also about conditions that are part of moral and political philosophy. However, this domain still raises many issues and poses several challenges and risks. Below, we summarize some of the research problems that have driven this manuscript:

RP1. Paucity of research: although in recent years a undeniably growing interest in this domain has occurred, this field of research is still in its beginning, suffering in many corners of a shortage of literature;

RP2. Lack of validation methods: this highly interdisciplinary domain requires validation methods pertaining to different disciplines. However, evaluation of automated decision-making systems is still far to be cross-disciplinary;

RP3. Separate entities problem: the implementation of equitable ADM systems requires data and algorithms are not treated as two separate entities. Many

studies have shown that a fair algorithm can often only partially compensate for the unfairness of the data. Despite this premise, a substantial portion of previous work addresses this problem one at a time, treating the data and the algorithm as distinct entities.

1.2 Thesis Contributions

The current manuscript lies at the interface of science, technology and society. The undertaken approach to automated decision-making systems with big data focuses on linking data and automated tools to human-technology interaction in social justice domain, offering an interdisciplinary perspective on fairness and bias in automated systems. This work is not intended to be purely technical; in many chapters theories and studies from the social sciences and humanities are drawn upon. The manuscript represents an attempt to provide a wide-ranging perspective on issues related to fairness and data-driven automatic decision systems, which from different points of view presents several problems and challenges related to the interdisciplinary nature of the research domain. Indeed, both fairness and bias should not be considered as a purely technical domain, but as a domain in which technical boundaries are opened to make room for social considerations, and thus as such is treated in the manuscript. The primary contribution of this thesis is therefore to offer an interdisciplinary perspective on fairness and bias in automated decision-making systems across specific application domains. In this work, general issues related to fairness in ADM are addressed by refining the research questions in each of the selected application contexts. In this way, the discussion about fairness and bias is approached from different perspectives. Since this domain has received increasing attention in recent years, this manuscript is unable to address all fairness issues brought up in automated systems. Therefore, the discussion is initially limited to data-driven automated decision-making systems and some specific applications. We describe below the application contexts and the main contributions this manuscript has brought to each of them.

Mainstream AI This study analyses the limitations of the mainstream definition of Artificial Intelligence (AI) as a rational agent, which currently drives the development of most AI systems. The need of a wider range of driving ethical principles for designing more socially responsible AI agents is drawn. An experience-based line of reasoning by argument to identify the limitations of the mainstream definition of AI is followed, which is based on the concept of rational agents that select, among their designed actions, those which produce the maximum expected utility in the environment in which they operate. Then, taking as an example the problem of biases in the data used by AI, a small proof of concept with real datasets is provided.

It is observed that biases measurements on the datasets are sufficient to demonstrate potential risks of discrimination when using those data in AI rational agents. Starting from this example, we discuss other open issues connected to AI rational agents and provide a few general ethical principles. The study contributes to the scientific debate on the governance and the ethics of Artificial Intelligence with a novel perspective, which is taken from an analysis of the mainstream definition of AI.

Data Bias Awareness In this study we propose a method of data annotation based on Bayesian statistical inference that aims to warn about the risk of discriminatory results of a given data set. In fact, although the process of rigorous data collection and analysis is fundamental in the model design, this step is still largely overlooked by the machine learning community. Our method aims to deepen knowledge and promote awareness about the sampling practices employed to create the training set, highlighting that the probability of success or failure conditioned to a minority membership is given by the structure of the data available. We empirically test our system on three datasets commonly accessed by the machine learning community and we investigate the risk of racial discrimination. The empirical findings in this study provide a new perspective on data annotation practices by showing that Bayesian inferences may reveal the risk of bias in three different widespread dataset.

Fairness in Ranking Systems In this study we develop a decision-making model to mitigate potential discriminatory effects of ranking systems. We introduce AFteRS, an Automated Fair-Distributive Ranking System, that has the objective of determining the best top-N-ranking in a set of candidates while simultaneously satisfying fairness constraints and preserving the general utility of the system. The approach takes inspiration from Roemer's Equality of Opportunity theory. We implement three fairness criteria, each one based on a different dimension of the distributive justice theory, namely equity, equality, and need. We test the system in an hypothetical scenario of a university selection process in which the decision-maker determines which students are suitable on the basis of their personal qualifications and achievements, so as to maximize the institution utility. In such a context, we examine the expected outcome for groups of individuals in the ranking system before and after the application of our distributive fairness approach, and we explore the trade-off between the three different fairness policies in relation to the obtained rankings. Results of our research don't show an absolute predominance of one fairness criterion over another one, and that it is possible to achieve fairness constraints with a minimal impact on the general utility of the system.

Long-term Fairness in Decision Support Systems The study of long-term effects in automated decision-making systems is still a largely unexplored field of research. However, these systems when imposing fairness constraints are implicitly required to achieve an equilibrium or equality between two or more groups in the underlying population, often named majority and minority groups or privileged and unprivileged groups. How do decisions resulting from an automated decision-making process affect the underlying population? Do the fairness constraints keep their validity for as long as they act? How do individual dynamics in the long run affect system decisions? The current study is designed to answer these research questions. In this paper, we propose a Decision Support System that aims to ensure long-term fairness. Our methodology extends Decision Theory to automated decision-making systems by introducing a theoretical model to apply fairness to a binary partition of the target population. In the spirit of promoting fairer and more effective automated decision systems, the role of individual dynamics in automated decision-making is explored and integrated in our theoretical formalization. To offer a best understanding of our theoretical model, we set a simulation scenario of a university selection process, in which an institutional decision-maker has to select in a set of policies, the policy to be adopted in order to maximize the long-term selection. Our theoretical formulation allows to study how automated system decisions affect the population and groups. Results on the case-study indicate that: i) policies, although showing similar performance, have different influences on groups; ii) individual dynamics affect system's decisions and fairness constraints in long run.

In addition, the author wishes to clarify that a partial contribution of this thesis consists of previously published work. Specifically, the contribution in Chapter 3 was previously published in an academic journal. In contrast, from Chapters 4, 5, and 6, academic publications drawn concurrently with or subsequent to the thesis preparation were derived. The entire list of papers published during the Ph.D. program can be found in Section 1.5; papers that were drawn from the thesis work are currently in process of publication or under review, and the reporting year is 2021.

1.3 New Research Trajectories

As a further contribution, the following manuscript intends to provide some new research trajectories in the domain of fairness in ADMs. Indeed, although an increasing number of scholars are devoting a lot of efforts in this topic, the road to fairness in machine learning and AI is strewn with obstacles not all easy to overcome. As a result, several unanswered questions still remain. In light of the

questions that will be raised by this manuscript, some insights and new trajectories for future research will be drawn: *Cross-Disciplinary Validation, Multi-High-Interpretability and Systematic Ground Encoding.*

1.4 Organization

The remainder of this work is organized as follows. Chapter 2 provides the conceptual background in order to facilitate the understanding of the rest of the thesis work. It introduces the notion of Automatic Decision-Making systems (ADMs) and discrimination in ADMs, it discusses the causes of discrimination and the concept of bias, and presents the notion of fairness in ADMs. Chapter 3, 4, 5 and 6 are four themed chapters based on the application contexts described in Section 1.2. Each of these chapters addresses the research problems set forth in Section 1.1, refining the research questions to the specific application context. In Chapter 7 the Conclusions and New Research Trajectories are drawn.

1.5 Publications

i (Under Review) Beretta, E., Vetró, A., Lepri, B. and De Martin, J.C. (2021) *AFteRS: an Automated Fair-Distributive Ranking System for Social Justice in AI*, Theory and Decision: An International Journal for Multidisciplinary Advances in Decision Science.

ii (Under Review) Beretta, E., Lepri, B. and De Martin, J.C., *Evolutionary Individual Dynamics in Long-Term Fairness Assessment.* In: Proceedings of 4th AAAI/ACM Conference on AI, Ethics, and Society, 2021.

iii (In printing) Beretta, E., Vetró, A., Lepri, B. and De Martin, J.C.„ *Equality of Opportunity in Ranking: a Fair-Distributive Model.* In: Bias 2021, Second International Workshop on Algorithmic Bias in Search and Recommendation, April 2021.

iv (In printing) Beretta, E., Vetró, A., Lepri, B. and De Martin, J.C.„ *Detecting discriminatory risk through data annotation based on Bayesian inferences.* In: ACM FAccT 2021, Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency, March 2021.

v Vetró, A., Santangelo, A., Beretta, E. and De Martin, J.C. (2019), *AI: from rational agents to socially responsible agents*, Digital Policy, Regulation and Governance, Vol. 21 No. 3, pp. 291-304. DOI: https://doi.org/10.1108/DPRG-08-2018-0049.

vi Beretta, E., Vetró, A., Lepri, B. and De Martin, J.C., *The Invisible Power of Fairness. How Machine Learning Shapes Democracy*. In: Advances in Artificial Intelligence, Proceedings of 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019. Ed. by Marie-Jean Meurs and Frank Rudzicz. Vol. 11489. Kingston, ON,Canada: Springer, Cham, 2019, pp. 238–250. DOI: https://doi.org/10.1007/978-3-030-18305-9_19.

vii Beretta, E., Vetró, A., Lepri, B. and De Martin, J.C. (2019), *Ethical and Socially-Aware Data Labels*. In: Lossio-Ventura J., Muñante D., Alatrista-Salas H. (eds) Information Management and Big Data. SIMBig 2018. Communications in Computer and Information Science, vol 898. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-11680-4_30.

Chapter 2

Background

2.1 Automated Decision-Making in the Algorithmic Era

Artificial intelligence (AI) is based on quantitative methods that have been ubiquitous in society for at least a hundred years [147], and it is a form of automated decision-making that recently started to have a significant impact on decision-making and human behavior [54]. In fact, automation has become increasingly used in decision-making over the past decade in a variety of domains, for instance, whether to figure out which YouTube video to recommend or to decide whether to grant a loan. In a such a context, Automated Decision-making Data-Driven Systems have become widespread and widely employed to rule on various aspects of individuals' lives [167]. An automated decision system is a system that uses data, algorithms, or computer programs to replace or support the human decision-making process [10]. In some cases, it is a way of integrating certain technologies and big data into the decision-making process. In other cases, they leverage machine learning systems [14]; we define them as Automated Decision-making Data-Driven System (ADMs) [111]. Unlike the first case that use traditional automated decision-making, machine learning systems do not follow explicit rules written by humans. Rather, the machine derives its own rules based on the data it has been trained on, such as data from a variety of sources [116]. A series of inputs flow into the system and its algorithm explains an answer, a verdict or a result. In this case, decisions are made automatically without human involvement (e.g. an algorithm that decides whether to get a loan or automated processing to reduce human decision-making costs). In general, decision support can improve decision quality in semi-structured situations [174] where decision automation is not possible or desirable, but is typically considered a "programmed" decision situation. Furthermore, automated decisions by means of algorithms makes practical and economic sense when decisions need to be taken very quickly, when the risk of a wrong decision is

perceived as low or when the result is reversible.

2.1.1 Discrimination in Automated Decision-Making

However, unrecognized or ignored, potential distortions could prevent machine learning from delivering on its promise to significantly improve the accuracy and fairness of automated decision-making systems. When data are used to actually drive decisions, the decisions are not suddenly neutral, objective, or right, rather they are vulnerable to inheriting many of the same kinds of biases that we thought they could eliminate [72]. Since automated data-driven decision-making systems are written and maintained by humans, AI and machine learning algorithms are not devoid of human influence but adjust their own behavior based on that of humans [33]. As a result, algorithms can reinforce human biases and especially that types of biases that reinforce human prejudices [189] . In this sense, we have seen a radical shift in recent years: whereas at to begin with there was a part of eagerness and trust approximately the possibility that machine learning and algorithms more generally would be a force for advancing civil rights and protecting people against discrimination, as of late there has been a realization that this is not an automatic feature of using this technology and that there are many ways in which it can end up having the same kinds of problems as the human decision-making it could replace [142].

About the Term *Bias*

The meaning of the term bias in this area can be easily confused. In statistics and computer science, this term is in fact used to identify a few concepts: a biased mode of data collection, i.e., when the sample collected is not representative of the population; a biased estimator, i.e., a systematic prediction bias. Whereas within the broader domain of machine learning systems the term discrimination, or differentiation, refers to the system's capability to distinguish between instances in order to determine which class a particular example belongs to, in this particular domain the term discrimination is used in a different way, to which we refer as unjustified discrimination [131]. Both concepts explained above can effectively contribute to the problem of unfair bias or unjustified discrimination, but the term bias in the algorithmic ethics domain refers to a somewhat different concept [14, 13]. The term bias in this specific domain is used when there is unjustified discrimination: for example, when a particular characteristic of individuals - such as gender or ethnicity - is not relevant to the decision but is nonetheless assigned a causal effect; or when this said characteristic is used by the system because it is predictive, despite the law stating that certain characteristics should not be considered by the system because they are ethically irrelevant.

About Discrimination in Law and how to explain it in ADMs

From a legal perspective, discrimination is conceptualized by two different doctrines, called Disparate Treatment - or Differential Treatment - and Disparate Impact [210]. These doctrines address how to treat certain protected characteristics of individuals in high stakes decisions. Later in this Section, we will formally describe what these characteristics are and what decisions are considered high-stakes. Under the first doctrine, discrimination is considered intentional when there is a deliberate intent to treat two subgroups of the population differently. Accordingly, the law states that it is illegal to consider protected characteristics in high-stakes decision making. For example, in a credit context, Disparate Treatment exists if the ethnicity of individuals is used to determine the granting of a loan. In ADMs, this is formally defined as the inclusion of individuals' protected characteristics in the system such that different subgroups of people are intentionally treated differently. A second way of understanding this type of disparity in ADMs is through the use of proxy variables [209, 131]. For instance, suppose that there exists a very convenient proxy variable, a kind of factor that is closely related to ethnicity, but that is not ethnicity itself, that is used in credit decision making precisely because it is a proxy variable. It would be something like intentional discrimination, even though the system is not directly considering the prescribed characteristic. However, under the second doctrine, this distinction does not appear so clear-cut. Specifically, the doctrine focuses on the seemingly neutral factors, i. e., that set of factors that apparently seem non-harmful, but which conceal the perpetration of an outcome disparity for certain protected characteristics [29]. In this spirit, the doctrine states that the decision-maker is held responsible for the disparity in some specific areas unless the use of such factors is justifiable. In other words, this means that unless there are good reasons that justify the use of such factors in the decision-making process, which is actually creating a disparity of outcome, then this use should not be justified. This means that the decision-maker is responsible about avoiding this outcome. In automated systems, this doctrine translates to not using these factors and minimizing outcome disparity [113, 6]. While the first prescription has been shown to be ineffective in avoiding discriminatory outcomes in these systems, several solutions have arisen on the second prescription, which we will describe later in this Chapter (Section 2.3). Table 2.1 provides a schematic summary of these doctrines and how to apply them in ADMs.

Protected Characteristics and High Stakes Decisions

As highlighted above, the law has established that there exist some protected characteristics based on certain characteristics of individuals. In ADMs these characteristics are called *protected* or *sensitive attributes*. The regulation of protected attributes and how they should or should not be used in decision-making is established through a number of civil rights laws in high-stakes areas, for example,

Law		**ADMs**	
Doctrine	*Type of discrimination*	*Use of features*	*Principle*
Disparate Treatment	Formal or intentional	Protected characteristics explicitly encoded as input; Proxy variables	Equality of opportunity
Disparate Impact	Avoidable or unjustified	Protected characteristics not necessarily explicitly included as inputs	Minimize inequality of results

Table 2.1: Application of Disparate Treatment and Disparate Impact doctrines in Automated Decision Data-Driven Systems

credit, education, housing, and employment. From a legal perspective, in both the United States [66] and the European Union [126], there is no a universal and unique law regulating this issue, but rather a series of laws in each of the high-stakes areas have been drawn [165]. A list of protected characteristics for both legal systems is provided in Table 2.2.

European Union	**United States**
Race	Race
Color	Color
Religion or creed	Religion
National origin or ancestry	
Sex	Sex
Sexual orientation	
Gender reassignment	
Age	Age
Physical or mental disability	Disability status
Veteran status	Veteran status
Genetic information	
Citizenship	Citizenship
Marriage and civil partnership	Familial status
Pregnancy and maternity	Pregnancy
	Genetic information

Table 2.2: List of *protected* or *sensitive attributes* in US and European Union

We emphasize that the list of these characteristics has not always been this way since its inception, but has undergone changes in both the characteristics themselves and the areas of application, reflecting changes in society over time. So discrimination both from a legal point of view and in ADMs has to do with some

unjustified or questionable basis on which important decisions are to be taken. So it has to do with considering as discriminatory all those decisions that are made on the basis of very specific characteristics and in very specific areas, often because these have been the kind of basis on which certain populations have been held in a subordinate position for unjustified reasons. These kinds of decisions are crucial and often determine the course of people's lives and their ability to access fundamental opportunities in their lives.

2.2 How Systems Learn to Discriminate

The ways in which systems learn to discriminate are many [24]. The following list shows some well-known cases of bias in literature and their discriminatory results:

- **Skewed Samples**: is due to biased samples and is a direct result of using an already biased data collection process. This process can lead to distorted patterns resulting in negative impacts that will reinforce the pre-existing data bias. This bias is also commonly known as a biased feedback loop. In the same way that sampling from a skewed distribution reinforces skewness by excluding less frequent observations, biased models will only have the best results on the privileged class, and their output will tend to exclude non-privileged classes. If decisions about biased ADMs impact the data collection process, the information collected will be increasingly focused on the privileged class. As a result, a vicious cycle will be established in which injustice will continue to grow in the process. An example of Skewed Sample can be found in the work of Lum *et al.* [123], who describe how predictive policing systems, which are increasingly being used by law enforcement to prevent crime before it occurs work. In particular, the authors highlight how these models create bias feedback loops through the suggestion that police should be employed in a particular type of attack. The fundamental problem with these models is that instead of predicting crime wherever it occurs, the model is more likely to predict a crime where the police have been able to observe it in the past. This problem is confirmed by the fact that even though they observe crime in a particular area predicted by the model, police are still called from other areas, and so there is less chance of observing crime in areas other than those where police had previously conducted attacks. In this way, the false negative rate remains very low, even though it does not match the actual data, and the predictions are confirmed without giving the model a chance to learn from other types of data as well;

- **Sample Size Disparity**: is due to the lower availability of data for the minority group. This implies that strong of size differences are present in the

sample for the majority and minority group [151]. As results, the system tends to show worse performance on minorities than on the overall population if it fails to generalize to these groups. The difference in sample size is exacerbated when when the features used by the classifier behave differently, even oppositely, in the majority and minority groups.

- **Tainted examples**: is due to the incorrect definition of the target of the model. It occurs when the model's target is arbitrarily defined causing the characteristics of some individuals to be more suitable than others. Consider as an example a scenario in which an ADMs is trained on data containing past hiring decisions to predict who to hire in the future [77]. To train this type of model an objective must be defined, i. e. a target to understand which individuals to select. Therefore the decision-maker could choose as target of the model the score achieved by the employees in the annual reviews. However, the attempt to formalize in an objective and neutral way the target variable fails, since in the human decision-making process the implicit biases end up influencing the evaluation [59]. So what the model is learning to do is not to predict the actual job performance of employees, but it is learning to predict how a human manager would evaluate that person [150]. In addition, it may turn out that women, for example, show lower evaluations on average, which is not necessarily due to their actual performance; it may be that the target variable chosen for the model is actually influenced by contextual variables, such as an environment that is unwelcoming or hostile to women. In this way, the model would be biased in favor of one part of the population;

- **Limited features**: It occurs when a model is trained on unrepresentative data. It is due to the diversity of data creation and collection, thus leading to signal problems where some subgroups are overlooked and thus underrepresented [19, 131]. In fact, data collection can only partially capture highly sophisticated real-life phenomena. This loss of information can be so severe that the data collected will not be granular enough to capture the differences between subgroups, resulting in poorer data quality for observations of those groups. As a result, the difference in data quality will have a direct impact on model performance; models with poor data quality will likely be biased toward subpopulation. This kind of bias is very common, since there is often very little information about certain parts of the population, as there is on the credit data [49, 138]. In this way, some factors end up being more informative for some subgroups than others, affecting the distribution of results;

- **Proxies**: is due to the use of proxy variables, i.e., when characteristics that are not directly considered as protected attributes are intentionally or unintentionally used to produce decisions [209]. As mentioned in Section 1,

removing protected attributes from the model does not guarantee a non-discriminatory outcome, since performance may still turn out to be poorer on minority classes than in the overall population. This phenomenon is called *redundant encodings* [141], and occurs when a particular protected attribute is encoded across one or more features in the data, making it unnecessary to remove the attribute itself. In this case, the model will catch the inequalities encoded in the data and learn how to reproduce them.

2.2.1 Other Sources of Bias

A list of other possible causes of bias is provided below [190]:

- **Historical bias**: when there is a misalignment between real world expectations and model outcomes. It requires understanding and studying the application and generation of data over time. Even if we have perfectly-measured features, they might still reflect historical factors, for instance conditions that we find only in poorer neighborhoods. Accordingly to this, we can even reflect the world perfectly, but still inflict harm on a population;
- **Representation bias**: when defining and sampling a population still underdevelopment, then certain part of the input space are under represented and others are over represented. This bias may be due to two main reasons: i) the sampling methods reach only a part of the population and ii) the target population has changed or is anyway different from the original training population;
- **Measurement bias**: when measuring features and labels in a prediction problem, random noise is added. This might be caused by three main reasons: i) different measurement process for different groups or ii) different data quality for different groups and iii) oversimplification of the classification model (e.g. selecting or having available too few features for a good prediction);
- **Aggregation bias**: when a single model is used for all groups, which require in fact different specific models, due to different conditional distributions, backgrounds, cultures etc. Usually with this kind of bias the model will be optimal only for the dominant population;
- **Evaluation bias**: when the evaluation and/or test data for an algorithm does not represent the target population. Misrepresented test data lead to the development of models that are optimized only for a subset of the population;
- **Deployment bias**: when there is a misalignment between the problem a model is intended to solve and the way in which it is actually carried out after deployment (e.g. when we adapt a specific model to a generic task).

2.3 Fairness and Ethics in ADM systems

Since the emergence of adverse outcomes has been shown in several application domains, an increasing number of researchers are focusing on the way algorithms encode prejudices and lead to disproportionate results [142]. As a result, many solutions to overcome the problem of discrimination in ADMs by embedding the idea of fairness in the algorithm's structure have arisen. In the recent years, several formal definitions of fairness have been suggested by the machine learning community. In Table 2.3, we report the most widespread ones grouped by similar characteristics: in particular, the first column indicates the name of the partitioning, while the second one the extended name of the fairness definition; the third column contains scientific references. Some of these definitions will be used throughout this manuscript.

First of all, we provide some mathematical notations that compose a typical setup in a machine learning domain:

- X denotes the features of an individual;
- Y denotes the target variable;
- A denotes a sensitive attribute (i.e. gender, race, etc.);
- C denotes a classifier;
- S denotes a score function or a conditional expectation. For example, the frequency of an event given certain observed characteristics can be written as $S = E[Y|X]$;
- t is a threshold. In case of binary classifiers, the score value causes the acceptance of classifier outputs when it is above t, otherwise causes the rejection.

We introduce and briefly describe the fairness definitions listed in Table 2.3, supplied with examples regarding risk assessment in the criminal justice domain. Individuals rated high risk of re-offending are classified by 0, otherwise 1 - that means low risk of recidivism.

The variable *race* has been considered as a sensitive and protected attribute.

Group fairness. Below, we introduce several formal definitions of *group fairness*.

Statistical parity. Classifier C satisfies *statistical parity* if $\mathrm{P}_a(\mathrm{C} = 1) = \mathrm{P}_b(\mathrm{C} = 1)$ for all groups *a, b* - i.e. *a = black, b = white*. This means that both black and white people should have equal probability to be classified as low risk.

Partition	Definition	Reference
Group fairness	Statistical parity	[55, 53, 110]
	Accuracy parity	[51]
	False positive parity	[44, 40]
	Positive rate parity	[83, 23, 211]
	Predictive parity	[176, 40]
	Predictive value parity	[53]
	Equal opportunity	[83, 40, 110]
	Equal threshold	[83, 40]
	Well-calibration	[106]
	Balance for positive class	[106]
	Balance for negative class	[106]
Individual fairness		[55]
Counterfactual fairness		[110]
Preference-based fairness	Preferred treatment	[212]
	Preferred impact	[212]
Fairness through unawareness		[55, 83, 110]

Table 2.3: Fairness in Machine Learning literature

Accuracy parity. Classifier C satisfies *accuracy parity* if $\mathrm{P}_a(\mathrm{C} = \mathrm{Y}) = \mathrm{P}_b(\mathrm{C} = \mathrm{Y})$ for all groups a, b. This means that both black and white people should have equal probability to be correctly classified as low risk, if belonging to actual low risk rate, and correctly classified as high risk, if belonging to actual high risk rate.

False positive parity. Classifier C satisfies *false positive parity* if $\mathrm{P}_a(\mathrm{C} = 1|\mathrm{Y} = 0) = \mathrm{P}_b(\mathrm{C} = 1|\mathrm{Y} = 0)$ for all groups a, b. This means that both black and white people with actual high risk rate should have equal probability to be incorrectly classified as low risk (False Positive Rate).

Positive rate parity. Classifier C satisfies *positive rate parity* if $\mathrm{P}_a(\mathrm{C} = 1|\mathrm{Y} = \mathrm{i}) = \mathrm{P}_b(\mathrm{C} = 1|\mathrm{Y} = \mathrm{i})$, $i \in$ *0, 1*, for all groups *a, b*. This means that both black and white people should have equal probability to be incorrectly classified as low risk - False Positive Rate - and to be correctly classified as low risk (True Positive Rate).

Predictive parity. Classifier C satisfies *predictive parity* if $\mathrm{P}_a(\mathrm{Y} = 1|\mathrm{C} = 1) = \mathrm{P}_b(\mathrm{Y} = 1|\mathrm{C} = 1)$, for all groups *a, b*. This means that both black and white people with low risk predicted score (Positive Predictive Value) should have equal probability to really belong to the low risk class.

Predictive value parity. Classifier C satisfies *predictive value parity* if $(\mathrm{P}_a(\mathrm{Y} = 1|\mathrm{C} = 1) = \mathrm{P}_b(\mathrm{Y} = 1|\mathrm{C} = 1)) \wedge (\mathrm{P}_a(\mathrm{Y} = 0|\mathrm{C} = 0) = \mathrm{P}_b(\mathrm{Y} = 0|\mathrm{C} = 0))$ for all groups *a, b*. This means that both black and white people with low risk predicted score (Positive Predicted Value) should have equal probability to really belong to low risk class, and both black and white people with high risk predicted score (Negative Predictive Value) should have equal probability to really belong to high risk class.

Equal opportunity. Classifier C satisfies *equal opportunity* if $\mathrm{P}_a(\mathrm{C} = 1|\mathrm{Y} = 1) = \mathrm{P}_b(\mathrm{C} = 1|\mathrm{Y} = 1)$ for all groups *a, b*. This means that both black and white people with actual low risk rate should have equal probability to be incorrectly classified as high risk (False Negative Rate). Since mathematically a classifier that satisfies False Negative Rate equity satisfies at the same time True Positive Rate equity, the definition also implies that both black and white people with actual low risk rate should have equal probability to be correctly classified as low risk.

Equal threshold. Classifier C satisfies *equal threshold* if $\mathrm{P}_a(\mathrm{Y} = 1|\mathrm{S} = \mathrm{s}) = \mathrm{P}_b(\mathrm{Y} = 1|\mathrm{S} = \mathrm{s})$, $s \in$ *[0, 1]*, for all groups *a, b*. This means that both black and white people should have equal score threshold t under which they are classified at low risk, and above which they are classified at high risk.

Well-calibration. Classifier C satisfies *well-calibration* if $\mathrm{P}_a(\mathrm{Y} = 1|\mathrm{S} = \mathrm{s}) = \mathrm{P}_b(\mathrm{Y} = 1|\mathrm{S} = \mathrm{s}) = \mathrm{s}$, $s \in$ *[0, 1]*, for all groups *a, b*. This means that both black and white people with the same score should be treated comparably "*with respect to the outcome, rather than treating black and white people with the same score differently based on the race group they belong to*"[106].

Balance for positive class. Classifier C satisfies *balance for positive class* if $\mathrm{E}_a(\mathrm{S}|\mathrm{Y} = 1) = \mathrm{E}_b(\mathrm{S}|\mathrm{Y} = 1)$, for all groups *a, b*. This means that both black and white people with an actual low risk rate should have the same expected value

assigned by the classifier C (a classifier uses the characteristics of individuals to identify which class - or group - they belong to). That is to say, it should not happen that the scoring process is "systematically more inaccurate for negative cases - high risk score - in one group than the other"[106].

Balance for negative class. Classifier C satisfies *balance for negative class* if $E_a(S|Y = 0) = E_b(S|Y = 0)$, for all groups *a, b*. This means that both black and white people with an actual high risk rate should have the same expected value assigned by the classifier C. That is to say, it should not be that the scoring process is "systematically more inaccurate for positive cases - low risk score - in one group than the other"[106].

Individual fairness. Given a set of individuals V and a set of outcomes $A = \{0, 1\}$, and considering a metric on individuals *d:* $V\ x\ V \rightarrow R$ and randomized mappings *M:* $V \rightarrow \Delta A$, *individuals fairness* is achieved if a randomized classifier, mapping individuals to distributions over outcomes, minimizes expected loss subject to the (D,*d*)-Lipschitz condition of $D(Mx, My) \leq d(x, Y)$ [55].
This means that two individuals are similarly classified if they are considered similar with respect to a particular task, such as to pay off a debt with a bank.

Counterfactual fairness. Classifier C satisfies *counterfactual fairness* if $P(C_{A\leftarrow a}(U^1) = y|X = x, A = a) = P(C_{A\leftarrow a'}(U) = y|X = x, A = a)$. That is, given a set of attributes (*education level, type of crime, drugs problems* and protected attribute A = *race*) and an outcome $\hat{Y}$ to be predicted (*recidivism*), a graph is counterfactually fair if *race* is not directly linked to $\hat{Y}$ through any other attributes. Intuitively, this means that a decision is fair towards an individual if it is the same in (i) the actual world and (ii) a counterfactual world where the individual belonged to a different demographic group (i.e. white instead of black).

Preference-based fairness. Here, we introduce new formalization of fairness [211] that are inspired by the concepts of fair division in economics and game theory [197, 21].

Preferred treatment. Classifier C satisfies *preferred treatment* if $B_a(C_a) \geq B_{a'}(C_{a'})$, for all a, a' $\in$ A[2]. This means that the preferred condition is preserved if each group obtains more benefit from their own classifier then it would be assigned

[1] "*U is a set of latent background variables, which are factors not caused by any variable in the set V of observable variables*"[110]

[2] B_a *is the fraction of beneficial outcomes received by users sharing a certain value of the sensitive attribute a*[211]

from any other classifier. In other words, both black and white people should prefer *"the set of decisions they receive over the set of decisions they would have received had they collectively presented themselves to the system as members of a different sensitive group"*[211].

Preferred impact. Classifier C satisfies *preferred impact* if $B_a(C) \geq B_a(C')$, for all $a \in A$. This means that the preferred condition is preserved if a classifier C, with respect to any other classifier, assigns at least the same benefit for all groups. In other words, both black and white people should prefer *"the set of decisions they receive over the set of decisions they would have received under the criterion of impact parity"*[211].

Fairness through unawareness. Classifier C satisfies *fairness through unawareness* if X: $X_a = X_b \rightarrow C_a = C_b$ for both individuals *a, b*. This means that for example the attribute *race* should not be used to train the classifier and thus to take a decision (i.e. granting or not a loan).

2.3.1 Algorithmic Fairness in Classification

Recently, a comprehensive discussion on fairness and discrimination in machine learning algorithms has been provided by *Barocas et al.* [14]. In particular, the authors summarize the advances in this field and classify the main definitions of fairness into three macro categories: (i) *independence*, (ii) *separation*, and (iii) *sufficiency*. Table 2.4 provides the most widespread definitions of fairness belonging to the above macro categories. As shown in second column, mostly of them are equivalent. In third column we give examples of their application in a classification task by considering the gender as sensitive attribute.

The following paragraphs provide a more detailed explanation of the above definitions.

Independence criterion. A *fairness* definition satisfies the *independence* criterion if sensitive attributes are statistically independent with regard to the classifier. The more widely adopted definitions falling under this *independence* criterion are (i) *demographic parity*, also known as *statistical parity* or *group fairness* [55], and (ii) *conditional statistical parity* [44].

Separation criterion. Regarding the *separation* criterion, it requires that the correlation between sensitive attributes and a classifier is *"justified by the target variable"* [14]. Equalized odds [83], conditional procedure accuracy [20], and avoiding disparate mistreatment [211] codify the same principle by which both members of the unprotected and protected group should have the same probability of being

Category	Fairness Definition	Example
Independence	Demographic Parity	Both females and males must have the same probability to be classified with good scores
	Conditional Statistical Parity	
Separation	Equalized Odds	In both groups, individuals with actual good scores and individuals with actual bad scores must have the same probability to be correctly and incorrectly classified with good scores respectively
	Conditional Procedure Accuracy	
	Avoiding Disparate Mistreatment	
	Predictive Equality	The model must classify individuals with actual good scores in the same way in both groups
	Balance for the Positive Class	
	Equal Opportunity	Both groups must have the same probability to be incorrectly classified with bad scores
	Balance for the Negative Class	
Sufficiency	Conditional Use Accuracy	The probability of belonging to the good/bad score class having a good/bad predictive score must be the same for both groups
	Predictive Parity	The probability of belonging to the good score class having a good predictive score must be the same for both groups

Table 2.4: Examples of the most widespread fairness definitions

correctly classified positively and the same probability of being wrongly classified positively. Predictive equality [40] and balance for the positive class [106] constitute a relaxation in the *separation* category, providing that both groups have the same false positive rate, i.e. the same probability of being incorrectly classified positively. Along the same line, the equal opportunity [83] and the balance for the negative class [106] provide a false negative rate for both groups, i.e. the same probability of being incorrectly classified negatively.

Sufficiency criterion. Finally, conditional use accuracy [20] and predictive parity [40] belong to the *sufficiency* category that establishes sensitive attributes and

target variable are statistically independent [14]. According to conditional use accuracy, both the unprotected and the protected groups should show the same accuracy for the predicted outcome, while the predictive parity constitutes a more relaxed approach where equal accuracy is required only for the positive predicted value.

Barocas *et al.* [14] observe that the majority of the proposed fairness definitions in machine learning literature are an approximation of these criteria. The key assumption behind the concept of fairness is based on the idea that some types of bias can be eliminated, especially those related to human error.

2.3.2 The Cost of Fairness

Recent contributions have analyzed the trade-off of implementing different types of fairness, and have shown the mathematical impossibility of satisfying more than one fairness criterion simultaneously [70], [106], [14]. From these findings, we can derive that no universally accepted notion of *fairness* can exist. Indeed, each measure of fairness embodies a different criterion of equity. *Friedler et al.* [70] demonstration is crucial: by proving the impossibility of simultaneously satisfying the mathematical constraints of multiple definition of fairness, they show the impossibility of simultaneously satisfying different criteria of equity. Hence, this result paves the way to a major challenge in the design, development and evaluation of machine learning systems.

2.3.3 Social Notions to design Algorithms

Achieving fairness doesn't merely involve the process of planning and engineering algorithms that satisfy mathematical and statistical properties. These algorithms indeed should also explicitly encode specific values and equity criteria. As a result, a significant ethical and political challenge arises for those who are responsible to decide which measures of fairness and which values an algorithm should embody. Several recent studies have drawn attention to this issue related to the implementation of machine learning systems. Evidence emerging from these studies suggests that fairness should be considered as a trade-off process whereby the system background priorities are established. In fact, since the beginning of the first studies on fairness in the field of machine learning, the main challenge has been to define what *fairness* means [23]: the large number of fairness measurements appeared in the literature is due to this effort, and, as already mentioned above, conciliating different metrics of fairness might be mathematically not achievable, except under constrained special cases [106]. As a consequence, choosing a fairness metric not only involves mathematical aspects or technical requirements the model is supposed to exhibit, but also conditions belonging to moral and political philosophy [98], [82], as well as issues of human perception of fairness metrics [185], thereby shifting the focus from purely technical requirements to a multi-facet problem.

Several of the difficulties on this issue are in fact related to two kinds of concerns historically debated in moral and political philosophy: the problem of theoretically defining what equity means, and the problem of defining what is fair in a given context [28].

In machine learning, several works on fairness actually refer to a technical trade-off [211], [218], i.e. losing accuracy in favor of a fairer classification [44]. This means sacrificing a part of utility, because there is a loss of accuracy in prediction, in favor of a common good, in other words, the good of the individuals concerned. A number of studies have postulated a convergence between social context and selection process of fairness metric, although the majority of contributions are still aimed at assessing the more technical aspects of automatic learning models [3], [81], [125]. According to Farnadi et al. [62], fairness cannot be achieved only by taking into account the individual's attributes, whereas the individual's relational context should also be considered. Heidari et al. [85] analyze the similarities arising among some fairness definitions and different economic models of Equality of Opportunity, proving fairness metrics actually embed justice criteria, although these are not clarified. Finally, a broader perspective has been adopted by Beretta et al. [19], who argue that fairness definitions are affected by specific ideas of democracy.

An important conclusion emerges from the studies discussed so far: assessing fairness is a procedure which embeds ethical, political and social aspects. Therefore, it is a process which is only apparently technical.

Chapter 3

AI: from rational agents to socially responsible agents

3.1 What kind of Rationality for AI Systems?

The expression *Artificial Intelligence* is gaining considerable attention from both private and public sector [162]. The hype is very high and, as it often happens in such situations, all this attention has generated confusion, even among experts, who refer to Artificial Intelligence to talk about very different things. We refer to AI following the mainstream definition of Russell and Norvig [164]: it is "*the study of designing and building intelligent agents* (p.30), where *agent* is "*anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators*" (p.34). An intelligent agent "takes the best possible action in a situation" (p.30), i.e. it is a rational agent the one which, for each possible percept sequence, is supposed to "*select an action that is expected to maximise its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has*" (p.37). An advantage of this definition is that the few concepts above are the building blocks for designing AI systems with scalable complexity, e.g. from a "simple" vocal translator to an autonomous vehicle. However, this definition of AI is based on a very precise, and in a sense narrow, vision of the concept of intelligence, which is bound to a particular type of rationality. In fact, if the actions undertaken by an agent must always maximize a performance measure, it is clear that the functionality and the effectiveness of such actions are strictly dependent on the form of knowledge the agent itself incorporates: an action is always the consequence of a certain vision of the world, of the world's rules that are considered to be true and for this reason are embedded in the form of algorithms elaborating data, and of a precise conviction about what the world should become according to that logic. There is a vast amount of evidence showing that designing and building AI agents according to such deterministic perspective is producing relevant negative social effects. Recently, the

investigative website ProPublica discovered the COMPAS algorithmic tool - Correctional Offender Management Profiling for Alternative Sanctions - widely spread in the American criminal justice to prevent recidivist behaviors, was biased against black defendants, revealing the tool was assigning them a higher risk rate generally [84]; Latanya Sweeney [191] has highlighted that delivery of ads by Google AdSense was biased in the sense that in a Google search for an individual's name an arrest record was suggesting by algorithm on the basis of racial association, and Cathy O'Neil underlined a large-amount of case-studies in which people are subjected to racial, gender, or any other kind of discrimination, in AI ground [142]. In addition, there are already many examples, in this regard, from AI agents that help financial institutions decide to which category of people to lend money, and that are based on the idea - implicitly embedded in the code and in the data used by the software- that it is better to favor white, educated citizens residing in certain specific areas of the cities, and especially males [188]; or other examples that include AI agents deciding (or recommending) to grant probation to prisoners, which once again favor individuals belonging to certain ethnic groups, or targeting to men more than to women job offers that are more economically advantageous [184]. Evidently, the "forms of knowledge" on which the algorithms that "animate" these machines are based, are the result of databases (or, in the simplest of cases, statistical surveys), which, even if accurate, they represent certain distortions of our society. So the question is: do we want these distortions to increase and to be perpetuated by our Artificial Intelligence tools, or do we prefer to create instruments that may help us diminish the unjust situations we live in? In this sense, the scientific world is taking important steps to include other perspectives, such as the ethical one , at the center of Artificial Intelligence programming, in order to avoid giving rise to a world in which we can design certainly effective and high-performative AI agents, but at the same time let them decidedly unfair, and in our place. This work is part of this community effort, and we advocate the need of a wider range of designing principles for AI agents, which goes beyond the perspective of the mainstream definition of AI.

3.2 Bias of the Forms of Knowledge governing AI

The problem of bias in the data used by AI systems is well represented in the following excerpt of Cathy O'Neal's book "Weapons of Math Destruction" [142]: "*if the admission models to American universities had been trained on the basis of data from the 1960s, we would probably now have very few women enrolled, because the models would have been trained to recognize successful white males*". The observation made by O'Neal entails an important, more general, reasoning: not only how AI collects and elaborates data has ethical consequences, but, before that stage, also the input data properties (percepts, in the terminology analyzed in the previous

section) are connected to important ethical interdisciplinary issues. The characteristics of the "forms of knowledge" involve ethical issues [69], and those problems propagate downwards throughout all subsequent phases of the data life cycle in AI systems, until affecting the output, i.e. the decisions or recommendations made by the software. Therefore, certain data characteristics may lead to discriminatory decisions and therefore it is important to identify them and show the potential risks. We take as reference two characteristics of input data: disproportions and collinearity.

3.2.1 Disproportionate Datasets

AI systems work on the base of large amount of historical data, very often elaborated with machine learning models. Problems of fairness and discrimination may arise due to disproportionate datasets, which lead to disproportionate results, generating problems of representativeness when the data are sampled - thus leading to an underestimation or an overestimation of the groups - and of imbalance when the dataset used has not been generated using the classical sampling methods. Simple random sampling - which is the most widely used method in statistical surveys - requires that the probability of sample extraction is known and not zero, and that not only each element but also each combination of elements (of equal number) has the same probability of being extracted. A biased sample leads to biased estimates. For this reason, statistical sampling is a fundamental step. However, in the era of Big Data, many of the data used today have not been generated using probabilistic sampling, but are rather selected through non probabilistic methods (very often acquired from third parties, or with opportunistic methods, thanks to the pervasiveness of digital technologies), which do not provide to each unit of the population the same opportunity to be part of the sample; this means that some groups or individuals are more likely to be chosen, others less. Representativeness is a property of the outcome of the extraction process, which itself has randomness as its property. For this reason, it is essential to keep this aspect under control in non-probabilistic samples. In general, solutions relating to demographic or statistical parity are useful in cases where there is no deliberate and legitimate intention to differentiate a group considered protected, which would otherwise be penalized [55]. It should therefore be borne in mind that the solutions vary according to both the nature and use of the data. Take as an example a type of analysis that includes in its attributes the individual income. If the choice to include in the sample only individuals with a high income is voluntary, no representativeness problems arise, since the choice of a given group is based on the purposes of the analysis. However, if the probability of being included in the sample is lower as the income is lower, the sample income will on average be higher in the overall income of the population.

3.2.2 Collinearity

In statistics two variables x_1 and x_1 are called collinear variables when one is the linear transformation of the other and therefore there is a high correlation. Collinearity is a group phenomenon involving at least two regressors and which may affect, in different extents, different groups of regressors. In general, there are always relationships between regressors that involve a certain degree of linear dependence, but it is good practice to consider the correlation value 0.9 as the limit beyond which singularity or almost singularity in the matrix of regressors is observed; over this threshold the estimation of parameters in Ordinary Least Squares are to be considered not reliable. In general, the main causes of collinearity are due to data collection techniques, such as similar measurement errors on different regressors; spurious correlations; inconsistency of a regressor data with the model specification, e.g. when using a higher than necessary polynomial; or application of a model to a small number of cases. The attempt to contain the negative effects is mostly due to the fact that collinearity damages the estimates of parameters and their precision. To prevent this effect some researchers adopt a naive approach that precludes the use of sensitive attributes such as gender, race, religion and family information, but may not be effective in case of multicollinearity. The use of geographic attributes, for example, is reported to be unsuitable when the use of protected data is to be foreclosed, because it easily leads to tracing protected attributes, such as race [116]. Hardt *et al.* [83] points out that the condition of non-collinearity requires that the predictor ($\hat{Y}$) and the protected attribute (A) are independent conditional on Y - e.g., the variable to be predicted, income, must be independent of the gender variable. In practice, it is encouraged to use features that allow to directly predict Y, but prohibits abusing A as a proxy for Y. Another common error is “mistake correlation with causation"; a high entropy dataset can induce thinking that the large number of features is sufficient to explain causality. Cause-effect ratios are often confused with correlations when features are used as proxies to explain variables to be predicted. For example, the IQ test is a test that measures logical-cognitive abilities, but if used as a proxy to select the smarter students for admission to a university course, it would almost certainly reveal itself as an imperfect proxy, since intelligence is a too broad concept to be measured by a number only. As a consequence, the test of the IQ is not sufficient to explain the variable to be predicted. To avoid the risks mentioned above, the following thresholds, defined on the base of literature and experience, are useful to identify cases of collinearity:

- correlation values higher than 0.9 should be avoided;
- the absence of high correlations does not exclude collinearity; it is therefore always good to also consider the value of R^2, in the case of $R^2 = 1$, we are in presence of multicollinearity;

- in case of collinearity there is no increase in the explained deviance which is certainly attributable to the effect of a specific regressor.

In addition, an effective method to identify collinearity is to calculate the Variance Inflation Factors that indicate how much parameter variability depends on the regressors. *VIFs* are calculated in this way:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{3.1}$$

If the i-th regressor is not linearly linked to the others, $R_i^2 = 0$ and *VIF* will be equal to one. High levels of *VIFs* indicate the presence of a relation of linearity: it is commonly assumed that for $VIF(\beta_i) > 10$ multicollinearity is strong. Finally, since correlation measurements can only be used for quantitative variables, the degree of dependency between categorical data is measured using the estimation of Pearson residuals[1], which is a commonly accepted measure of discrepancy between observed and expected values [214].

3.3 Overview of Data sets used

We applied the metrics defined above to the following three datasets, each referring to a different application domain. Table 3.1 shows to which datasets which measure was applied.

3.3.1 Credit Card Default dataset

This dataset [118] contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. The dataset is composed by 24 variables, of which four demographic ones that can be considered as protected attributes (sex, age, education, marital status).

[1]Pearson residuals are widespread in statistic domain to study the linear relationship among two categorical variables. When two categorical variables are analyzed, the correlation is called association; therefore, Pearson's residuals measure the strength and direction of the association between two categorical variables and is particularly appropriate when one of the two is dichotomous categorical. The absolute value of the residuals indicates the strength of the association, while the direction is indicated by the sign. See more at: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient and at: https://en.wikipedia.org/wiki/Pearson%27s_chi−squared_test

3.3.2 COMPAS Recidivism Racial Bias dataset

Data [112] contains variables used by the COMPAS algorithm in scoring defendants, along with their outcomes within two years of the decision, for over 10,000 criminal defendants in Broward County, Florida. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm used by judges to assess the probability of recidivism of defendants. Three subsets of the data are provided, including a subset of only violent recidivism, as opposed to, e.g. being re-incarcerated for non-violent offenses such as vagrancy or Marijuana. The original dataset contains 28 variables, of which eight are considered protected attributes: Last name, first name, middle name, sex, race, date of birth, spoken language, marital status. The dataset is well-known because of a study of the U.S. non-profit organization Pro Publica [84] that showed that the COMPAS algorithm is distorted in favor of white individuals, thus exposing black people to a risk of distorted recidivism, because it would be higher than it actually was.

3.3.3 Student Alcohol Consumption dataset

The data [45] were obtained in a survey done by students of mathematics and Portuguese language courses in secondary school. It contains social, gender and study information about students. Two datasets are provided: The one containing the students of the mathematics course contains 395 observations, the one relating to the Portuguese language course contains 649 observations. Both contain 33 variables, most of which are protected attributes describing demographics, such as school, context of belonging (urban, rural), family indications, etc.

	Disproportion	Collinearity
Credit card default dataset	x	x
COMPAS Recidivism racial bias dataset	x	
Student alcohol consumption dataset		x

Table 3.1: Measures and Datasets

3.4 Applications

3.4.1 Credit Card Default dataset

The field of creditworthiness often appears in the literature alongside issues related to ethical decisions [207], [158]. Recently, some studies have shown that access to credit for black people is modulated by certain attributes such as race, rather than by information about the payer's status [25], [15], [39]. The dataset that we use does not contain the protected attribute race, however it contains other

personal information that can be used in a discriminatory way if applied to assess creditworthiness, such as gender and level of education.

Disproportion

Figure 3.1 reports the frequency of variables gender, marital status, age, education, expressed as a percentage for each of their categories. The data shows that 60% of individuals are women, 46.7% of individuals have attended university, the age group most represented is that of 25 to 40 years, the proportion of married individuals is the same for single individuals. Although we do not have information neither on the real frequencies of protected attributes in the source population nor on the sampling method used (if any), the results of the analysis of disproportions suggests to use the age variable with caution: in fact the variable age shows a more considerable disproportion compared with the other protected attributes, exposing a potential risk of discrimination (e.g., if the dataset is used to automate decisions or recommendations on the capability to repay a debt, and attribute age is one of the predictors).

Collinearity

We perform the analysis for each protected attribute in the Credit Card default dataset, in relation to default payment (1 = yes, 0 = no). We report on Figure 3.2 the mosaic plots[2] for the attributes education, marital status and gender: blue indicates cases in which there are more observations in that cell than would be expected under the null model of independence between attribute education and attribute default payment; red means there are fewer observations than would have been expected; eventually, grey indicates that observations are coherent with the assumption of independence. Figure 3.2 shows that:

- default payment is highly correlated to the education level, for all its levels;
- the correlation between the protected attributes and the default payment variable is significant for the gender variable (both male and female);
- the correlation is significant for the marital status variable in correspondence with the default payment group = yes;
- in addition, Pearson residuals show that the most correlated categories are: the education variable and the male, both in correspondence with default payment = yes.

[2]A mosaic plot is an area proportional visualization of a (possibly higher-dimensional) table of expected frequencies. It visualizes Pearson Residuals.

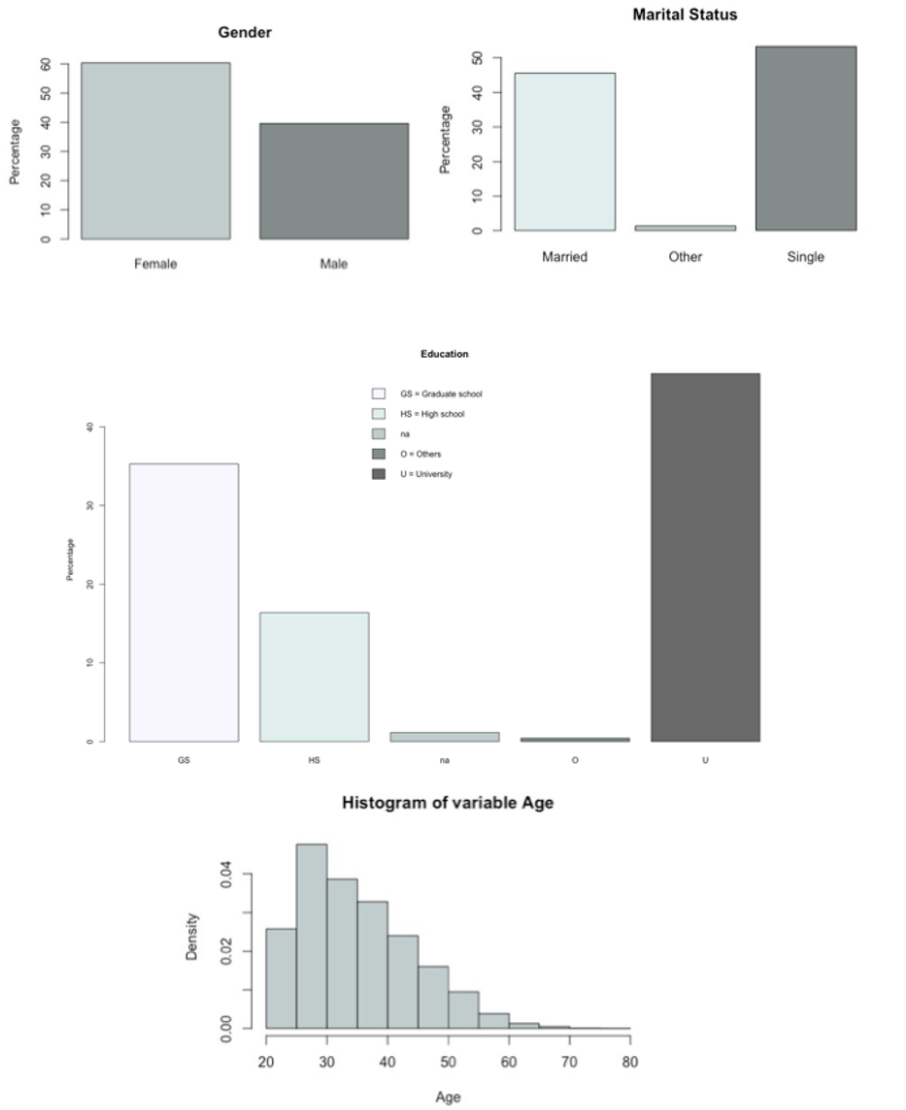


Figure 3.1: Distributions of selected attributes in the Credit Card Default dataset

As a consequence of the analysis, the identified correlations should be taken into account when using the dataset in an algorithm that supports or automate decisions.

3.4.2 COMPAS Recidivism Racial Bias dataset

Disproportion

As reported above, previous research has shown that the data in the COMPAS dataset is unbalanced in favor of white people. Table 3.2 shows the variability

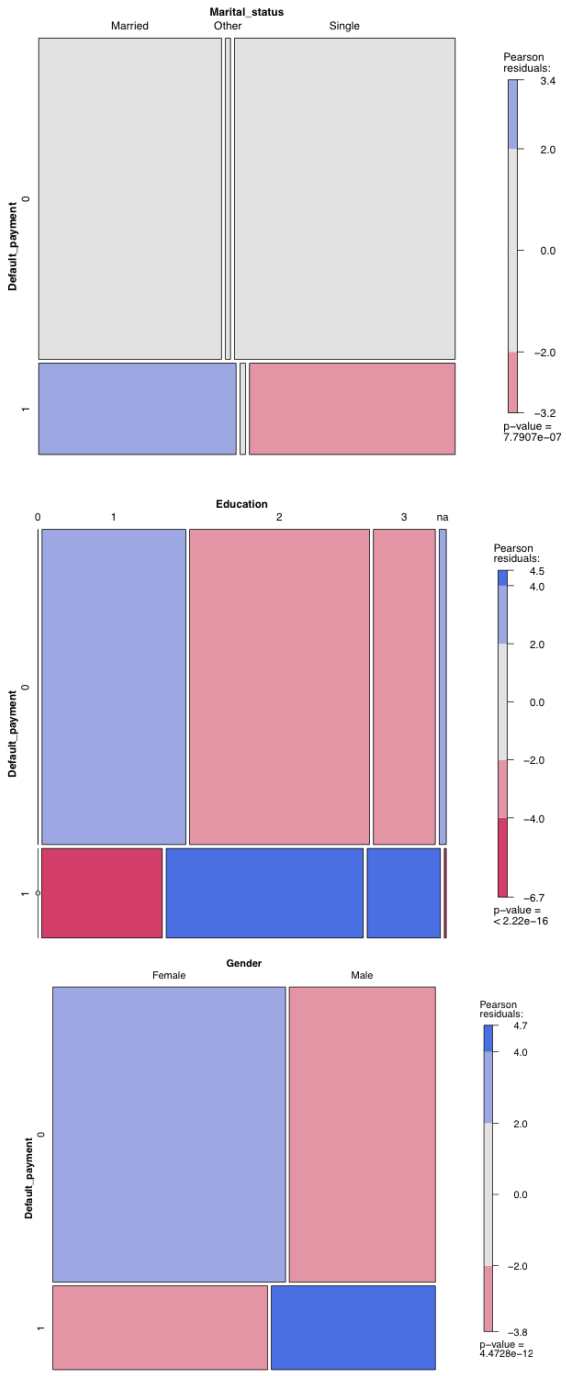


Figure 3.2: Analysis of collinearity with mosaic plots for selected attributes in the Credit Card Default dataset

in race attribute, which is the underlying reason of the findings of the previous study: the highest levels of reoffending are observed in black individuals. Moreover,

33.22% of the dataset's observations refer to white people, while 53.45% refer to black people, indicating that there may be an over-estimation of the race attribute - against black people - which would contribute to the estimation of recidivism (confirmed by follow-up analyses showing a highly dependence between that the race variable and the level of recidivism).

Ethnic Group	High	Low	Medium	N/A
African-American	3400	3369	3010	12
Asian	9	50	12	0
Caucasian	943	3554	1579	10
Hispanic	191	945	315	0
Native American	15	26	16	0
Other	56	653	150	1

Table 3.2: Distribution of ethnic groups within the COMPAS database, by level of risk of recidivism

3.4.3 Student Alcohol Consumption dataset

The dataset is composed of 33 variables, principally qualitative; 649 observations for the dataset referring to students of Portuguese, 395 for that referring to students of mathematics.

Collinearity

We randomly chose 4 quantitative variables to predict workday alcohol consumption and calculated the Variance Inflation Factor for each of the regressors, and report results on Table 3.3. The variables indicate: i) number of school absences (numeric: from 0 to 93); ii) current health status (numeric: from 1 - very bad to 5 - very good); iii) quality of family relationships (numeric: from 1 - very bad to 5 - excellent); iv) age. The average of *VIF* is equal to 1.02, therefore among the variables considered a relationship of collinearity is only moderate. However, we observe that while some attributes in some contexts are considered protected, in others are essential to avoid situations of risk or damage; in the case of alcohol abuse among students, personal information are useful to identify areas of intervention and define appropriate social policies. We underline once again how the intended scope of the AI plays a fundamental role in the choice of considering some attributes as protected or not.

	Absences	Health	Fam Rel	Age
VIF	1.02	1.01	1.01	1.03

Table 3.3: Variance Inflation Factor for selected attributes in Student Alcohol Consumption dataset

3.5 For Artificial Intelligence as a Socially Responsible Agent

The measurements reported in the previous section highlight that when biases are incidentally encoded into the AI agents which the only purpose is to maximize some performance measures, certain injustices and prejudices can be perpetuated and exacerbated. In recent years, both in the scientific community and in civil society a lively debate on this issue has arisen, as the use of big data and automatic learning tools has brought many changes in various fields, including that of Artificial Intelligence. As we have seen before, a relevant problem in this field is related to unbalanced datasets, which overestimate or underestimate the weight of some variables in the reconstruction of the cause-effect relationship needed to predict events, as happened with some algorithms used by the American police to prevent crimes [123]. In this work, the authors showed that the algorithm used by police patrols for predicting future drug crimes, were fed with data that were under-representative of the white consumers of drug. As a result, predictions of the software constantly pointed to areas of the cities where non-white people resided, and police would follow those recommendations to focus crime prevention activities. Arrests would then be concentrated in those areas, and on the non-white people, creating a feedback loop that reinforces the initial bias. Furthermore, there are cases where biases can be injected into the data during the agent training process [13], as it occurs with supervised learning techniques that require humans to label the data. For instance, Kate Crawford [46] showed that differences in gender or ethnic and social background can produce different biases in assessing the meaning of an image or concept. However, the ethical issues raised by the functioning of AI go well beyond the composition of its databases. In some cases Artificial Intelligence poses problems of transparency and openness, since data, algorithms and the architectural functioning are often opaque. This can be dangerous in many areas. For instance, in the job recruiting domain some concerning are arising over the use of Artificial Intelligence tools in the selection and management of personnel, the mechanisms of which are unknown to employees and intermediate bodies. For this reason, an attempt to pursue a policy linked to the promotion of open data and of the open code has been carried on in some countries [149]. However, there are still situations in which transparency and openness do not imply two other desired properties of AI: explainability and understandability. It is the case for example of neural networks

and black-box algorithms [146]. This issue has been regulated by the new General Data Protection Regulation [1] at Article 22, that provides a general prohibition of solely automated decision-making (that means with no significant human intervention in any phase of the data processing) with legal effects on the individual; furthermore, Article 29 of the *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679* specifies that transparency criterion is mandatory also in cases of high complexity of the technology concerned, as the transparency conditions require that detailed explanations or disclosures of the whole algorithm are not to be disclosed, but rather the underlying logic in order to clarify the criteria leading to a particular decision. Finally, the counterbalance of openness and transparency is the need to protect the privacy of individuals, leading to the setting of boundaries beyond which transparency cannot be pursued. One of the typical nodes, in the field of Artificial Intelligence and not only, is for example that of the *mosaic effect*, linked to the secondary use of certain data, very frequent in health research, which is not easy to be predicted from the beginning and which, for this reason, makes informed consent complex to implement. Consensus needs to be questioned as well to determine in which situations individuals can refuse to be subjected to tools and processes that make use of Artificial Intelligence. This issue has been regulated by the new General Data Protection Regulation at Article 22, that provides a general prohibition of solely automated decision-making (that means with no significant human intervention in any phase of the data processing) with legal effects on the individual (see also Art. 29 of the *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*). In light of our results, we point out that a deterministic approach to the design and construction of AI agents does not avoid the risk of discriminatory outcomes. Alongside the definition of Artificial Intelligence as a *rational agent*, a more comprehensive approach is needed that takes into account a complementary definition, that of a *socially responsible agent*. AI should be rational in augmenting social fairness. The White Paper *Artificial Intelligence at the service of the citizen* [52] provides some useful and general principles to achieve more equitable AI systems, suggesting to consider AI as a humanistic and anthropocentric process. Furthermore, it suggests "principles of equity, such as procedural (non-arbitrariness of procedures), formal (equal treatment for equal individuals or groups) and substantive (effective removal of obstacles of an economic-social nature)" [52](p.38). In this study, it has been shown that an improperly designed AI may easily violate all these criteria of equity. More in general, when people is the target of AI decisions, the respect of universal human rights should be the ultimate reference [154], [103]. Along with technical developments in AI, a discussion of how the human-AI cooperation can be most efficiently managed without letting rational AI agents decide on fundamental issues should be included.

Chapter 4

Detecting discriminatory risk through Data Annotation based on Bayesian inferences

4.1 Introduction

In the last decades machine learning systems are widely spreading in different academic domains, as well as many public and private sectors are increasing the exploitation of these systems. Their widespread and pervasiveness is mainly driven by the exponential growth of computational power and the extensive availability of large amounts of data [106]. Supervised machine learning models are also particularly widespread and now deeply rooted in different sectors due to their usage versatility. The predictive ability of supervised machine learning systems is deployed in disparate areas of application: credit reliability [106], justice system [20], job recommendations [181], university selection process [100], cultural contents [168],[92] and purchases recommendations [145]. The key ingredient that supervised machine learning models have in common is the availability of a set of labeled data used to train the model in elaborating a response related to past events [75]. Since the known properties of the available set of data is used to create a classifier that makes predictions about new entities of the same type, the structure, properties and quality of the data are aspects that largely and directly influence the quality of the model and the results it produces [142], [5]. Although data-driven decision models have been shown to produce both economic and social benefits, many researchers have highlighted several problems and damages related to their use in different areas, especially if they are built on partial or incomplete data [83], [55]. As a matter of fact, in recent years several studies have found a convergence of issues related to the ethics and transparency of these systems in the process of data collection and in the way they are recorded [127]. While the process of rigorous data collection and analysis is fundamental to the design of the model, this step

is still largely overlooked by the machine learning community [19], [96]. As the practice of removing protected attributes from available data has been shown to potentially exacerbate further discrimination [203] - making bias even more difficult to detect - practices related to data collection, data transparency and data explainability become even more relevant and urgent. The aim of our work is to provide a data annotation system that serves as a diagnostic framework containing immediate information about the data appropriateness in order to more accurately assess the quality of the available data used in training models. We propose a data annotation method based on Bayesian statistical inference that aims to warn of the risk of discriminatory results of a given data set. In particular, our method aims to deepen the statistical knowledge related to the information contained in the available data, and to promote awareness of the sampling practices used to create the training set, highlighting that the probability of a discriminatory result is strongly influenced by the structure of the available data. We test our data annotation systems on three dataset widely spread in machine learning community: the COMPAS dataset [84], the Drug Consumption dataset [63], [64] and the Adult dataset [107].

4.1.1 Problem Statement

The majority of machine learning systems are based on historical data processing [140]. This is particularly true in supervised machine learning models. Several studies have shown evidence that many equity and discrimination issues are due to input data properties [18]. Most of today data sets used to train models are chosen through non-probabilistic methods, generating problems of data imbalance and representativeness [60], [140]. This means that different fractions of the population do not show the same opportunity to be represented within the sample - aka, training sets -, leading some groups of individuals to have a lower probability of being represented. Common observed effects of a bad sampling are underestimation and overestimation of some groups [14]. Undetected distortions in data may also easily represent a spurious statistical noise. This happens when the data structure induces dependence between two variables that are not linked by a real cause-effect relationship.

Data Sampling A key moment in the pipeline of a machine learning model is when the programmed algorithm is supplied with training data representing the entities on which the model itself trains its knowledge to make predictions. The quality of the data used in this phase is fundamental for the desired result, according to the principle of "garbage in - garbage out": even the most sophisticated models can present distorted results in the presence of low quality data [194]. One of the main causes of data distortion is the way the data is selected and provided to the algorithm displaying problems related to inaccuracy, lack of update or inadequate

representativeness. However, while knowledge of bias typologies has proliferated over the years, less attention is paid to issues concerning data collection, notation and sampling [129]. In the spirit of fostering a broader awareness of data handling, we provide a reasoned list of issues that may arise during this phase:

i *Data Selcetion:* the large proliferation of data sets availability on the same kind of problem to be analyzed, make hard the a priori choice of a given data set [75];

ii *Inadequate sampling methods:* most models are trained with data sets that have been "found" and not subjected to probabilistic sampling methods, leading to limited or no data control [11];

iii *Cost and Time Limit:* collecting large amounts of data that present proportional representations of each property with respect to a sensitive attribute is time consuming and often costly and labor-intensive [33];

iv *Data set validation:* in the design of a machine learning model, more attention is paid to the mathematical basis of the classifier, restricting the data formation process to a black box [78], [76];

v *Validation planning:* data validation, when applied, is often performed only after the model has been trained and used, making the feedback cycle inefficient and often ineffective [87];

vi *Lack of statistical rigorousness:* the suitability of the data set varies depending on the task for which the data are prepared. For instance, models based on linear regression imply assumptions of normality on the measurement error [75], [200]. This specificity is often absent in the pipeline of machine learning models.

Miss-dependency Two-dimensional or bivariate statistics is the study of the degree to which two distinct characters of the same statistical unit are connected. However, the connection only measures the degree of statistical dependency without inducing a cause-effect relationship or dependency between the variables. For instance, it can be shown that people with small feet make more spelling mistakes than people with large feet. However, this statistical dependency does not indicate that having small feet is the cause of spelling mistakes; the greater frequency of spelling mistakes may in fact be due to the younger age of people with small feet. In this case there could be a third variable, age, responsible for the cause-effect relationship. While in a human-centered model - where the human makes the decisions - this distinction is quite evident, in a machine learning model miss-dependency is not always deductible. This depends on two reasons: i) the machine does not recognize the meaning of the instance but looks at the properties of the

variables; ii) the way in which the data are structured modifies the interpretation that the machine is having regarding the relation of statistical dependency. This means that, while in a human-centered model it is the human to verify that the relationships of statistical dependence detected in the available data are leading to a cause-effect relationship, in machine learning models the machine is not always able to recognize a spurious connection, erroneously assigning to two or more variables a cause relationship. In other words, the structure of the available data is responsible for the successful or failed relationships established with the protected attributes (ethnicity, gender, etc.) in the data. In addition, the rapid growth and spread of current machine learning systems is due in part to the ease of design of the models themselves, which thanks to modern software allows the construction of predictive models avoiding the understanding and adoption of rigorous statistical analysis. The simplicity of design has therefore created a gap between predictive and analytical-explicative power, favoring misinterpretation between causality and statistical dependence. The distinction between statistical dependence and causal dependence in data is therefore a primary issue in machine learning models, especially to determine the causes of failure, potential biases encoded in the data and the reliability of application.

4.2 Research Questions

Based on the problems highlighted, our contribution aims to answer the following research questions:

RQ1 Is it possible to establish the probability of composition of the training data from the available data set?

RQ2 Do the available data known to the machine learning community present a discriminatory future risk based on their structure?

4.3 Background

When machine learning model decisions are based on historical records, they tend to embed distortions that exist in reality and crystallize them. Prejudices and human bias therefore become part of the technology itself. This is particularly evident with regard to ethnic discrimination. Over the last years, the rise of machine learning models in various sectors is leading to a dramatic increase of discriminatory outcomes for ethnic minorities, across different fields of application. A striking and well known case is the COMPAS software, used in U.S. court to estimate the probability of defendants' recidivism, which has been shown to underestimate the risk of recidivism for white defendants and overestimate it for black

defendants [84]. However, the COMPAS case is not an isolated phenomenon. In a 2017 experiment conducted on the Airbnb platform, applications from guests with typically African American names were found to be 16% less likely to be accepted than identical guests with typically white names [57]. Also in 2017, a geo-statistical analysis revealed that the design of the popular Pokémon GO game strengthens existing geographical prejudices, for example by benefiting urban areas and neighborhoods with smaller minority populations, economically disadvantaging ethnic minority areas [42]. Several studies have demonstrated the discriminatory potential of targeting advertising [183], [182], which is only recently receiving interventions to remove the prejudicial content of the model. For example, Facebook after years of scandals related to ads that exclude people based on race [8] has finally removed the racial targeting option for ads [109]. In a 2019 study, the commercial algorithm widely used in the U.S. health care system to guide health care decisions was found to discriminate against black patients [143]. The algorithm falsely assigned a healthier condition to black patients despite the risk of complications being the same for white patients, making black people less likely to receive more financial resources for extra care. Although facial recognition technologies are now used in several domains, they still present many discriminatory issues related to differences in margins of error - generally software has a 20% higher margin of recognition error for black women [152] -. As an example, we report what happened recently with Google Vision AI, a computer vision service for image labeling [104]. By providing the system with two images of people holding a body temperature thermometer, it labeled the image containing the white person as an "electronic device", while in the image containing the black person the device held was labeled as a "gun". In a later experiment it was shown that it was sufficient to apply a pink mask on the black person's hand in order the software labeled the image as "tool". Racial bias encoded in machine learning systems is likely to spread silently and like wild fire in everyday technologies. The increasing and ubiquitous spread of such models also intended to make allocative decisions about people's lives makes the problem of prejudice and rational discrimination more urgent than ever. For this reason and for the historical moment we are experiencing, our work intends to focus on rational discrimination in data.

4.4 Motivating Example

Given a population composed of 60% Caucasians, 35% black people and 15% Asian people, the probability of positive outcome for the respective ethnic groups is 70% for Caucasians, 20% for Blacks and 60% for Asians. What is the probability of failure with respect to the protected attribute Ethnicity?

In this example the probabilities are given rather than the numerosity in order to simplify the following notation. To offer a better a better understanding of the

Methodology this data will be used in Section 4.5. The data gives the probability of success, but the similar reasoning is also valid for cases where the probability of failure is known. The intent is to verify whether the probabilities of success or failure of a subgroup are influenced by group membership - and vice versa - and more specifically how these probabilities affect the composition of the training set.

4.5 Methodology

Our data annotation system is based on four modules:

I Dependence: assesses the degree of connection among the protected attribute - in our study, the ethnicity - and the target variable;

II Diverseness: provides the training diversification probability in respect to each level of the protected attribute and the target variable;

III Inclusiveness: provides the probability that two properties are simultaneously included in the training set;

IV Training Likelihood: provides the occurrence likelihood of the protected attribute levels given the target variable levels - and vice versa - before the training set is sampled.

4.5.1 Quantifying Dependence

Excluding some specific domains where the dependence of some protected attributes with the response variable is not considered problematic, but rather it is fundamental for the understanding of a certain problem (for example the gender attribute in the medical field in the detection of particular diseases [47]), in the broad field of machine learning systems the dependence between the protected attribute and the response variable has caused severe consequences [142], [140]. The dependence between the protected attribute and the response variable is therefore one of the major causes of discrimination and as such must be rigorously examined. The first step for a correct bias detection within the data is given by the dependency analysis between the different modalities of a protected attribute and the response variable. In statistics, the measurement of the degree of dependence of two qualitative variables is called contingency; contingency measures the degree of connection of two categorical variables. To determine the degree of connection, the marginal frequencies and the combined frequencies of the bivariate table are used. Given two categorical variables x_i and y_i, the dependency or independence is established through the theoretical independence table $f'(x_i, y_j)$ once the table

of the observed real data $f'(x_i, y_j)$ is given. The contingency $C(x_i; y_j)$ is therefore given by the difference between the observed and theoretical frequencies:

$$C(x_i; y_j) = f(x_i, y_j) - f'(x_i, y_j) \tag{4.1}$$

If the table of the observed real data and the theoretical table of independence coincide - that is if for each cell the value is null - then the two variables are independent. Otherwise, it is necessary to measure the degree of connection between the variables. The degree of connection between two categorical variables is commonly measured by the Pearson connection index, obtained as the sum of the relative quadratic contingencies. The index assumes a value of zero in case of independence in distribution and increases as the degree of connection between variables increases:

$$\chi^2 = \sum_{i,j} \frac{C^2(x_i; y_j)}{n_{i,j}} = n\Big(\sum_{i,j} \frac{n_{i,j}^2}{n_{i0} n_{0j}}\Big) \tag{4.2}$$

In order to support Pearson's connection index, the contingency coefficient is adopted with the purpose of reducing the χ^2 in the range [0;1]:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{4.3}$$

However, the effect size of the degree of connection between two categorical variables is not always easy to interpret, where by effect size we mean a quantitative measure of the magnitude of a phenomenon. To offer a better understanding of the relationship of dependency between two variables, several simplified methods of interpretation have been proposed, especially to guide social scientists in the interpretation of statistical test results. In the spirit of simplifying the interpretation of the dependency between the response variable and the protected categories for a data set user, we introduce the concept of the Effect Size Index w (ES w):

$$w = \sqrt{\sum_{i=1} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}, \tag{4.4}$$

where $p_0 i$ and $p_1 i$ are the value of the ith cells. Notice that unlike the contingency coefficient, the ES w is not derived from frequencies but from proportions. The relationship between the Pearson connection index, the contingency coefficient and the ES Index is given by the following formula:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{w^2}{w^2 + 1}} \tag{4.5}$$

Alternatively to the Formula 4.4 it is also possible to calculate the ES w from the contingency coefficient:

$$w = \sqrt{\frac{C^2}{1 - C^2}} \tag{4.6}$$

The size of the ES w between two variables is then evaluated through the use of Table 4.1, which relates the magnitude of the ES with a nominal label. The

Magnitude	**Value**
SMALL	w = 0.1
MEDIUM	w = 0.3
LARGE	w = 0.5

Table 4.1: Conventional definitions of Effect Size Index w magnitude

advantage of using the conventional conversion table for the user of the data set is that the magnitude of the dependency is displayed quickly and immediately without the need for more complex statistical tests.

4.5.2 Estimating Diverseness

Intuitively, the probability of an event represents how likely the event will occur. According to the classical definition the probability is given by the following ratio:

$$P = \frac{number\ of\ favorable\ cases}{number\ of\ possible\ cases} \tag{4.7}$$

We now apply this elementary theory to the problem of data collection in machine learning. When the data set is partitioned into training and test sets, a split with a more or less standard ratio (70/30 or 80/20) is generally performed, i.e. a sampling is performed on the available data. Let's consider the case in which the training data set is generated by random sampling on the original data set without considering further techniques (stratification or re-sampling) - for example in the case of a non expert user -. The probability an event occurs turns into the probability that the training set shows some existing properties contained in the original data set:

$$P = \frac{number\ of\ favorable\ properties}{number\ of\ possible\ properties} \tag{4.8}$$

In our data annotation this ratio is introduced to allow the dataset user to answer questions like: *"If I perform a random sampling on the original dataset, what is the probability that the training set is mainly composed of positive examples? What is the probability of belonging to a certain group with respect to the target variable?"*

Prior Probabilities The a priori probability of a data property is the degree of belief of the property in the absence of other information, also known as the unconditional probability. The degree of belief is the probability of a property to be true in an uncertain environment. The probability is referred to the belief and

not to the truth of the fact, as it is not possible for the user to know exactly the truth, that is if the original data are representative of the real world. Since the user does not have access to the complete information, several hypotheses on how the real data is structured have to be drawn, assigning to each of them a probability of being true. Formally:

$$\begin{aligned} P &= (Y = y) \\ P &= (A = a) \end{aligned} \tag{4.9}$$

We estimate the prior probabilities by using the data of the problem introduced in Section 4.4, where the target variable Y assumes value 1 in case of negative outcome, otherwise 0. In this specific case, the prior probabilities indicate that

Formula	**Probability**
$P(Y = 0)$	P = 0.48
$P(Y = 1)$	P = 0.52
$P(A = white)$	P = 0.6
$P(A = black)$	P = 0.35
$P(A = Asian)$	P = 0.15

Table 4.2: Example of prior probabilities

the training set has probability 0.48 to be composed by individuals who display a positive outcome and 0.52 to be composed by individuals who display a negative outcome; finally, the probabilities that it is formed by individuals of white, black and Asian ethnicity are respectively 0.6, 0.35 and 0.15 (Table 4.2).

4.5.3 Estimating Inclusiveness

Posterior Probabilities Given two events A and B, the probability $P(A|B)$ is said posterior probability because it allows to calculate the probability of A, knowing that B occurred. In our case the posterior probability means to compute the probability that $Y = y$, knowing that $A = a$ has occurred (and vice versa). In other words, the probability that the training set shows the property Y = y, knowing the property $A = a$ has occurred (and vice versa). We start by estimating the probability that two events occur simultaneously. From the definition of conditional probability:

$$\begin{aligned} P(A = a \cap Y = y) &= P(A = a)P(Y = y|A = a) \\ P(Y = y \cap A = a) &= P(Y = y)P(A = a|Y = y) \end{aligned} \tag{4.10}$$

Since from Compound Probability Theorem [161] $P(A = a \cap Y = y)$ is equal to $P(Y = y \cap A = a)$, i.e. the probability of both properties occurring is the same, either of the two formulas can be employed indistinctly.

Formula	Probability
$P(Y = 0 \cap A = white)$	P = 0.42
$P(Y = 0 \cap A = black)$	P = 0.07
$P(Y = 0 \cap A = Asian)$	P = 0.09
$P(Y = 1 \cap A = white)$	P = 0.18
$P(Y = 1 \cap A = black)$	P = 0.28
$P(Y = 1 \cap A = Asian)$	P = 0.06

Table 4.3: Example of properties occurring simultaneously

4.5.4 Estimating Training Likelihood

From the definition of conditional probability, we derive the Bayes Theorem for the properties of the training set:

$$\begin{aligned} P(A = a|Y = y) &= \frac{P(A = a)P(Y = y|A = a)}{P(Y = y)} \\ P(Y = y|A = a) &= \frac{P(Y = y)P(A = a|Y = y)}{P(A = a)} \end{aligned} \tag{4.11}$$

In the case of binary classification and in the case of protected attributes we are in the presence of a certain event partition. This means that the events are disjointed from each other $Y_i \cap Y_j = \emptyset$ and $A_i \cap A_j = \emptyset$ if $i \neq j$ and that as a whole they are the only ones possible, i. e., if a certain property occurs, one and only one certainly appeared. In other words, it is not possible that the training set is composed of individuals who belong simultaneously to the black and white ethnic group, or who simultaneously show a positive and negative outcome. The union of the occurrence of the single properties is therefore the whole set of possible properties. For the properties outcome and ethnicity the generalization formula are respectively:

$$\begin{aligned} \Omega : \cup_{i=1}^{N} Y_i = \Omega,\ hence \sum_{i=1}^{N} P(Y_i) = P(\cup_{i=1}^{N} Y_i) \\ \Omega : \cup_{i=1}^{N} A_i = \Omega,\ hence \sum_{i=1}^{N} P(A_i) = P(\cup_{i=1}^{N} A_i) \end{aligned} \tag{4.12}$$

By applying Formulas 4.10 and 4.12 the Bayes Theorem can be generalized for each property of the training set:

$$\begin{aligned} P(Y = y|A) &= \frac{P(Y = y)P(A|Y = y)}{P(A)} = \frac{P(Y = y)P(A|Y = y)}{\sum_{i=1}^{N} P(A|Y_i)P(Y_i)} \\ P(A = a|Y) &= \frac{P(A = a)P(Y|A = a)}{P(Y)} = \frac{P(A = a)P(Y|A = a)}{\sum_{i=1}^{N} P(Y|A_i)P(A_i)} \end{aligned} \tag{4.13}$$

The first equation in Formula 4.13 derives the probability of the outcome property given the ethnic property, while the second equation derives the probability of the ethnic property given the outcome property. In other words, it derives the probability of composition of the training set based on the posterior probabilities of the outcome and ethnicity properties. Carried out a random sampling on the original data, the Formula answers the following questions:

i In the sampled training set what is the probability of belonging to an ethnic group with respect to the outcome variable?

ii In the sampled training set what is the probability of obtaining a certain outcome with respect to the ethnic group?

Complementarily, the two equations can be interpreted as the probability of bias within the training set.

Formula	**Probability**
$P(Y = 0\|A = white)$	P = 0.7
$P(Y = 0\|A = black)$	P = 0.2
$P(Y = 0\|A = Asian)$	P = 0.6
$P(Y = 1\|A = white)$	P = 0.3
$P(Y = 1\|A = black)$	P = 0.8
$P(Y = 1\|A = Asian)$	P = 0.4
$P(A = white\|Y = 1)$	P = 0.34
$P(A = white\|Y = 0)$	P = 0.87
$P(A = black\|Y = 1)$	P = 0.53
$P(A = black\|Y = 0)$	P = 0.15
$P(A = Asian\|Y = 1)$	P = 0.11
$P(A = Asian\|Y = 0)$	P = 0.18

Table 4.4: Example of posterior probabilities

4.6 Case Studies Datasets

COMPAS (Correctional Offender Management Profling for Alternative Sanctions)[1] is a popular tool used by U.S. court to estimate the defendants' probability of recidivism. This dataset displays the probability of reoffending based on two year of further studies. The dataset has been shown to underestimate the risk of

[1]Retrieved from:
https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

recidivism for white defendants and overestimate it for black defendants [84].
Drug Consumption [63], [64] contains information on the consumption of 18 drugs based on personality traits and socio-economic attribute. For simplicity of analysis we assumed the consumption of Cannabis as target variable but the annotation of the dataset can be made on each target drug.
Adult Dataset [107] The data set contains adult income annual census from the US Census Bureau. It is commonly employed in forecasting tasks in order to predict the factors leading to income below or above $50,000.

Property	**COMPAS**	**Drug Consumption**	**Adult Dataset**
Size	6172x9	1885x31	48842x15
Target variable	$0 \rightarrow$ no $1 \rightarrow$ yes	$0 \rightarrow$ non user $1 \rightarrow$ user	$0 \rightarrow > 50K$ $1 \rightarrow \leq 50K$
Levels of ethnicity attribute	Asian Black Caucasian Hispanic NA[c] Other	Asian Black Black/Asian Caucasian White/Asian White/Black Other	AIE [a] API [b] Black Caucasian Other

Table 4.5: Summary of Datasets Prominent Properties
[a] American-Indian/Eskimo, [b] Asian-Pac-Islander, [c] Native American

4.7 Results

We performed the analyses that constitute our data annotation system for each of the datasets presented in Section 4.6. Sub-sections 4.7.1, 4.7.2, 4.7.3 and 4.7.4 report the analysis for each module - dependency, diverseness, inclusiveness, training likelihood, respectively - and contain an example graphic module. Figure 4.4 shows an illustrative example of the graphical visualization for the complete notation.

4.7.1 Dependence

This module aims to analyze the connection relationships between the protected attribute Ethnicity and the target variable that are established and depend on the available data. For instance, for the COMPAS dataset the module highlights the dependency relationships between recidivism and different ethnic minorities. Summary results for dependence module are shown in Table 4.6.

	COMPAS	Drug Consumption	Adult Dataset
Contingency coefficient	0.1413	0.1558	0.0994
Effect size w variable	0.1427	0.1578	0.0999
Magnitude of Effect size w	SMALL	SMALL	VERY SMALL

Table 4.6: Summary of Dependence Prominent Properties

None of the three datasets displays worrying dependency values among the protected attribute Ethnicity and the target variable, showing the magnitude of the Effect Size w as small or very small. However, the results of the COMPAS dataset - which is proven to contain bias - indicate that this module alone is not sufficient to show a latent bias risk. The degree of bias depends on the sample size and the value of the contingency coefficient of the target variable and the protected attribute [216]. Smaller samples lead to more bias and higher variance [217] and therefore the results of the dependency must be analyzed in relation to the amount of data available. In order to facilitate the interpretation of the connection relations, we propose a graphic notation for dependence. Figure 4.1 shows the graphical representations of the dependency modules based on different connection magnitude.

4.7.2 Diverseness

This module aims to analyze the diverseness of the data available by estimating prior probabilities. They determine the probability that training set will display an a priori environment based on the original data available, i.e. they show the probability of training set composition stratified by each of target variable and protected attribute levels. For example, in our case study the module highlights the probability that training set will be equally composed by ethnic minorities and ethnic majorities. Summary results for diverseness module are shown in Table 4.7. In terms of target variable probabilities, the results show strong distortions for the Drug Consumption and Adult datasets with a high probability of positive examples - i.e. showing a negative outcome - while the probabilities of the COMPAS dataset are quite homogeneous. Regarding the probabilities of the protected attribute ethnicity, the distortions are even more pronounced than the target variable ones, revealing a very high probability of composition for the Caucasian ethnicity in the Drug Consumption and Adult datasets. In the case of the COMPAS dataset the probabilities are indeed distorted, although still not such as to predict at this point

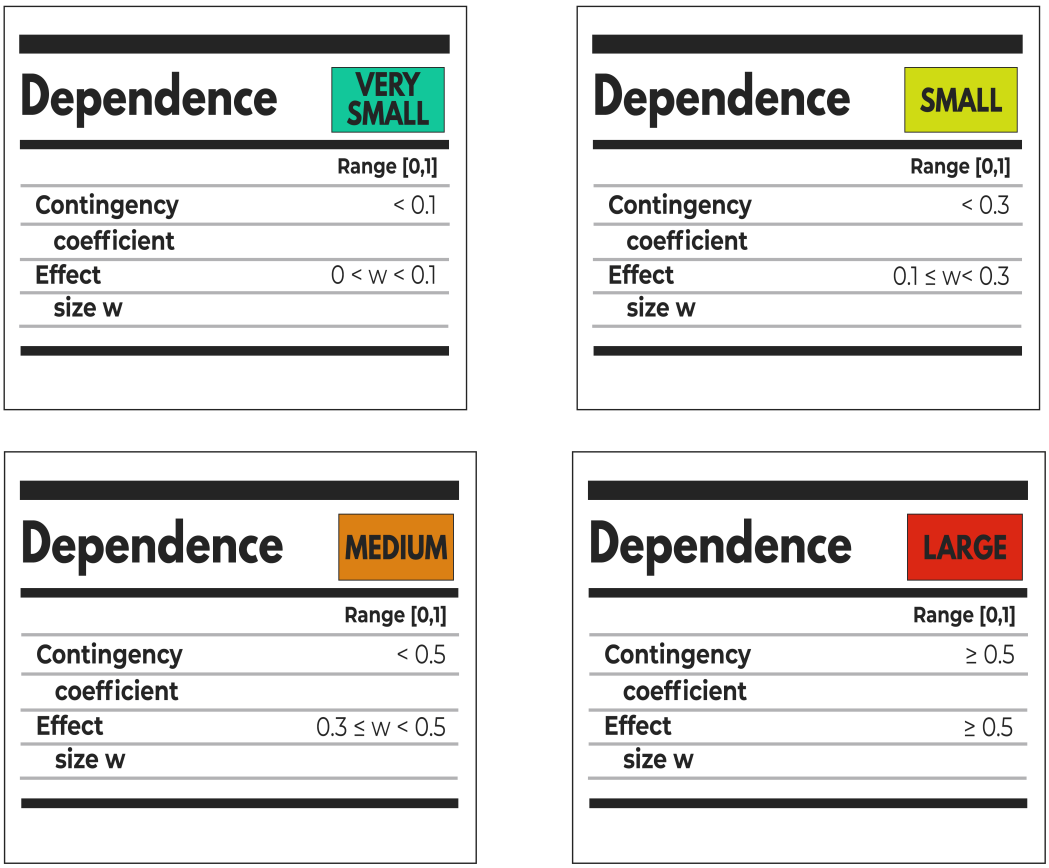


Figure 4.1: Example of Dependence graphic visualization

of the analysis more severe future distortions, which is why more in-depth analysis are required. Figure 4.5.2 shows the graphical representation of the diverseness module that simplifies the display of prior probabilities. In the example is given the notation for a dataset where both the levels of the target variable and those of the protected attribute ethnicity are equiprobable.

4.7.3 Inclusiveness

This module aims to analyze the inclusiveness of the data available by estimating the simultaneously probabilities. They determine the probability that training set will simultaneously display two by two the target variable and the protected attribute properties. For instance, in our case study the module highlights the probability that in training set the property Asian appears simultaneously with property success. Summary results for diverseness module are shown in Table 4.8. The results of this module show that the probability that two properties will occur

	COMPAS	Drug Consumption	Adult Dataset
0	0.545	0.329	0.239
1	0.455	0.671	0.761
Caucasian	0.341	0.912	0.855
Black	0.514	0.018	0.096
Asian	0.005	0.014	
Hispanic	0.082		
Native American	0.002		
Other	0.056	0.033	0.008
White/Black		0.011	
White/Asian		0.011	
Black/Asian		0.002	
Amer-Indian-Eskimo			0.010
Asian-Pac-Islander			0.031

Table 4.7: Summary of Diverseness Analysis Results

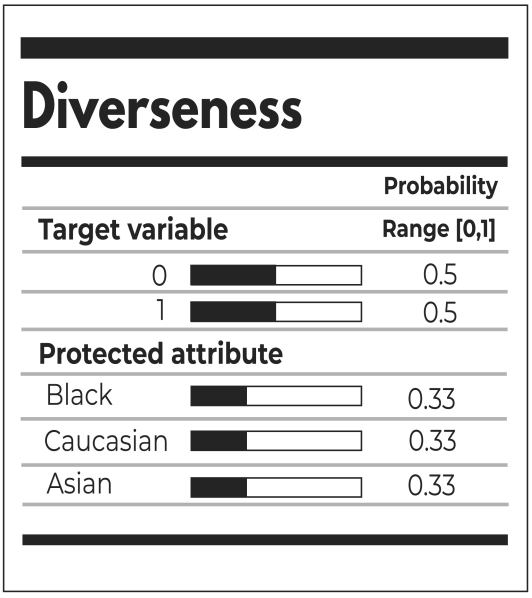


Figure 4.2: Example of Diverseness graphic visualization

simultaneously is related to the sample size. Evidence of this can be found in the results of the Drug Consumption and Adult datasets, where the highest probabilities of simultaneous events involve the Caucasian property. The COMPAS dataset

	COMPAS	**Drug Consumption**	**Adult Dataset**
0∩AIE[a]			0.0006
0∩Asian	0.0023	0.0019	
0∩API[b]			0.0041
0∩Black	0.1514	0.0023	0.0057
0∩Black/Asian		0.0000	
0∩Caucasian	0.1281	0.0555	0.1061
0∩Hispanic	0.0320		
0∩NA[c]	0.0006		
0∩Other	0.0219	0.0013	0.0005
0∩White/Asian		0.0004	
0∩White/Black		0.0006	
1∩AIE			0.0042
1∩Asian	0.0008	0.0007	
1∩API			0.0111
1∩Black	0.1661	0.0010	0.0412
1∩Black/Asian		0.0003	
1∩Caucasian	0.0822	0.1165	0.3115
1∩Hispanic	0.0189		
1∩NA	0.0005		
1∩Other	0.0124	0.0050	0.0036
1∩White/Asian		0.0016	
1∩White/Black		0.0014	

Table 4.8: Summary of Inclusiveness Analysis Results
[a] American-Indian/Eskimo, [b] Asian-Pac-Islander, [c] Native American

shows quite homogeneous probabilities especially with regard to the Black property, while for the Caucasian property the highest probabilities are related to the simultaneous occurrence with the Non-recidivist property. Since the simultaneous probabilities depend on the number of examples within the available data and the sample size, this result alone is not sufficient to establish a priori the certain presence of serious data distortions, although some evidence can already be seen. Figure 4.3 shows the graphical representation of the inclusiveness module that simplifies the display of simultaneously probabilities. In the example is given the notation for a dataset where all the properties of the target variable and those of the protected attribute ethnicity are equiprobable.

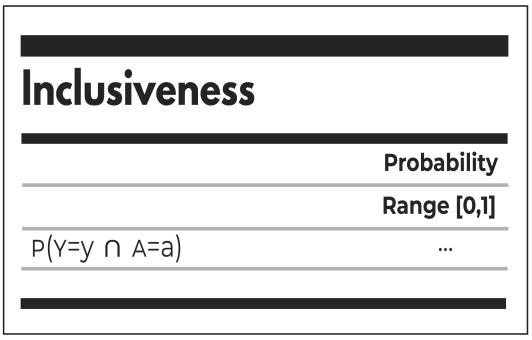


Figure 4.3: Example of Inclusiveness graphic visualization

4.7.4 Training Likelihood

This module aims to analyze the training likelihood of the data available by estimating the posterior probabilities. They determine the probability that in the training set the occurrence of the properties of the protected attribute is given by the properties of the target variable - and vice versa - . For example, in the COMPAS dataset they determine the probability that the occurrence of reoffending is given by the properties of the protected attribute ethnicity. Summary results for training likelihood module are shown in Table 4.9.

The results of this module show that the posterior probabilities of target variable and protected attribute ethnicity are quite skewed in all dataset. In the case of the Adult dataset given as occurred event 1 or event 0, the probability of occurrence of the Caucasian ethnic group is respectively 0.908 and 0.839, - i.e. very high for both events - while the probabilities of all other ethnic groups conditioned to the target variable are all significantly lower; this means that the original data contain many examples of individuals belonging to the Caucasian ethnic group. In the case of Drug Consumption, a similar reasoning can be carried out for the ethnicity probabilities conditioned to the target variable; moreover, notice that given the property Black/Asian, the probability of occurrence of event 1, i. e. that the individual is a consumer, is 1 - while the probability of 0 is 0 - which means that in the available data there are no examples of individuals belonging to the ethnic group Black/Asian showing a positive outcome - i. e. negative examples -. Figure 4.4 shows the graphical visualization of our data annotation system for the COMPAS dataset. The analysis of the COMPAS dataset shows that if an individual is randomly sampled from the original data for the training set, the probability that this individual is black knowing that the re-offending property has occurred - i.e. knowing the outcome of the re-offending event - is 0.591, while the probability that the individual is white knowing that the re-offending property has occurred is 0.293. Instead, given as occurred the property Black the probability that the

	COMPAS	**Drug Consumption**	**Adult Dataset**
0|AIE[a]			0.117
0|Asian	0.742	0.731	
0|API[b]			0.269
0|Black	0.477	0.697	0.121
0|Black/Asian		0.000	
0|Caucasian	0.609	0.323	0.254
0|Hispanic	0.629		
0|NA[c]	0.545		
0|Other	0.638	0.206	0.123
0|White/Asian		0.200	
0|White/Black		0.300	
1|AIE			0.883
1|Asian	0.258	0.269	
1|API			0.731
1|Black	0.523	0.303	0.879
1|Black/Asian		1.000	
1|Caucasian	0.391	0.677	0.746
1|Hispanic	0.371		
1|NA	0.455		
1|Other	0.362	0.794	0.877
1|White/Asian		0.800	
1|White/Black		0.700	
AIE|0			0.005
AIE|1			0.011
Asian|0	0.007	0.031	
Asian|1	0.003	0.006	
API|0			0.035
API|1			0.030
Black|0	0.450	0.037	0.048
Black|1	0.591	0.008	0.111
Black/Asian|0		0.000	
Black/Asian|1		0.002	
Caucasian|0	0.381	0.895	0.908
Caucasian|1	0.293	0.921	0.839
Hispanic|0	0.095		
Hispanic|1	0.067		
NA|0	0.002		
NA|1	0.002		
Other|0	0.065	0.021	0.004
Other|1	0.044	0.040	0.010
White/Asian|0		0.006	
White/Asian|1		0.013	
White/Black|0		0.010	
White/Black|1		0.011	

Table 4.9: Summary of Training Likelihood Analysis Results
[a] American-Indian/Eskimo, [b] Asian-Pac-Islander, [c] Native American

individual has not reoffended is 0.477, while the probability that the individual has reoffended is 0.523; given the property Caucasian, the probability that the individual has not reoffended is 0.609, while the probability that the individual has reoffended is 0.391, that is significantly lower. This means that in this dataset the reoffending is related to ethnicity, and that success or failure are determined by the

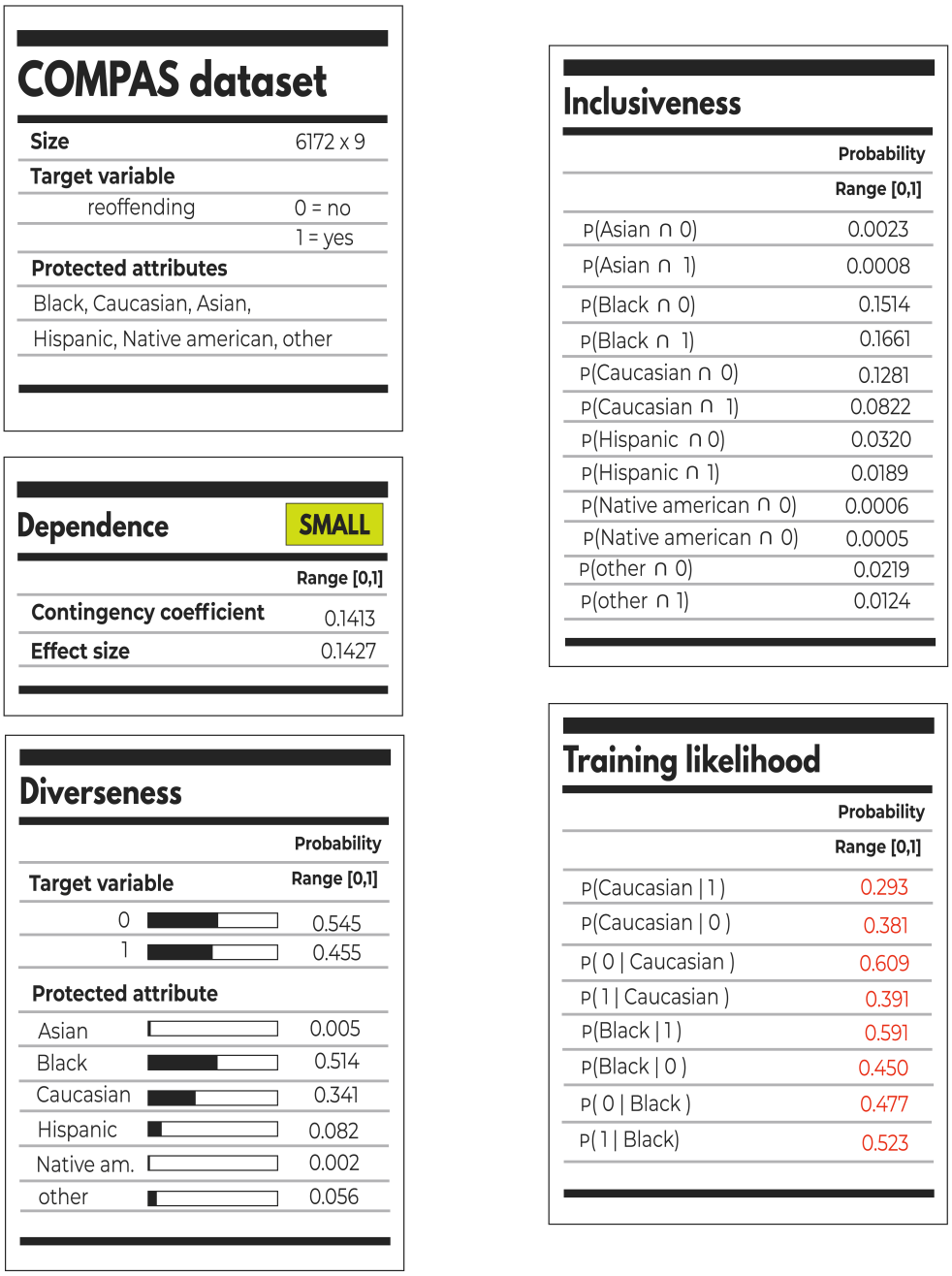


Figure 4.4: Data annotation visualization for COMPAS dataset

membership to a specific ethnic group. The differences in probability between the properties highlight the risk of future bias, and in the case of the COMPAS dataset they anticipate the underestimation of recidivism for the Caucasian ethnic group and the overestimation of recidivism for the Black ethnic group proven in recent

studies [84].

4.7.5 Final Remarks

RQ1: in traditional sampling practices, instead of observing all the units of a population, only a subset of a population is detected, which must show certain probabilistic characteristics. In machine learning models the training set is sampled not from the real population but from the available data. While in classical sampling the empirical knowledge alone is effectively of a sample nature, in machine learning systems the available data are often of sample nature too, precisely due to the fact that it is not possible to make assumptions on the real population. Considering a random sampling from the available data, we have shown that the probability of composition of the training set can be predicted, highlighting that the structure of the data directly affects the probability of properties distribution;

RQ2: we analyzed three datasets frequently accessed by machine learning community. Of these, all three showed more or less pronounced distortions for the protected attribute Ethnicity. Although the COMPAS dataset is the sole one that has been shown to discriminate against black people, the Drug Compsuntion and Adult datasets reveal possible future bias in the detriment of ethnic minorities.

4.8 Relations to Related Work and Limitations

4.8.1 Data Labeling

Although there are a number of papers that for ethical purposes deal with data annotation they are all very recent, indicating that this field of study is still partially explored and has only recently received considerable attention. Our contribution differs from the others because it induces a probabilistic reasoning on the causes of model discrimination based on sampling problems; our intention is to deepen the knowledge of data validation analysis, focusing on the meaning of probabilities. From a graphical point of view, our work has been inspired by the Data Nutrition Labels [87], a data labeling system mainly based on descriptive data statistics. A similar approach is addressed in [19], where an operational framework is proposed to identify the bias risks of automatic decision systems. In [75] the authors propose a data labeling system based on discursive data sheets. In [37] the authors propose a collaborative crowdsourcing system to improve the quality of the labels.

Since ethically data annotation represent a quite new field of study, there are several works that provide different types of labels. We believe that at present the focus should not be on achieving a unified data annotation system in the short term,

but rather on the fact that the fair machine learning community is working together to focus attention on the data collection problem. Especially because awareness of data issues is often not rooted outside of this community. It is important that this field and this work inspire greater awareness of the possible causes of discrimination due to the fundamental ingredient that all users and designers of machine learning systems (from the most to the least experienced) use, data.

4.8.2 Data Bias and Conditional Probabilities

As Friedler *et al.* [70] observe, the goal of fairness algorithms is to ensure that the mapping between the *feature space* - the input - and the *decision space* - the output - is a transparent process and that the points defined between these spaces are properly measured. To better explain this concept, we will make use of the definitions of spaces provided by Friedler:

1. **Construct Space**. The *construct Space* is a metric space consisting of individuals and a distance between them. It is assumed that the distance correctly captures closeness with respect to the task [70]. This space is not always directly observable, since the desired variable is not always measurable. To overcome this problem, proxy variables belonging to the *Observed Space* are often employed;

2. **Observed Space**. The *Observed Space* is a metric space consisting of the set of features directly measured or observed in the data;

3. **Decision Space**. The *Decision Space* is a metric space consisting of the process of finding a map from the *Observed Space* to the space of outcomes.

Nowadays, automatic decision-making systems generally use a type of data in which the constructed space is typically unobservable. On the other hand, traditional statistical models make use of surveys and sampling that, starting from the *Observed Space*, allow to infer the *Decision Space* - or prediction space - and the margin of error derived from the observation of the actual population and the observed population. Since the data used today rarely make use of these tools, the biggest gap in predictive systems is characterized precisely by the inability to quantify the inference error. To provide a better perspective of this problem, consider the following cases:

1. **Case 1: the actual properties differ across groups**. In this case, the distributions of the results correspond with the distribution of the properties of interest. For example, in college admissions tests, the decision space might indicate that students under 20 are more likely to complete their studies. Thus, Case 1 occurs when this prediction coincides with the distribution of these properties in the real population;

2. **Case 2: the actual properties are different from those observed**. In this case the distribution of the results does not correspond with the distribution of the properties of interest. Following the example above, students under 20 are found to be more likely to complete their studies but this result does not match with the actual distribution of the property in the real population.

The case 2 is the most common case today and is the one that gives rise to the bias results. This is because the *Observed Space* is not representative of the real population and/or does not coincide with the *Construct Space*; thus the *Decision Space* is constructed through erroneous mapping.

Our method estimates the composition probability of the training set from the *Observed Space*. In this context, our approach does not directly solve the problem of *Construct Space* and *Observed Space*, but addresses the problem by considering that since the two cases are indistinguishable from each other, if Case 2 - the worst case - were true, then the model would incorrectly infer the *Construct Space* by providing an incorrect mapping of the *Observed Space* into the *Decision Space*. In this sense, the Bayesian inference is used as a worst-case inference prediction. In the future, our method could be improved to provide a margin of error for inference based on a posteriori probabilities. This would mean that, from the *Observed Space*, an a posteriori error probability is derived based on the a priori probability that the *Observed Space* and the *Construct Space* coincide. The margin could be constructed from these two cases and be expressed in continuous form: in the first case, the spaces are coincident - the limiting case in which the entire population is considered by the model -, in the second case, the two spaces are completely disjoint. Starting from the a posteriori probability of the features space derived from the available dataset, the margin of error would then indicate the maximum margin of error - and thus the future bias of the *Decision Space* - if the distance between the *Observed Space* and the *Construct Space* were maximum.

Since the actual population is not known, the distance between the two spaces is not directly observable. A solution to overcome this problem might be to perform tests to produce a set of possible populations, in other words, to provide the a priori probability of the *Construct Space*. From this perspective, our approach does not address the removal and mitigation of bias within the data, since it stands within the literature as a data annotation and labeling system. As part of work improvement, data resampling techniques - such as oversampling, undersampling, smooting processes, and non-probabilistic methods - could be used to define margins of error on the prediction space, particularly on protected attributes.

4.9 Discussion and Future Work

The purpose of the current study was to detect the potential race discriminatory risk for future machine learning system by providing a data annotation system based on Bayesian Inference. Our notation serves as a diagnostic framework to immediately visualize data appropriateness and potential bias occurring when sampling the training set from an available dataset. The investigation of the probabilities of the training set sampling has shown that it is possible to establish a risk of future bias by observing prior and posterior probabilities of the ethnicity and target variable properties. The empirical findings in this study provide a new perspective on data annotation practices by showing that Bayesian inferences may reveal the risk of bias in three different widespread dataset. Furthermore, this study has raised important questions about the awareness of most widely data sampling practices in machine learning community. The findings of this investigation complement those of earlier studies. Our data annotation system is limited to the binary case and to the analysis of categorical variables for classification tasks. This would be a fruitful area for further work. Our intent is to expand the work in the following directions: i) extend the notation to multiple protected attributes - the probabilities of the training set will then be given by the vectors of the protected attribute combinations - ; ii) extend the notation to the non-binary case - for prediction tasks involving regression analysis for example - ; iii) extend the probabilistic notation to non-labeled data.

Chapter 5

Achieving Fairness in Ranking Systems

5.1 Introduction

Over the last decade, we have witnessed a large diffusion of increasingly sophisticated predictive models for decision-making, which exploit an ever-growing amount of personal data for suggesting or directly taking actions [106]. The algorithmic decisions are the result of massive profiling [169] and classification mechanisms [55] [83]. Algorithms are involved in a wide range of cases: deciding whether we are reliable enough to receive a mortgage or a loan [106]; suggesting whether a convicted individual is inclined to re-offend [20]; identifying the best candidates for a job [181] or to attend a particular university [100]; recommending what we should buy next [145], what music we should listen to [168] or what movie we should watch [92].

There are advantages that can hardly be denied when using algorithmic decision-making systems: compared to people, algorithms don't get tired, sick or bored; they can perform tasks in a shorter time than human beings; they can work with a much larger amount of information than people can deal with. However, several works show that, like humans, algorithms are susceptible to biases [18] [14]. In this regard, researchers, practitioners and commentators have raised a number of issues on the results of automated decision processes, which have been reported to be discriminatory especially for disadvantaged groups [60] [140]. As a consequence, research communities are devoting relevant effort to study how to include the notion of fairness in automated-decision-making systems (in particular those based on machine learning) in order to produce more equitable results and to avoid discrimination [5]. Current ongoing researches formalize the concept of fairness with different approaches, and a large and diverse range of solutions has been proposed in different contexts [14].

This type of research on algorithmic fairness is part of a wider debate on how

software systems implicitly propagates certain political, economic or cultural visions [119], [19]. With this perspective in mind, we can conceptually organize the space of current researches in two strands: i) investigating instrumental value of software systems and wondering what types of discrimination they could create or worsen [199], and ii) exploring human involvement in shaping the systems, aiming at modeling neutrality or trying to codify affirmative policies to reduce social inequalities [171]. Although several efforts have been made to achieve fairness in both areas, there are still important gaps. The first one concerns ranking systems, that are the base of most of today's automatic decision-making systems, for example in asset pricing, housing, health care, university admission, job recruitment, just to mention the most recurring applications. Similarly to other types of systems, also ranking systems have been found to reflect some of the biases of our society [102] [39]. Hence, although ranking systems are widely used and suffer of the same issues of discrimination observed in other types of automated decision-making, fairness in ranking systems is much less explored than in other fields, as for example supervised machine learning[1]. The second gap that we identified in the literature of fairness in automated decision-making systems regards the fact that most of the studies focused on providing a definition of equity, rather than giving a solution to inequality. Many approaches actually provide mechanisms for avoiding a disadvantage to individuals or minority groups, but they do not provide compensatory mechanisms for those groups that have suffered an unfair outcome. One of the main causes is due to the fact that most of the proposed fairness solutions are the result of a restorative process which aims at removing biases in models that have caused discriminatory outcomes, instead of compensating for the real causes of biases exacerbated by the automatic process. However, several theories of justice and economic theories (see Section 5.4) establish that the disadvantageous circumstances of an individual or group of individuals should lead to redistribution or compensation by society: in practice, minimal research is done to integrate distributive justice and equality of opportunity theories in automatic decision-making systems.

Our research contributes to tackling these two issues identified in the literature: we focus on fairness in ranking systems by turning into practice the principles of distributive justice. In particular, we refer to Roemer's Equality of Opportunity as the basis for defining fairness and inequality and we design a ranking system based on the notions of distributive fairness. The purpose of such a ranking procedure is to re-allocate resources: for this reason other criteria of justice -such as procedural, interpersonal and informational justice- are not taken into consideration.

In this manuscript we introduce AFteRS, an Automated Fair-Distributive Ranking System that implements three fairness criteria, each one based on a different dimension of the distributive justice theory, namely *equity*, *equality*, and *need* [43].

[1]For an overview of fairness in supervised machine learning, see [14]

Each fairness criterion provides diverse ranking results as well as different effects on individuals and groups of individuals. We test the system in an hypothetical scenario of a university selection process in which the decision-maker determines which students are suitable on the basis of their personal qualifications and achievements, so as to maximize the institution utility. In such a context, we examine the expected outcome for groups of individuals in the ranking system before and after the application of our distributive fairness approach, and we explore the trade-off between the three different fairness policies in relation to the obtained rankings. Results of our research doesn't show an absolute predominance of one fairness criterion over another one, and that it is possible to achieve fairness constraints with a minimal impact on the general utility of the system.

The manuscript is structured as follows: Section 5.3 provides an overview of researches on bias and fairness in supervised machine learning, and analyzes limitations and open challenges of the current solutions. In Section 5.4, we provide a comprehensive background on distributive justice theories (Section 5.4.1) and, in particular, on the Roemer's Equality of Opportunity approach (Section 5.4.2). Section 5.5 introduces our Automated Fair-Distributive Ranking System (AFteRS), while Section 5.6 introduces a set of new fairness metrics (i.e.,, ranking, inequality, and distributive justice metrics) to evaluate our approach in a university selection process (Section 5.6.1) as well as the results of our evaluation (Section 5.6.2). In Section 5.6.2 we describe a Fair Ranking Policy Simulation scenario that has the goal of supporting a human decision maker in selecting the criteria of distributive justice to apply to obtain a fair ranking. Lastly, Section 5.7 provide the discussion and limitations about our study.

5.2 Research Questions

A vast majority of previous works in ranking systems has established fairness constraints by exploiting statistical parity or by assessing equity of exposure according to group membership or merit. One of the aims of our work is to show that these practices don't necessarily lead to fair outcomes. In this study we face the problem of assessing fairness in rankings by proposing an Automated Fair-Distributive Ranking System based on distributive justice and Roemer's EOp theory (Section 5.5.1). We provide details of the model design (Section 5.5.2) that takes inspiration from Hothorn 2006 and Brunori et al. 2020 in implementing specific parts of Roemer's EOp theory (Sections 5.4.2 and 5.4.2). Finally, we introduce three different policies for rankings (5.5.2).

The present study is based on the three main research questions shown in Table 5.1.

	Research Question	Sub-group
RQ1	Are decision-making systems based on the theories of distributive justice and Equality of Opportunity able to provide fairness acceptable results while preserving the utility and the accuracy of the decision?	Fairness in automated decision-making
RQ2	Are ranking systems based on a distributive fairness constraint able to preserve the accuracy of the ranking and the model's overall utility by providing a ranking of the best candidates?	Fairness in ranking systems
RQ3	What are the factors affecting the fairness utility trade-off in a fairness constrained ranking system?	Model evaluation

Table 5.1: Research questions overview

5.3 Related Work

Recently, a growing attention has been paid to model fairness in automated decision-making systems, in response of a increasing number of studies and journalistic investigations showing that those systems can reverberate the same bias and discrimination occurring in society. However, limited efforts have been made to investigate the theoretical and moral assumptions underlying the proposed fairness techniques and definitions. Concerning distributive justice and equality of opportunity theories, scientific communities are still far from making them operational and they mostly focus on supervised machine learning algorithms. For instance, Hardt's definition of equality of opportunity [83] considers the case of binary classification in machine learning and it states that fairness is satisfied if both groups - minority and majority - have the same probability of being correctly classified in a positive way. Although such implementation is correct per se, it lacks all the theoretical foundation of equality of opportunity theory, where the socio-economic substratum of individuals is considered in a broader perspective. The attempt to make these theories operational has led to a hyper simplification, misinterpreting the concept of opportunity which in distributive theories refers to an intrinsic condition of the individual - for example, the birth context. In the same spirit, in the ranking systems domain, distributive and equal opportunities theories are simplified through the binary codification of some properties, such as relevance, defining as opportunity the probability that groups of users or items are relevant to a certain query. For instance, Singh *et al.*[178] define equality of opportunity in ranking systems as the probability of being seen, i.e., the exposure, which must be similar for all groups of users or items. Our work substantially differs from the implementations

proposed so far, providing a more comprehensive encoding of distributive and equal opportunity theories in ranking systems. Since these theories have not been implemented in a substantial and complete way by recent work or substantial strands of computer science research, the remaining literature section offers a general overview of fairness studies in recommendation and ranking systems.

5.3.1 Fairness in Recommendation and Ranking Systems

In the last decade, a growing number of recommendation and ranking systems is being used in a wide range of areas with an invasive impact on people's lives. As a consequence, several studies on fairness in recommendation and ranking systems have been emerged. In this Section, we provide a review of these approaches, highlighting their limitations and the open challenges in integrating fairness criteria in recommendation and ranking systems.

Recommendation Systems. Yao *et al.* [208] have identified two main types of biases: (i) observation bias and (ii) bias arising from unbalanced data. The first type occurs when -due to the feedback mechanism- the model is reinforced with a given classification and the user receives recommendations that are always very similar to the previous ones. Therefore, if a user is never exposed to an element, she will never be able to give an opinion on it in order to re-calibrate the model [61]. This type of problem has been addressed in several works that propose to manage the observation bias by increasing the diversification of the proposed elements [86], [193], [132]. However, this type of intervention does not directly address the issue of fairness: Leonhardt *et al.* [115] point out that these studies have largely focused on *individual diversity*, which deals with diversifying recommendations to users, and with *aggregate diversity*, whose objective is to improve the diversity of the items between users. Although these aspects can be considered as part of the problem of fairness in recommendation systems, they do not deal with the discrimination effects of the recommendations on the users.

On the contrary, a strand of recent studies focused more explicitly on fairness, and in particular on the problem of bias, distinguishing user-related biases from item-related biases: as a matter of fact, Farnadi *et al.* [61] use multiple user-user and item-item similarity measures to assess fairness. Other approaches involve more dimensions: Burke [31] defines the problem of fairness in recommendation systems as a multi-sided problem, where the primary need is to recognize the necessity of differentiation between subjects (i.e.,, users) and objects (i.e.,, items). This work [31] also highlights that recommendations are often made into multidimensional platforms, causing problems of multidimensional equity, and distinguishes three classes of systems which differ in terms of equality issues that arise in relation to different groups, namely consumers (C-fairness), providers (P-fairness) and both (CP-fairness). Recently, Edizel et al. [58] propose a post-processing method for

mitigating bias by predicting it in recommendations connected to a pool of sensitive attributes.

Further expanding the scope or aspects of fairness, we also mention Yao *et al.* [208] that define four types of metrics to measure unfairness and improve equity among users in collaborative filtering, dealing with prediction errors between protected and unprotected groups and using the matrix factorization method. A similar approach is recently adopted by Burke *et al.* [32], where the opinions of different users are aggregated with the aim of ensuring a fair representation of the protected group. Along the same line, Ning *et al.* [139] address the problem by using sparse linear methods, while Xiao et al. [205] define the problem as a multiple objective optimization to maximize user satisfaction among groups. Finally, Tsintzou *et al.* [195] study the long-term effects of recommendations to achieve fairness for items.

Ranking Systems. A variety of studies incorporates the notion of statistical parity [55], establishing that the demographic set of individuals' attributes, with any outcome distribution, has to be the same as the demographic set of the population as a whole [213], [206], [178], [35], [12]. In particular, Yang *et al.* [206] have treated the statistical parity measure in ranking by comparing the outcome distributions of minority and majority groups through averaging the differences in the top-N-ranking, similarly as in the Normalized Discounted Cumulative Gain[2]. Although in a preliminary way, they have formulated the issue of fairness as a multi-objective programming problem where fairness is achieved while keeping accuracy acceptable. In contrast, Zehlike *et al.* [213] and Celis *et al.* [35] have adopted statistical parity by the perspective of outcome diversification. Both works have proposed to apply a fairness constraint while maximizing the positional general utility, i.e., positioning more qualified individuals in higher position. Finally, Asudeh *et al.* [12] have proposed a class of fair scoring functions in training phase to produce free-bias rankings.

Another class of studies have faced the fairness issue by the perspective of equalizing exposure [179], [22], [180]. For example, Singh *et al.* [179] introduced an optimal probabilistic ranking to equalize exposure among minority and majority groups, while Biega *et al.* [22] investigated exposure allocation by satisfying an individual fairness constraint and by keeping the general utility. In a more recent work, *et al.* [180] advanced their previous study by developing a Learning-to-Rank algorithm and by optimizing the ranking utility and allocating exposure according merit.

We conclude this overview by observing that:

[2]The Normalized Discounted Cumulative Gain is a wide spread measure of ranking quality in Information Retrieval (see [95] for additional details).

i Unlike the supervised machine learning domain where scholars are trying to unify their approaches 2.3, the domain of fairness in ranking systems is still highly fragmented;

ii Current solutions in machine learning and ranking domains actually focus on removing bias;

iii Distributive fairness solutions are almost completely unexplored.

Drawing upon the previously described strands of research, our work explores, for the first time, the usage of distributive justice mechanisms for ranking tasks in order to provide a more equitable system by compensating individuals' disadvantageous circumstances. Therefore, this study makes a major contribution to research on fairness by assuming the individuals' score as a transferable resource in order to equalize the individuals' opportunities; thus meaning that it can be re-allocated according to a certain policy.

5.4 Background

The current Section provides a theoretical background of distributive justice theories (Section 5.4.1) and describes in detail the Roemer's Equality of Opportunity (EOp) approach (Section 5.4.2). In addition, we also discuss how notions from EOp can be integrated with machine learning tools, as shown by the works of Hothorn *et al.* [88] and Brunori *et al.* [30]. The aim of the following Section is to provide a comprehensive perspective of the theories that constitutes the philosophical ground of our work.

5.4.1 Distributive Justice Theories

Historically, distributive justice has been a widely debated topic first by moral and political philosophers and then by economists. In particular, the way in which goods should be distributed among individuals gave rise to various egalitarian theories. According to traditional social welfare theories [148], egalitarianism means equality of well-being or utility. However, since the work of the moral and political philosopher John Rawls in the 1970s [155], many scholars have argued that this type of equality is not ethically desirable because individuals are not held responsible for their choices or preferences. Hence, a new approach to egalitarianism has flourished rapidly, establishing individual responsibility as an important foundation of distributive justice [155], [157], [156]. For example, Rawls advocated the equality of *primary goods*, as income or rights, and recognized the redistribution of these goods strictly linked to social roles and responsibility. In the 1980s, the Nobel Prize economist Amartya Sen 1997 introduced the theory of *capabilities*, arguing that

equality does not lie in primary goods as defined by Rawls, but rather in functions. According to Sen equality is achieved when individuals have the same possibility to realize themselves and their values, i.e., to develop all the capabilities needed to actively provide for self-improvement. A remarkable progress in distributive justice theory is due to the jurist and philosopher Ronald Dworkin 1981, who distinguishes between preferences and resources by arguing that inequalities in outcomes are to be considered ethically unacceptable and require a distributive policy when they result from unequal resources, while they should not be redistributed when they are the result of individual preferences. As a result of Dworkin's contribution, scholars have begun to explore distributive mechanisms with respect to the individual's initial circumstances, one of the fundamentals of Equality of Opportunity (EOp) theory. Along this line, the first major studies on individual circumstances are due to the political philosophers Richard Arneson 1989 and Gerald Cohen 1989, who define equality as the possibility of obtaining a resource if it is sought. Although this definition may seem a nuance in theory, it is actually a significant way of separating individual responsibility and individual's circumstances. Both these two aspects affect the way and the possibility of an individual to achieve a goal or to obtain a resource. Thereafter, a large part of research has focused on defining which aspects lie within individual's circumstances and which ones on individual responsibility. By way of example, Barry 1991 considers individual choices fully determined by social circumstances, and considers minimal the ability of individuals to make choices outside of their own circumstances. This more progressive view is in opposition to a more conservative view that tends to treat individual choices as entirely belonging to the sphere of responsibility.

In our work, we focus the attention on the above mentioned post-Rawls literature which has identified *Equality of Opportunity*, rather than *Equality of Outcomes*, as the minimum goal for the egalitarianism. The most prominent formalization of Equality of Opportunity (EOp) in the economic field is due to John Roemer 1993, who has radically influenced the economists' approach to assess inequality. The key principle of Roemer's theory is based on the assumption that the resources obtained by individuals depend on two factors, namely (i) *individual choices*, which lie within the sphere of personal responsibility, and (ii) *circumstances*, which are exogenous to individual control [160]. Roemer therefore defines inequality of opportunity as the inequality of opportunity systematically associated with circumstances, and suggests measuring the degree of inequality through the effort made to achieve a certain objective.

5.4.2 Roemer's Formalization of the Equality of Opportunity Theory

The idea of Equality of Opportunity (EOp) formalized by Roemer [159] is based on the basic principle that the individual's achievement should depend on choice,

effort, and ability, and not on the circumstances of birth. Hence, four key principles characterize this theory: (i) *circumstances*, (ii) *effort*, (iii) *responsibility* and (iv) *reward*.

The first assumption that Roemer formulates on the idea of equality is referred to the so-called *principle of compensation*. He claims that if inequalities in a group of individuals are caused by birth *circumstances*, which include variables such as gender, race, or familiar socio-economic status, then these are morally unacceptable and must be compensated by society - *reward*. The second assumption is based instead on individual utility, or well-being, in relation to individual *responsibility*, also called the *principle of responsibility*. In fact, Roemer argues that the effort that individuals invest in achieving the acts they perform and for which they are fully responsible, in addition to the circumstances of birth, plays a key role. Therefore, a society that guarantees equal opportunities is a society in which results, well-being, or utility, are distributed independently to *circumstances*, and in which individual *responsibility* and *effort* are fully recognized. According to Roemer's general theory of EOp, policies should be oriented to equalize the opportunities that different *types*, or groups of individuals, categorized in accordance with diverse circumstances, need to have in order to achieve a given goal. A *type* is a group of individuals sharing the same circumstances, while the group of individuals characterised by the same degree of effort is called a *tranche*.

It is worth noting that one of the reasons why Equality of Opportunity is often associated with Roemer is due to the fact that he did not only propose and clarify its theoretical framework, but he was the first scholar to devise an operational algorithm that gave rise to an interesting empirical literature to which he contributed significantly. A first distinction between the various nuances deriving from the literature concerns the partitioning of individual characteristics into two categories, *effort* and *circumstances*. Explaining the differences in the various theories is beyond the scope of this work; for our purpose it is sufficient to point out that different partitions correspond to different notions of EOp.

More generally, the statistical approach suggested by Roemer to measure Equality of Opportunity is valid for any nuance of the theory. He assumes that each individual outcome y can be expressed as the result of a combination of effort e ($e_i \in \Phi$, where Φ is the set of all possible levels of effort) and circumstances c ($c_i \in \Omega$, where Ω is the set of all possible circumstances). The individual outcome is therefore produced by the function $g : \Omega \times \Phi \Rightarrow \mathbb{R}$ such that:

$$y_i = g(c_i, e_i) \tag{5.1}$$

The model presented is a purely deterministic model in which measurement errors or random components are neglected, as suggested by several authors [65], [68], [124], [153]. This problem is due to the fact that effort (e) is not a directly

observable datum, as well as the g function. To overcome some issues Roemer supposes that the g function is fixed and identical for each individual and introduces two basic hypotheses:

Hypothesis 1 (H1)**.** *The g function is monotonically increasing in effort (while subjective utility is commonly considered decreasing in standard notions of effort).*

Hypothesis 2 (H2)**.** *The distribution of effort is independent of circumstances.*

We will resume the treatment of the hypotheses thus formulated in Section 5.4.2. A second differentiation in the different approaches for the estimation of EOp is related to the partitioning of individuals into *types* and *tranches*.

$$M_{type,effort} = M_{i,j} = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,j} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ m_{i,1} & m_{i,2} & \cdots & m_{i,j} \end{pmatrix}$$

For the *ex ante approach*, or *type-compensation principle*, EOp occurs if the set of opportunities of different individuals is identical, independently from circumstances. Roemer states that *"it is good to transfer from an advantaged type to a disadvantaged type, provided that the ranking of types is respected. Suppose that between two types, one is unambiguously better off than the other, that is, the outcomes can be ranked unambiguously according to first-order stochastic dominance. Then a transfer from the dominant type to the dominated type for some effort level, ceteris paribus, is EOp enhancing"*[160]. The type approach focuses on differences in the perspectives of ex ante outcomes for classes of individuals with identical circumstances, thus focusing on inequalities between types and being neutral towards inequalities within types.
For the *ex post approach*, or *tranche-compensation principle*, EOp occurs if all those who spend the same level of effort achieve the same result. Roemer states that *"the closer each column is to a constant vector, the better. If for some effort (column), the inequality of outcome across types is reduced, and everything else remains unchanged, EOp has been improved"*[160]. In contrast to the type approach, the tranches approach focuses on ex post inequalities in classes of individuals with the same degree of effort. Consequently, the approach focuses on the distribution of inequality in outcomes within tranches.
Roemer's definition of EOp can therefore be summarized in the following model:

Population. Let consider a population of $1, ..., N$, individuals i, with an outcome y_i, assigned to a finite set of types $t = 1, ..., T$;

Fraction of population. Let f^t be the fraction of the population of type t;

Objective. Let an objective be given, i.e., a threshold set by the decision maker to reach EOp. The value of the degree to which an individual achieves an objective is a function of circumstances, effort and social policy θ ($\theta \in \Theta$, where Θ is the set of social policies)

$$u^t(e_i, \theta), \tag{5.2}$$

where u^t is the average achievement of the objective in type t that spend effort e when the policy is θ.

Effort distribution. Let $G^t_\theta(e_i)$ be the function of effort distribution in type t when the policy is θ.

Therefore, with the available set of data $T, G^t_\theta(e), f^t, u, \theta$ we can then rewrite the equation (5.1) in this way:

$$y_i = G^t_\theta(e_i) \tag{5.3}$$

Circumstances and Types

The identification of types and efforts requires society to have at least a similar, if not unified, view of how to distinguish actions and variables that belong to the sphere of individual responsibility and circumstances. Roemer's approach for measuring inequality of opportunity involves considering a situation as unequal if two individuals, who have both made the same choices and had different birth circumstances, have obtained a different outcome. The first step to make Roemer's method effective is to identify types, i.e., to identify the combinations of the circumstances' realization that partition the population into N subsets, in which each individual is included once and only once. The simplest empirical approach identifies types on the basis of socio-economic uniform features, such as gender, ethnicity, income, and then compute the value of opportunities according to the outcomes obtained by the individuals belonging to each type. Many machine learning systems actually adopt this methodology to achieve a fairness result, and the definition of discriminating circumstances is made on the basis of a historical discrimination that has led individuals belonging to these minority categories to be in a disadvantaged position [131]. Minority categories are therefore defined by identifying variables or proxy variables of real discrimination, and these variables, such as gender, ethnicity, place of birth, are called protected or sensitive attributes [198]. This kind of approach actually displays several methodological problems in the correct identification of types. Although straightforward and simple, the described method does not allow to take into consideration all those variables that contribute to shaping both the responsibility of the individual and the circumstances of birth. In general, Roemer does not address the problem of identification of types and circumstances, but over

the years several important empirical contributions have been provided to trace the structure of the method. Some of the most relevant are the inferential conditional trees proposed by Hothorn *et al.* [88], the non-parametric method by Checchi *et al.* [38], and the latent class models by Li Donni *et al.* [117]. It is beyond the scope of this work to analyse and discuss the trade-offs between the various methodologies proposed, therefore we focus only on the Hothorn methodology [88] that we deemed most effective in determining types. To the best of our knowledge, the sole work involving the algorithm proposed by Hothorn [88] was applied by Brunori *et al.* [30] to study socio-economic differences on panel data. In its general meaning, the algorithm for the determination of types exploits the permutation test theory developed by Strasser *et al.* [187] to generate recurring binary partitions overcoming the problem of overfitting and variable selection. In fact, recursion takes advantages of the conditional distribution of statistics that measure the correlation or association between the response variable and its covariates, and performs multiple hypothesis tests to determine the significance of the correlation or association. If it is not possible to identify a statistically significant correlation or association between the response variable and any of the covariates, recursion stops. In the algorithm we have implemented in our approach we use conditional inference trees to recursively partition the Euclidean space of the variables of the individuals in convex sets of hyperplanes. The convexity of sets is a fundamental property of this methodology because it allows us to affirm that individuals belong to one and only one subset, and therefore to one and only one type. We briefly describe below the steps of Hothorn's algorithm for conditional recursive inference trees to perform the identification of Roemer types (Algorithm 1 - Step 1).
Given a response variable Y and a set of covariates $X(x_1, ..., x_m)$ we assume that the conditional distribution of the response variable $P(Y|X)$ given the covariates is a function f of the covariates such that $P(Y|f(X))$. At each step the algorithm tests the partial null hypothesis of independence $H^0_{partial} : P(Y|X) = P(Y)$ between the response variable and any of the covariates, and stops if the hypothesis cannot be rejected at a certain level of α[3] previously selected; otherwise, it selects the covariate x_M with the highest correlation or association to Y through the *Simple Bonferroni-adjusted P-values*[4] that indicate the deviation from the partial hypothesis $H^0_{partial}$. The test is performed on each covariate to test the global null hypothesis. At the end of the procedure a set of N types is obtained, i.e., after multiple independence tests on each circumstance of individuals are executed.

[3]The value of α controls the probability of falsely rejecting H_0 at each node, and its use is the same to conventionally control Type I and Type II errors in hypothesis tests [88].

[4]Use *t tests* to make pair comparisons between group means, but check the overall error rate by setting the error rate of each test to the experimental error rate divided by the total number of tests. In this way, the level of significance observed is adjusted considering multiple comparisons are being performed (for further details see [26])

Circumstances and Effort

To discuss the effort estimate, we resume the assumptions H1 and H2 expressed in Section 5.4.2. The first hypothesis does not present particular problems, the second one poses more issues: individuals with more advantageous circumstances may consequently be more inclined to exert a greater degree of effort. In any case, it would be difficult to assign to an individual the accountability of her/his level of outcome if the degree of effort depends on exogenous circumstances. Thus, from a computational point of view estimating effort is one of the most complex aspects, as its difficulty in being observed is the result of a process of maximizing individual preferences. Since we assume that the effort is not directly observable, it is necessary to deduce its value from observable behaviours, i.e., a proxy measure is needed to measure and compare the effort of different individuals. The definition and measurement of effort by Roemer has changed over time. The definition to which we refer in this manuscript considers the relative individual effort determined not only by the variable of preference (the degree of effort); on the contrary the individual effort is determined by all the elements that establish the location of each individual in the distribution of the advantages that characterize the given type. Roemer argues that it exists an effort distribution function that characterizes the entire subgroup within which the location of the individual is set and what is needed is a measure of effort that is comparable between different types. The hypothesis at the basis of this assumption is that two individuals belonging to a different type t who occupy the same position in their respective distribution functions have exerted the same level of effort - and therefore of responsibility. Since, under the same circumstances, individuals who make different choices exercise different degrees of effort and thus achieve a different outcome. The differences in outcome within the same type are by definition determined by different degrees of effort, and therefore are not considered in the computation of the EOp. In general, Roemer states that to estimate effort it is necessary to aggregate individuals according to their circumstances (see *type estimation* in Section 5.4.2), to compare outcome distributions, and to measure the degree of effort an individual has exerted using the quantile she occupies in her type distribution. Since for H1 the outcome function is monotonous by definition and for H2 the effort is orthogonal to the circumstances, it is possible to measure the effort of an individual belonging to a generic type by the rank or quantile of the effort distribution in which that individual is positioned. Therefore, all the individuals positioned at the same quantile in the distribution of the respective type are by assumption characterized by the same level of effort. As we have highlighted in Section 5.4.2, the *ex-ante* and *ex-post* approaches express two different methods of achieving EOp. Hereafter we will refer to the *ex-post* approach, or *tranche-compensation principle*, which is the methodology we adopted.
Let the tranche vector $Y_{t,\lambda}$ be the set of outcomes enclosed in a given quantile λ of a type t: it expresses the different outcome values of individuals who exercised the

same degree of effort. Since the inequality in outcome within $Y_{t,\lambda}$ is not explained by this methodology, several papers propose to apply a smoothing function to eliminate this unexplained inequality [38], [30]. The standardized distribution of the outcome of an individual i, belonging to type t and located at quantile λ, is obtained by scaling each average outcome-tranche until all have the same mean of the total distribution, and it is expressed by the following equations:

$$y^t(G^t_\theta(e)) = y^t(\lambda) \quad \Rightarrow \quad F^t(y) \vdash y^t(\lambda), \tag{5.4}$$

where $F^t(y)$ is the cumulative distribution of outcomes in type t,

$$\tilde{y}^t_i(\lambda) = y^t_i(\lambda)\frac{\mu}{\mu^\lambda} \quad \Rightarrow \quad \tilde{F}^t(y) \vdash \tilde{y}^t_i(\lambda), \tag{5.5}$$

where $y^t(\lambda)$ is the outcome of an individual i in type t at given quantile λ, derived from the cumulative distribution of the type-specific cumulative distribution in Equation 5.4, μ is the mean of population's outcome, μ^λ is the mean of individual's outcome located at quantile λ over all types t.
In this way, observed inequalities are exclusively due to circumstances or degrees of effort; therefore, only inequalities resulting from exogenous circumstances are observed and not those arising from the responsibility of individuals. As Brunori *et al.* [30] suggest, for the smoothing process we adopt one of the proposed Bernstein's polynomial approximation applications [114], [215] and thus we obtain the standardized distribution of tranche vectors $Y_{t,\lambda}$. The methodology is described below.
The outcomes y of individuals can be considered as a sequence of random variables having a density function f supported by a closed interval $[a, b]$ and a cumulative distribution function F, where $y \in [a, b]$ and y is a positive continuous variable. The continuous density function f defined on $[a, b]$ can be approximated by a linear combination of Bernstein's polynomial bases of degree m, defined by the formula:

$$\tilde{f}_m(y) = \mathbb{B}_m(y, a, b) = \sum_{i=0}^{m} f(\frac{i}{m}) b_{i,m}(y, a, b), \qquad a \leq y \leq b \tag{5.6}$$

where $b_{i,m}(y, a, b)$ are binomial probabilities defining the Bernstein basis polynomials in a generalized polynomial space:

$$b_{i,m}(y, a, b) = \frac{1}{(b-a)^m}\binom{m}{i}(y-a)^i(b-t)^m - i, \qquad \forall i = 1, ..., m \tag{5.7}$$

The cumulative smoothed distribution of the outcome for type t $F^t(y)$ [Equation 5.4] with Bernstein's approximation is simply derived by estimating the density function for each type t, by approximating each function with Bernstein polynomials, and than by computing the integral function of $\tilde{f}_m(y)$:

$$F^t(y) = \int_a^b \tilde{f}_m^t(y)\, dy \tag{5.8}$$

To determine the degree of the polynomial that best approximates the function $\tilde{f}_m(y)$, we use the degree of the polynomial that maximizes the out-of-sample log likelihood by ten-fold cross-validation, as suggested by Brunori *et al.* [30].

5.4.3 Measurement of Inequality of Opportunity

In order to compute inequality of opportunity, an inequality index applied to the standardized distribution Y derived from the Equation 5.5 must be employed. The measurement of inequality of opportunity can be treated as a two-stage process:

1. The actual distribution of Y is transformed into a counterfactual distribution $\tilde{Y}$ which expresses the inequality in Y due to exogenous circumstances, while all the inequality due to individual responsibilities is removed;

2. Secondly, a measure of inequality is applied to $\tilde{Y}$, thus obtaining a measure of the "unfair" inequality due to circumstances.

However, computing the Equation 5.5 means getting an outcome vector Y in which the only inequality expressed is that within the tranches: an inequality index applied to this distribution captures exclusively and completely the outcome inequalities resulting from the circumstances, i.e., inequality of opportunity.
For this purpose we use the Gini index, a statistical concentration index that measures the degree of inequality of a distribution, commonly used to measure the distribution of income. The index lies in a range between 0 and 1: a low or equal to zero Gini index indicates the tendency to the equidistribution and expresses perfect equality; on the contrary, a high or equal to 1 value indicates the highest concentration and expresses the condition of maximum inequality. The Gini index calculus is based on the Lorenz curve of the distribution[5] (see Figure 5.1).

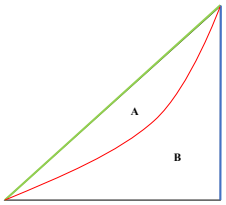


Figure 5.1: Graphical representation of the Gini index through Lorenz curve

[5]For further details on Gini index and Lorenz curve calculus see Lorenz [122], Gini [79] and Gastwirth [74]

The blue line represents the line of perfect inequality, the green line represents the line of perfect equality, or line of equidistribution, and the red line is the Lorenz curve. The area A between the lines of perfect equality and the Lorenz curve is called the concentration area and represents the deviation from perfect equality; Gini's index is the ratio between the area A and the total area:

$$GiniIndex = \frac{A}{A+B} \tag{5.9}$$

The inequality of opportunity through the application of the Gini index is therefore expressed by the following equation:

$$InequalityofOpportunity = GiniIndex(\tilde{Y}) \tag{5.10}$$

Inequality in Outcomes

Both Roemer's methodology and Gini's index provide a valuable contribution to the study of the concentration of transferable phenomena. A variable is called transferable when it can be transferred without fully transferring the unit itself; examples of non-transferable variables are the weight or height of individuals. The study of concentration, or study of transferability, is precisely the study of how a transferable phenomenon is distributed among the units, namely its attitude to concentrate in a reduced number of units. The economic-statistical disciplines are the fields where the study of concentration is most frequently pursued, especially to investigate the inequality in income and wealth distribution. When the concentration is high, hence there is an excess of this phenomenon, the condition of distributive inequality occurs. We draw on this theoretical premise to underline that Roemer's method necessarily leads to a reallocation of resources, which, although equitable for egalitarian theory, differs from the majority of studies in the field of fairness in machine learning systems. To the best of our knowledge, in this domain this is the unique study that considers the outcome of individuals as a transferable resource, in a way that it can be reallocated according to pre-established policies. Moreover, we specify that although the method necessarily leads to a reallocation of the individual outcome, the final goal of the methodology is not to equalize the results but rather to equalize the individuals' opportunities.

5.5 AFteRS: the Automated Fair-Distributive Ranking System

5.5.1 Problem Statement

Generally, the most popular measures to assess fairness in ranking systems are tied to three key concepts of Information Retrieval: (i) *utility*, (ii) *relevance*, and

(iii) *exposure*. The expected ranking r is the one that maximizes the general *utility* under a certain query q:

$$r = argmaxU(ranking_n|q) \tag{5.11}$$

Utility is commonly expressed as a mapping function β that under a certain query q maps the *relevance* of items to each user:

$$\beta(Rel(item_n|user_n, q)) \tag{5.12}$$

Exposure is computed after *relevance* has been established, and it indicates the probability of attention an item could get according to its *relevance* for the query q:

$$exposure = \frac{1}{\log_2(1+j)}, \tag{5.13}$$

where j is the position of the item i_n in the ranking r_n. When *relevance* is blind towards protected attributes, it is not rare that the average *exposure* for a minority group is substantially lower with respect to the majority group, despite differences in *relevance* are quite shallow. As highlighted by Singh and Joachims 2018, in a ranking where female individuals occupy slightly worse positions than men, they receive a significantly lower average of *exposure*. In this vein, a large majority of studies pose a fairness constraint so that average *exposure* is equally distributed among groups. Although valuable efforts have been made to establish an ethically acceptable fairness constraint while maintaining an adequate level of system utility, the most widespread fairness methodologies in ranking systems miss the key point of ethical programming. As a matter of fact, *exposure* is computed on the basis of item's position, and the positioning is derived by *relevance*. In order to obtain a bias-free ranking the way items embedding affects their positioning should be analyzed. As an example, we consider a ranking of eight potential candidates for a job, that belong to three different ethnic groups. White candidates have respectively 1, 0.98, 0.95 of *relevance*, Asian candidates have 0.93, 0.91, 0.88, and African-American candidates have 0.86, 0.84. A greater value of *relevance* determines a higher position in ranking, so that the best top-ranking is the one that places at first three positions the White candidates, at the middle the Asians, and at the bottom the African-Americans. As a result, the average aggregate *exposure* is 0.71, 0.39, 0.22 for Whites, Asians and African-Americans respectively. Figure 5.2 summarizes a generic *exposure* re-allocation process under demographic parity constraint. Given a query - *the most qualified candidates* - and items embedding - *candidates features* -, the best ranking under fairness of *exposure* constraint generate a re-allocation such that the average aggregate *exposure* is almost equalized. However, this widespread practice does not necessarily remove discrimination. As a matter of fact, although the *exposure* has been distributed in a fairly homogeneous

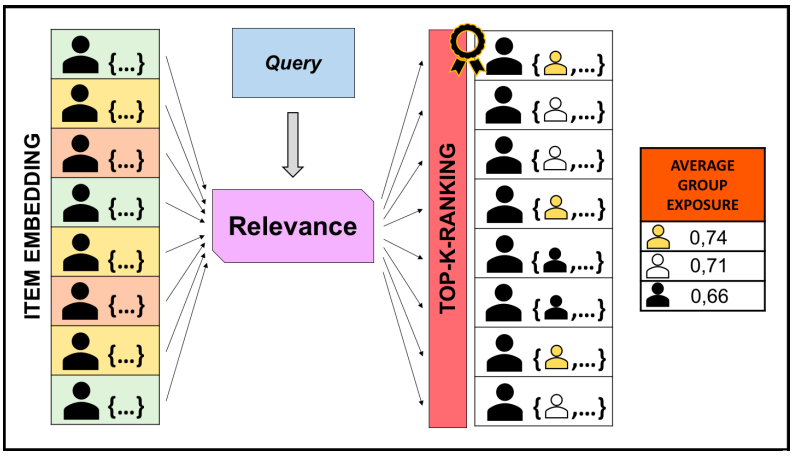


Figure 5.2: Demographic-Exposure Parity constraint. Item embedding colors define individuals with different sets of characteristics. The ranking position is determined by relevance. The top-k-ranking is the ranking achieved with the demographic parity fairness constraint. Although the aggregated exposure has been equalized, the ranking positions show that the relevance of the African-American minority remains lower than the other two groups (yellow: Asian; white: Caucasian; black: African-American.)

way among classes, the African-American minority group remains in the bottom-middle ranking positions (Figure 5.2). In fact, if the ranking is systematically influenced by protected attributes, in this case the ethnicity, the ranking positioning will not be affected by *exposure* re-allocation. The example presented so far supports the idea that the relevance and the positioning bias may constitute a basis to improve fairness in ranking systems domain. The methodology we propose aims at contributing to the debate in the following ways:

1. Exposure fairness constraint is substituted by positioning and relevance fairness constraints;

2. Merit of exposure is substituted by exerted effort, that is a proxy variable indicating in what extent individuals are responsible for their positioning;

3. Group membership is substituted by type membership, where *type* is a vector of the items embedding possible realizations.

5.5.2 Model

Figure 5.3 shows the steps of the Automated Fair Distributive Ranking system. Since AFteRS is based on Roemer's EOp theory (Section 5.4.2), before to compute ranking we need to derive the two dimensional list that fully describes individuals:

(i) the type vectors by partitioning the whole population, representing the set of circumstances beyond individual's control (Step 1, see also Section 5.4.2), and (ii) the effort vectors by computing quantiles of each type distribution, representing a list of attributes for which individuals are entirely responsible (Step 2, see also Section 5.4.2). After these two steps, the model generates the top-Fair-Distributive-ranking (Step 3) according to a set of policy Θ representing a pool of fairness constraints.

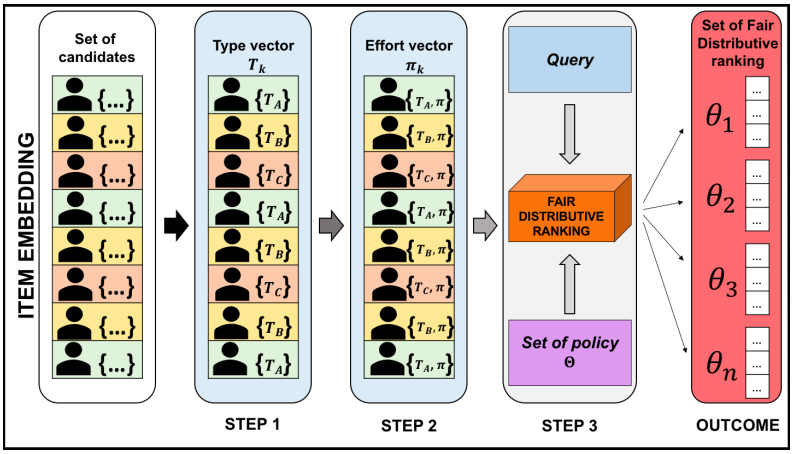


Figure 5.3: Graphical representation of the Automated Fair Distributive Ranking System

The Fair Distributive Ranking *Gamma* is derived by Equations 5.11 and 5.2, where utility is a function of type vector, effort vector and a policy given a query q:

$$\Gamma = argmax_{\theta \in \Theta} u^t(q|e_i(\lambda), \theta) \tag{5.14}$$

As a result, Equation 5.15 describes the fair exposure allocation among types, denoting the opportunity-equalizing policy under Roemer assumptions. The Γ ranking exposure is derived by maximizing the area below the lowest function exp^t, i.e., the type-exposure:

$$\max_{\theta \in \Theta} \int_0^1 min_t exp^t(\lambda, \theta)\, \mathrm{d}\lambda \tag{5.15}$$

The properties of the Γ ranking are the following:

(i) Type Fairness Constraint: ranking Γ is bounded to exploit the population partition by detecting a set of non-overlapping types. This process allows to overcome the a-priori protected attributes assignment and studying a wider set of individuals' traits;

(ii) Type-Effort Centered Ranking: ranking Γ is derived by scaling the type-outcome distribution to tranche mean outcome (Equation 5.5). In this way,

all the unexplained inequality is removed and a first re-allocation based on effort-inequalities is computed;

(iii) Ordered-Utility Ranking: ranking Γ is ordered by decreasing utility, i.e., it selects the best candidates from the whole population under specific policy constraints.

Policy

While in artificial intelligence and machine learning the terms justice and fairness are often used interchangeably, several other studies [2], [7], [43] define justice "as the perceived adherence to rules that reflect appropriateness in decision contexts" [43], while fairness is defined as "global perception of appropriateness" [43]. Our ranking system aims at following this framework. Generally, there are four criteria for measuring justice: (i) procedural, (ii) distributive, (iii) interpersonal, and (iv) informational, each of them subdivided into several other sub-criteria. Since the purpose of our model is to re-allocate the individuals' outcomes in a ranking, we focus only on the distributive criterion, which shows the following sub-criteria:

i *Equity*: "outcomes are allocated according to contributions" [43];

ii *Equality*: "outcomes are allocated equally" [43];

iii *Need*: "outcomes are allocated according to need" [43].

Thanks to the separation of circumstances and effort, Roemer's EOp theory is well suited to codify this methodological framework. In our model, each distributive sub-criterion represents a different policy actuation; therefore, each policy exploits Roemer's EOp theory to redistribute the outcome in a diverse manner. Table 1 summarizes the different fairness criterion embedded in each policy.

Policy	**Criterion**
Equity	Roemer's EOp Theory
Equality	Demographic Parity in Roemer's EOp Theory
Need	Demographic Parity in Supervised Learning

Table 5.2: Policies' Criteria

Equity. Once circumstances (Section 5.4.2) and effort (Section 5.4.2) are derived, an analysis of inequity between types is performed. Figure 5.4 and Figure 5.5 show the Cumulative Distribution functions of outcome for type A and B. For effort degree 0.8, type A and B get on average approximately 14 and 16 values of outcome

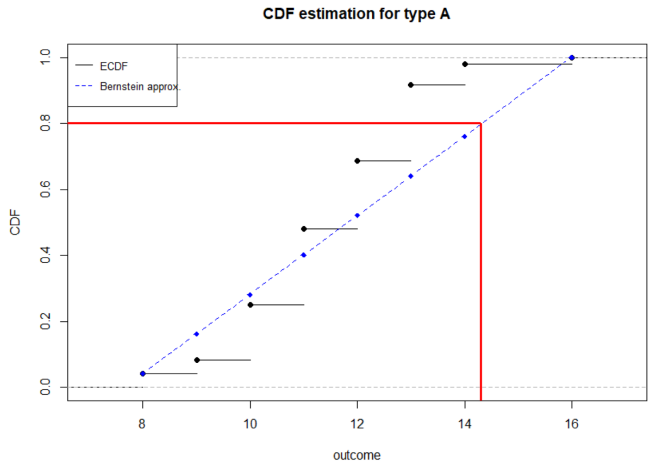


Figure 5.4: Effort-outcome Cumulative Distributions' Functions for type A. Example illustrating how the effort estimation method works.

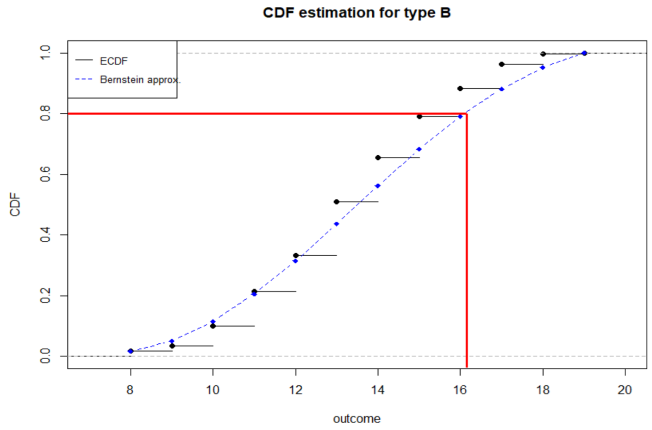


Figure 5.5: Effort-outcome Cumulative Distributions' Functions for type B. Example illustrating how the effort estimation method works

respectively. This means there exists inequity between type A and B. To measure the extent of the inequity, we perform an inequity decomposition of Gini Index by tranche approaches, as to say we calculate Gini for each type at every degree of effort. The purpose of the decomposition is to evaluate what are the types and effort degrees that most affect the Gini value. Once an association has been established, outcomes are re-allocated among types according to the decomposed Gini Index. The re-allocation produced a new outcome for all individuals, i.e., a counterfactual outcome $y_{\neg}$ that indicates the outcome individuals would have gotten if they had not belonged to their type. The counterfactual outcome is therefore a function of

a standardized distribution of tranche vectors $Y_{t,\lambda}$ (Equation 5.5) and decomposed Gini:

$$y_{\neg} = f(Y_{t,\lambda}, Gini_{t,\lambda}) \tag{5.16}$$

The Equity policy then reorder individuals in ranking according to counterfactual outcomes. Notice that the method used to develop the counterfactual outcome allows to evaluate the amount of both individual and aggregate redistributed outcomes. This procedure is possible when the outcome to be redistributed is a transferable variable, i.e., when ideally a statistical unit can cede a part or all the intensity possessed to another statistical unit.

Equality. The policy is derived from the Demographic Parity criterion in supervised learning [83]. The main differentiation is that, while the common definition assess parity among individuals belonging to a protected group (e.g., gender, ethnicity), here Equality Policy assesses parity among individuals belonging to Roemerian types. Once circumstances and effort have been established, the policy makes use of the standardized outcome $\tilde{y}$ (Equation 5.5) to reorder individuals in ranking by satisfying a parity-position constraint (Figure 5.6).

Notice that this policy doesn't produce a counterfactual outcome as the Equity one but provides a new ranking order.

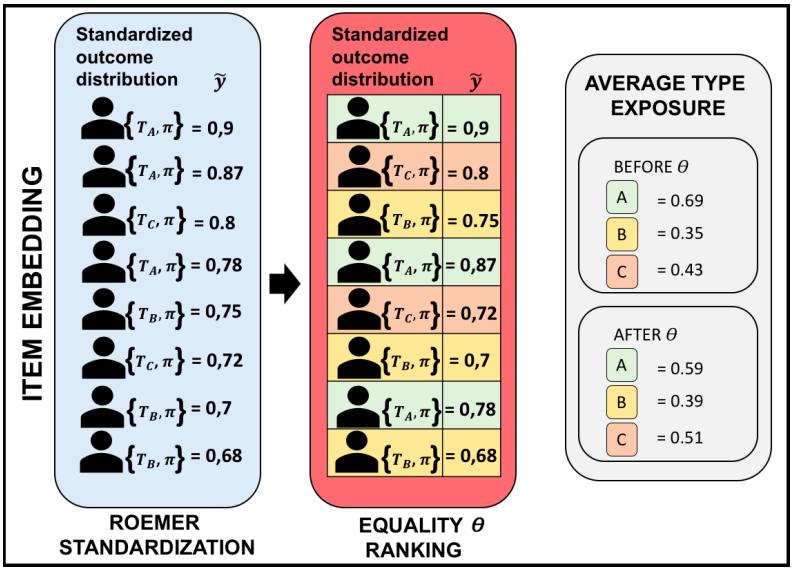


Figure 5.6: Example of a reordered ranking after Equality Policy computation

Need. While the Equality policy makes use of Roemerian types to assess the Demographic Parity criterion, the Need policy implements the original constraint through the employment of protected attributes. The Need policy process works

as the Equality one shown in Figure 5.6, with the exception that individuals are grouped by one or more sensitive attributes. The key difference among the two policies is that, while the Need one requires to a-priori establish protected attributes to reorder the ranking, the Equality one takes into account a wider range of attributes based on their association with the outcome (see Section 5.4.2). Notice that it is possible to apply Need policy in order to satisfy different multiple-constraints by changing the grouping attribute. For instance, it is possible grouping the population by qualification levels in order to constrain the policy to return the ranking with the maximum utility for the decision maker. In this way, the Need policy computes a more relaxed version of the concept of Need as a sub-criterion of distributive justice, as the sub-criterion is no longer applied to the individuals to be ranked but to the decision maker.

Algorithms 1 and 2 provide an overview of the entire process.

Algorithm 1 Automated Fair Distributive Ranking – Step 1-2 (Figure 5.3)
The algorithm partitions the population in n types (Section 5.4.2), derives effort (Section 5.4.2), and computes the Standardized Distribution (Equation 5.5)

Step 1
input: dataset D
output: non-overlapping subsets of D $\implies$ population partitioned in T_k types

for all $X_i \in D$ **do**
  Test the null hypothesis of independence between Y and all X_i
  if $H^0_{partial} : P(Y|X) = P(Y)$ couldn't be rejected **then**
    Stop
  else
    1. select X_i with the strongest association to Y (smallest *adj p-value)*
    2. find the splitting point $C*$ for X_i such that
      $S^{x_i}_n \subset \chi_i$ are all the possible disjoint sets of the sample space χ_i
  end if
end for
return T_k vectors $\subset D$

Step 2
input: T_k vectors $\subset$ D
output: Standardized Outcome $\tilde{y}^t_i(\lambda)$

partition each T_k in 10 sets Ψ_n, such that $\Psi_{k,n} \subset T_k$
training set $1^{st} - 9^{th}$ T_k sets
test set 10^{th} T_k set
for all $\Psi_{k,n} \subset T_k$ **do**
  1. perform the Bernstein polynomials log-likelihood on the training set to estimate
    the best type-distribution approximation $LL_B(p_m = \sum_{i=1}^{n} \log f_B(x_j, p_m))$
  2. predict the CDF of T_k on the test set
end for
estimate the *Standardize Distribution* $= y^t_i(\lambda)\frac{\mu}{\mu^\lambda}$
return $\tilde{y}^t_i(\lambda)$

Algorithm 2 Automated Fair Distributive Ranking – Step 3 (Figure 5.3)
The algorithm computes the Γ ranking based on policies Equity, Equality and Need

Step 3
input: Standardized Outcome $\tilde{y}_i^t(\lambda)$
output: ranking Γ constrained by a policy $\theta \in \Theta$

if $\theta = equity$ **then**
for all $\tilde{Y}_{t,\lambda} \in D$ **do**
    compute the counterfactual outcome from stnd. outcome and decomposed Gini
    $\Gamma \leftarrow$ ranking ordered by counterfactual outcome
end for

if $\theta = equality$ **then**
for all $T_k \in D$ **do**
    $sorted_{T_k} \leftarrow$ type-ranking ordered by decreasing stnd. outcome
end for
for all (j) $\in sorted_{T_k}$ **do**
    $row_n \leftarrow j$ element of $sorted_{T_k}$
    array[j] $\leftarrow row_n$ ordered by decreasing std. outcome
end for
Γ $\leftarrow$ *merge* all j array

if $\theta = need$ **then**
$G_k \leftarrow n$ subsets $\in D$ grouped by protected attribute A
for all (z) $\in sorted_{G_k}$ **do**
    $row_n \leftarrow z$ element of $sorted_{G_k}$
    array[z] $\leftarrow row_n$ ordered by decreasing std. outcome
end for
Γ $\leftarrow$ *merge* all z array

return Ranking Γ

5.6 Evaluating AFteRS

To evaluate our Ranking System we draw an experimental design that consists in implementing a set of rankings based on the policy set Θ (Section 5.5.2) and a benchmark ranking α without fairness constraints. We compare the entire set of rankings through diverse metrics (Section 5.6.2) to evaluate the performance of each policy in all the top-N-ranking.

Data. To develop our Fair Distributive Ranking, we use the Student Performance DataSet[6] [45] that consists of a single-year performance scores of students belonging to two Portuguese schools. The dataset contains 649 instances and 33 attributes.

5.6.1 Metrics

The ranking system evaluation is carried out through various metrics. Since our approach is built on a set of theories across several fields of study, we employ three types of metrics from different domains for a more comprehensive analysis: i) ranking domain metrics, ii) inequality domain metrics, and iii) a set of metrics we propose to study our fairness constraints. All metrics formulas are summarized in Table 5.3.

Ranking metrics. We employ three metrics from ranking domain: (i) expected ranking, (ii) relevance, and (iii) exposure. As shown in Section 5.5.2, after policy computation the general definitions of expected ranking and exposure are shaped as in Equations 5.14 and 5.15 respectively. The metric relevance is mostly involved in computing the overall utility, where utility denotes the system's capability to generate a relevant ranking with respect to a query q.

Inequality metrics. We compute six inequality metrics: (i) Gini index [79], (ii) Theil index [192], Richness, (iii) Margalef Index [128], (iv) Shannon-Wiener index [173], and (v) Simpson Index [177]. As shown in Section 5.4.3, Gini is a statistical concentration index used to measure the level of inequality in a distribution; it ranges from 0 to 1. We apply Gini in two different circumstances. First, Gini evaluates the degree of inequality in both original and final distribution. Then, we apply a decomposition on Gini index to compute the counterfactual score in equity policy (Section 5.5.2). Theil index is an entropy measure commonly widespread in statistical economics to study segregation, where a zero Theil value indicates perfect equality. We apply Richness, Margalef index, Shannon-Wiener index and Simpson index to perform the diversity analysis in types. Richness and Margalef

[6]url: https://archive.ics.uci.edu/ml/datasets/Student+Performance. Last access: 2020-05-04

Metric	Formula	Input
Expected ranking	$r = argmaxU(ranking_n \mid q)$	ScDistr.
Exposure	$\frac{1}{\log(1+j)}$	Original Distribution
Relevance	$\beta(Rel(item_n \mid user_n, q))$	ScDistr., adj ScDistr.
Expected ranking-policy	$\Gamma = argmax_{\theta \in \Theta} u^t(q \mid e_i(\lambda), \theta)$	Adj ScDistr.
Exposure-policy	$\max_{\theta \in \Theta} \int_0^1 min_t exp^t(\lambda, \theta)\, \mathrm{d}\lambda$	Adj ScDistr.
Gini Index	$1 - \frac{1}{\mu} \int_0^\infty (1 - \mathrm{F(y)})^2 \mathrm{d}y$	All distributions
Decomposed Gini	$Gini_\lambda^t$	Stand. ScDistr.
Richness	n^t	Types diversity
Margalef	$\frac{T-1}{lnN}$	Types diversity
Shannon-Wiener Index	$H = \sum_{i=1}^R p_i \ln p_i$	Types diversity
Simpson	$1 - \sum \frac{n^t(n^t-1)}{N(N-1)}$	Types diversity
Theil Index	$\frac{1}{N} \sum_{i=1}^N ln(\frac{\mu}{y_i})$	All Distributions
Opportunity-Types Profile	$min/max(y^t - \mu(y))$	ScDistr.
Opportunity-Types Rate	$y^t - \mu(y)$	ScDistr
Opportunity-L/G Profile	$min/max(y_\lambda^t - \mu(y_\lambda))$	StndDistr.
Opportunity-L/G Rate	$y_\lambda^t - \mu(y_\lambda)$	StndDistr.
Unexplained Inequality Rate	$\frac{1}{N} \sum y_i - \tilde{y}_i$	ScDistr, StndDistr.
Reward Profile	$min/max(j(y_\lambda^t) - j(adj(\tilde{y}_\lambda^t)))$	ScDistr, adj ScDistr.
Reward Rate	$j(y_\lambda^t) - j(adj(\tilde{y}_\lambda^t))$	ScDistr, adj ScDistr.

Table 5.3: Summary of metrics employed. *Notation*: F(y)= cumulative distribution function of the score, μ = mean score; R = number of types, p_i = frequency of types; y_λ^t = score distribution aggregated by type and quantile; $\tilde{y}_i$= standardized score; $adj(\tilde{y}_\lambda^t)$= adjusted mean-type score at each effort degree (after policy); j = ranking position

index are simple diversity measures we apply to record the number of individuals and to measure abundance in each Roemerian type. Shannon-Wiener is a statistical diversity index indicating abundance or lack of species in a given population; in our case is employed to study diversity in Roemerian types. Simpson index is a measure of dominance of the most common species. In our case, it indicates the Roemer types' dominance.

Distributive Fairness Metrics. We propose a set of new metrics to evaluate fairness in automated distributive systems. The Distributive Fairness metrics are shaped to specifically measure various dimensions of inequality in our model by exploiting the notions of Roemerian types and of distributive re-allocation. They can be easily adopted in more general contexts by replacing the notion of types with that of groups. The Opportunity-Types Rate and the Opportunity-Types Profile are employed to study inequality in types. They indicate respectively which score

each type reaches on average and which types are most and less advantaged based on mean score distribution. The Opportunity-Loss/Gain Set and the Opportunity-Loss/Gain Profile act in a similar way by computing the mean-types score on a standardized score distribution (Equation 5.5) at each effort degrees (Step 2 in Figure 5.3). As a result, the effective inequality in the original distribution can be assessed after the fair inequality has been removed, in other words, after removing inequality due to individuals' responsibility. The Unexplained Inequality Rate computes the amount of the total fair inequality removed. The Reward Profile and the Reward Rate are applied to final distributions to evaluate the extent of policies re-allocation (Step Outcome, Figure 5.3). They respectively calculate the most and less advantaged types by the scores re-allocation - i.e., after applying fairness constraints - and the average re-allocation of the scores for each type.

Purpose of Metrics for Research Questions. The great abundance in metrics we employ arises from the necessity to measure different interdisciplinary aspects of our approach. In order to offer a clearer comprehension of their use, we show their connection with research questions in Table 5.1.

Research Question	**Sub-group**	Metric domain
RQ1	Fairness in automated decision-making	Inequality metrics Distributive Fairness metrics
RQ2	Fairness in ranking	Ranking metrics
RQ3	Model evaluation Inequality metrics	Ranking metrics Distributive Fairness metrics

Table 5.4: Metrics and Research questions overview

5.6.2 Results

To provide a clearer understanding of the results in relation to the research questions (Tables 5.1 and 5.4), a schematic summary of each Subsection's contribution to answering the research questions is given below.

- **RQ1** Subsection 5.6.2 provides an in-depth perspective on the results of applying a distributive justice and equal opportunities perspective in an automatic decision-making system. The Subsection follows Algorithm 1 (Step 1 and 2) by applying the conditional inferential tree Hothorn's algorithm to

solve the effort estimation dilemma in Roemer's theory, and provides a detailed description of the difference resulting from applying a fair distributive algorithm compared to applying an algorithm that does not have this type of constraint. The fourth row of Table 5.8 shows a minimum deviation in the outcome of the two algorithms, a sign that the distributive justice algorithm does not substantially affect the quality of the results. This Subsection represents the convergence of the three research questions formulated in Table 5.1. In Subsection 5.6.2 the results of the simulation of the automatic decision-making system indicate that the trade-off between fairness and decision quality is subject to the type of policy that is employed and not to the distributive fairness constraint per se.

ii **RQ2** The assumptions of good quality of the results of a distributive fair ranking algorithm are verified alongside those of **RQ1** in Subsection 5.6.2. Further detailed analysis for the case of ranking systems is provided in subsection 5.6.2. Figure 5.8a shows a substantial decrease of the Gini Index in the distribution derived from the application of Algorithm 1, indicating that the distributive fair ranking algorithm improves the conditions of inequality. In Figure 5.8b the results of the algorithm show that the Opportunity Loss/Gain Rate after redistribution for each type is not high, suggesting distributive acceptable results in individual terms. In Figure 5.9 the tests on fairness utility trade-off show positive results especially for Equity and Equality policies. The assumptions of better and fair exposure results are confirmed as shown in Figure 5.10.

iii **RQ3** The evaluation of the factors that affect the trade-off is investigated in two steps. The first step consists of the preliminary analysis (Subsection 5.6.2) of the data and how the types' outcome is distributed; the second step involves the analysis of the trade-off with respect to the applied policies (Subsection 5.6.2.) The results show that the composition of the data is crucial in the redistribution process, suggesting that the more inequality exists in the initial data, the greater the trade-off between utility and fairness (Table 5.5). Moreover, the kind of applied distributive justice criterion (in our case, the policy) is also significant in the trade-off (Figure 5.9). More specifically, the trade-off between fairness and utility is lower in the Need policy, where the results are worse, while it is higher in Equity and Equality policies where both utility and fairness achieve better results.

Preliminary Analysis of Types

We perform a preliminary analysis of Roemerian types in order to both study types' composition and evaluate type-specific inequality. First, we compute the conditional inference tree (Section 5.4.2) to extract Roemerian types. The result of

the process is shown in Figure 5.7. A complete description of types' composition is

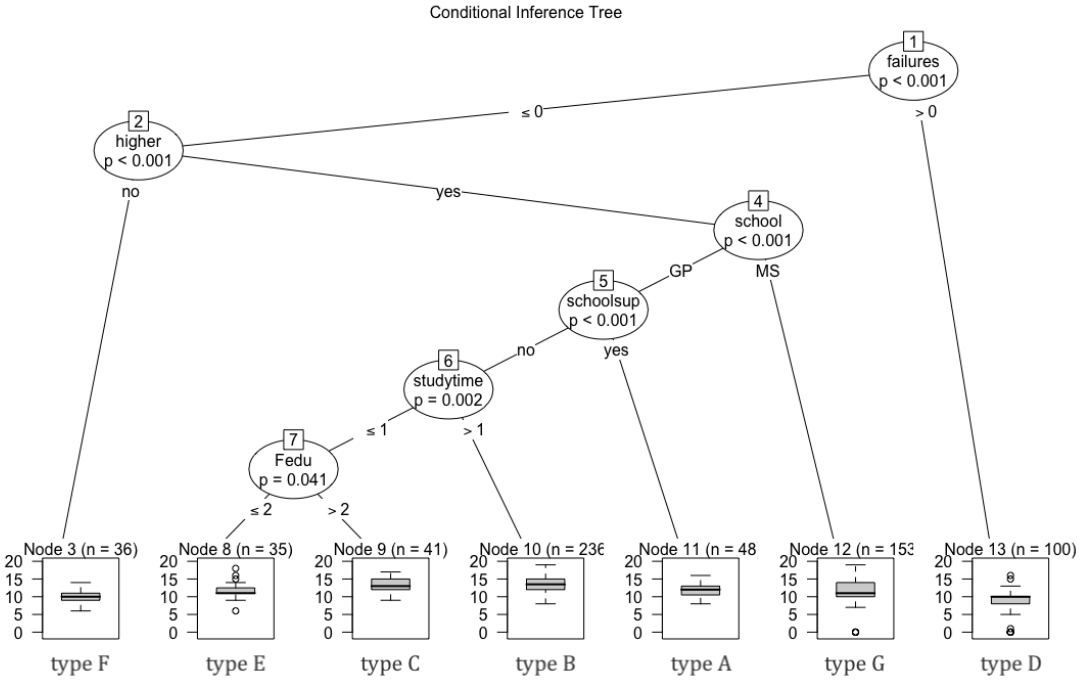


Figure 5.7: Conditional inference tree resulting from Step 1 of Algorithm 1

provided in the following list:

- **Type A**: consists of individuals with no failures in past class, that want to take higher education, belonging to the school "GP" and who need extra educational support.
- **Type B**: consists of individuals with no failures in past class, that want to take higher education, belonging to the school "GP", who do not need extra educational support and study more than 2 hours.
- **Type C**: consists of individuals who have no failures in past class, want to take higher education, belonging to the school "GP", do not need extra educational support, study less than 2 hours and have a father with secondary or higher education.
- **Type D**: consists of individuals with 1 or more failures in past class.
- **Type E**: consists of individuals who have no failures in past class, want to take higher education, belonging to the school "GP", do not need extra educational support, study less than 1 hour and have father with education between 5th and 9th grade, or primary education, or none.

- **Type F**: consists of individuals who have no failures in past class, do not want to take higher education.
- **Type G**: consists of individuals with no failures in past class, that want to take higher education, belonging to the school “MS".

Secondly, we perform an analysis of inequality through the set of metrics described in Section 5.6.1. Table 5.5 briefly summarizes the results of the preliminary analysis. This preliminary analysis on dataset shows that B and C types achieve on average a

Metric	**Results**							
	Types							Overall
	A	B	C	D	E	F	G	
μ outcome	11.58	13.59	13.22	8.6	11.63	9.89	11.76	11.9
Exposure	0.32	0.37	0.35	0.28	0.32	0.29	0.34	0.34
Relevance	0.61	0.72	0.7	0.45	0.61	0.52	0.62	0.63
Gini Index	0.0764	0.0935	0.0891	0.1860	0.0893	0.0961	0.1590	0.145
Shannon-W.	0	0.754	0.325	0.089	0.079	0.201	1	-
Theil Index	0.0097	0.0137	0.0125	0.0239	0.0153	0.0152	0.0242	0.025
Opp.-TR	0.1966	1	0.8506	-1	0.2146	-0.4808	0.2665	-

Table 5.5: Main results of preliminary analysis

higher outcome with respect to other types, while D and F achieve on average less. At first glance, it seems that a low outcome is associated with the willingness of the students to take higher education, while higher outcomes do not show a prevailing association. We set the query as “best score" and compute the ranking according to q to study exposure and initial relevance. Since the exposure (Equation 5.13) is expressed as position bias and is used on a logarithmic scale, a slight deviation of the value actually indicates a big change in the true exposure of individuals in the ranking. The Gini index decomposition highlights the residual inequality for each type, where a value tending to zero indicates perfect equality - in other words, outcomes are distributed equally inside types. At this early stage of the analysis, the above metrics are not very effective in estimating actual inequality, while they become robust when compared to the standardised distribution results. We employ Richness, Simpson, Margalef and Shannon-Wiener indexes to perform the diversity analysis of types. Results of diversity analysis are shown in Table 5.6. We use Richness and Margalef metrics to indicate the diversity of outcomes in each type. Richness is built by counting the number of different outcomes in each type, while Margalef considers also the density of each outcome. Types with more diversity of outcomes are considered more complex and therefore richer. By considering solely these two metrics, D and G result the more diversified types. Simpson’s and Shannon’s indices are more complex measures than Richness and Margalef metrics

	Richness	Margalef	ShannonW	Simpson
A	8	1.81	0.00	0.23
B	12	2.01	0.75	0.90
C	8	1.88	0.33	0.66
D	13	2.61	0.09	0.00
E	10	2.53	0.08	0.04
F	9	2.23	0.20	0.37
G	14	2.58	1.00	1.00

Table 5.6: Results of diversity analysis

that indicate how many levels of outcome are present in the types, considering both the Richness of the outcome and the dominance of one outcome over another within each type. The Shannon-Wiener index is more oriented to measure diversity, while the Simpson index to measure dominance. Specifically, the Simpson's index weighs the abundance of the most common outcomes within types, indicating the probability that two outcomes randomly extracted from the population belong to different types. Higher index values indicate greater diversity. We consider types A and C to show the joint interpretation of the indexes (Table 5.6). In this case both types show the same Richness value and a very similar Margalef value, i.e., they show the same richness of outcome. Type C, however, has a lesser dominance of outcome, i.e., the outcome is more evenly distributed within the type and this means that type C has a greater diversity. In this way, we see that D is actually the type with the least diversity, since it shows higher Richness values than the other types but very low diversity indexes. This means that within the type there is a strong dominance of some outcomes over others. Both diversity indices are expressed in a range between 0 and 1 to facilitate comparative analysis.

Effort and Standardized Distribution Analysis

We perform Bernstein-Likelihood polynomials in order to find the distributions that best approximate types (Section 5.4.2) and then we set 10 quantiles to exploit 10 levels of population efforts (Section 5.4.2). Each quantile is populated as reported in Table 5.7.

As highlighted by Brunori et al. 2020, a number of true quantiles doesn't exist, and thus they suggest performing the Bernstein polynomials and the Hothorn algorithm in order to find the best approximation of the unknown continuous distribution functions. The standardized distribution is derived from the cumulative distribution functions of each type. Table 5.8 shows the impact of the standardization process by reporting differences in type-effort outcome distributions. As a result of standardization the individual's outcome undergoes a transformation; in fact, unexplained inequality - i.e.„ inequality due to individual responsibility

	1	2	3	4	5	6	7	8	9	10
A	4	8	11	0	0	10	0	11	3	1
B	37	36	3	26	16	26	9	23	29	31
C	4	5	0	7	0	7	0	0	7	11
D	10	1	4	4	18	12	32	12	2	5
E	1	1	6	0	12	6	5	0	2	2
F	3	5	6	0	4	4	0	9	3	2
G	6	9	35	28	15	0	19	10	19	12

Table 5.7: Effort-Types frequency table

- is removed. The effect of the process is therefore to reduce the degree of inequality, which means carrying out a preliminary distributive treatment (Figure 5.8a). By analyzing the process results, it appears that standardization does not substantially affect one type specifically, but acts homogeneously on the tranches (type-effort vectors). The Unexplained Inequality Rate -0.128 is in fact quite small - computed by applying the formula listed in Table 5.3. Figure 5.8 reports the comparison among the original and the standardized distribution. The standardization process reduces the degree of inequality both in the overall and in the type distributions. Since distributive processes mitigate inequality by reallocating resources, we calculate the Opportunity-Loss/Gain rate to evaluate the reallocation effects on each type. Figure 5.8b shows the main analyses of the preliminary redistribution effects. The first two outputs indicate the type outcome's deviation from the population mean, which is decreased after standardization, as can be observed. The Opportunity-Loss/Gain rate measures the extent of the preliminary redistribution showing the average advantage and disadvantage per type. Analyses report that D is the only type that exhibits a positive difference in outcome after the process, although minimal. This result is due to two factors: first, the standardization process is more effective in more densely populated subgroups; and second, the process tends to overestimate the outcome in less diverse subgroups. Since the latter factor is actually the result of an intermediate step in our model, we do not bother to balance it at this system stage. The balancing of this factor is discussed in Section 5.6.2.

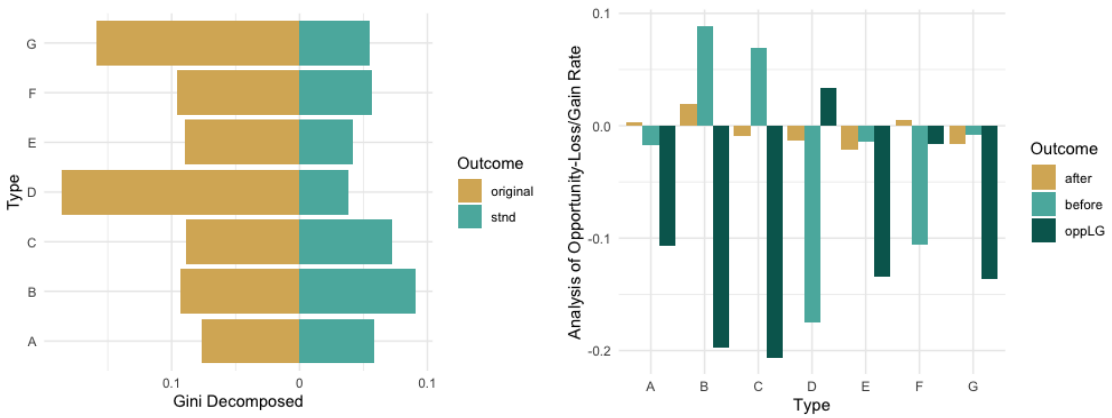
Top-N-Ranking Under Fairness Constraints

We perform three different policies belonging to three distributive justice subcriteria. The goal is to test the response of a ranking system to different distributive fairness theories. We test the policies on all top-N-ranking at 50 intervals, i.e.„ top-50-ranking, top-100-ranking, and so on. Figure 5.9 shows the inequality-utility trade-off for all ranking policies. Results show that Equity constraint performs bet-

t	1	2	3	4	5	6	7	8	9	10
A	0.40	0.49	0.55			0.49		0.50	0.53	0.57
A	0.03	0.00	0.04			-0.00		0.00	0.03	0.05
A	0.76	0.65	0.67			0.56		0.67	0.70	0.81
A	-0.36	-0.15	-0.13			-0.07		-0.18	-0.18	-0.25
B	0.62	0.51	0.43	0.51	0.47	0.51	0.47	0.51	0.50	0.49
B	0.25	0.01	-0.07	0.01	-0.04	0.02	-0.02	0.01	0.00	-0.02
B	0.89	0.75	0.59	0.87	0.78	0.80	0.85	0.94	0.99	1.00
B	-0.27	-0.24	-0.16	-0.36	-0.31	-0.29	-0.37	-0.43	-0.49	-0.51
C	0.49	0.50		0.49		0.51			0.46	0.50
C	0.12	0.01		-0.01		0.01			-0.04	-0.02
C	0.84	0.72		0.90		0.69			0.75	0.99
C	-0.35	-0.22		-0.42		-0.18			-0.29	-0.49
D	0.37	0.53	0.53	0.47	0.50	0.49	0.50	0.50	0.49	0.48
D	0.00	0.03	0.03	-0.02	-0.01	-0.01	0.01	0.00	-0.01	-0.03
D	0.47	0.59	0.24	0.16	0.41	0.59	0.32	0.14	0.00	0.01
D	-0.10	-0.07	0.29	0.31	0.09	-0.11	0.18	0.36	0.49	0.47
E	0.11	0.48	0.48		0.47	0.48	0.52		0.50	0.52
E	-0.26	-0.02	-0.02		-0.03	-0.01	0.02		-0.00	0.01
E	0.70	0.70	0.76		0.58	0.74	0.66		0.49	0.81
E	-0.59	-0.22	-0.28		-0.10	-0.25	-0.14		0.01	-0.29
F	0.43	0.48	0.54		0.56	0.48		0.49	0.53	0.53
F	0.06	-0.02	0.03		0.06	-0.01		-0.00	0.03	0.02
F	0.45	0.38	0.52		0.57	0.51		0.43	0.45	0.59
F	-0.02	0.10	0.01		-0.01	-0.03		0.07	0.07	-0.06
G	0.15	0.47	0.48	0.50	0.53		0.49	0.48	0.50	0.50
G	-0.21	-0.02	-0.02	0.01	0.02		-0.00	-0.02	0.00	-0.01
G	0.78	0.51	0.73	0.59	0.75		0.56	0.75	0.77	0.69
G	-0.62	-0.04	-0.24	-0.09	-0.22		-0.07	-0.27	-0.26	-0.18

Table 5.8: Standardized outcome descriptive statistics. Columns represent different levels of effort. For each type (row) a different result is displayed. First row: standardized outcome. Second row: deviation of type-effort standardized outcomes from the mean. Third row: original outcome. Fourth row: difference between standardized and original outcome. Missing values denotes non-populated tranches.

ter in keeping low levels of inequality but sacrificing general utility in the first top-N-ranking. On the contrary, Equality constraint performs better in keeping general utility but shows a higher degree of inequality for almost all the top-N-ranking with respect to Equity constraint. Finally, Need constraint, despite displaying almost



(a) Gini Index

(b) Opportunity-Loss/Gain Rate

Figure 5.8: Comparison of Gini Index and analysis of Opportunity-Loss/Gain rate before and after the standardization process

uniform inequality and utility levels (especially for top-n-ranking with lower density), shows overall worse performance rates than the other fairness constraints. In this case, Need policy has been applied in order to fill an hypothetical gender gap. In this way, the generated ranking must comply with the demographic parity criterion by exhibiting an equal number of female and male individuals - the examined dataset presents gender as a binary variable. Results of the ranking system show that in all three policies the exposure levels of the types are generally lower than those shown by the ranking built on the original distribution. More specifically, they show that the deviations of the mean types' exposure from the exposure of the population mean are more similar and lower, a sign that the redistribution has acted on this value by reallocating the exposure among types. As shown in Figure 5.10, type exposure means display more uniform distributions in the ranking under Equality constraint. This result is crucial as it indicates that the ranking system under this specific fairness constraint is able to keep a high level of general utility and a low level of inequality by simultaneously satisfying the required fairness constraint on the exposure metric. The type exposure means of remaining fairness constraints confirm their non-optimality with these specific data. As discussed in Section 5.4, distributive justice theories, and more specifically EOp theories, have as a fundamental moral premise that the reward, i.e.,, the redistribution rate, must be susceptible only to the effort that individuals exert to achieve a result. Therefore, in the theoretical vision of these methods, individuals' merit in achieving a specific result is solely commensurate with the effort they have spent. Given this premise, the theories of distributive justice therefore assume a broader meaning in which equality is not the primary objective. In our model rewards are distributed in diverse ways according to the given policy (Section 5.5.2). The Equity policy

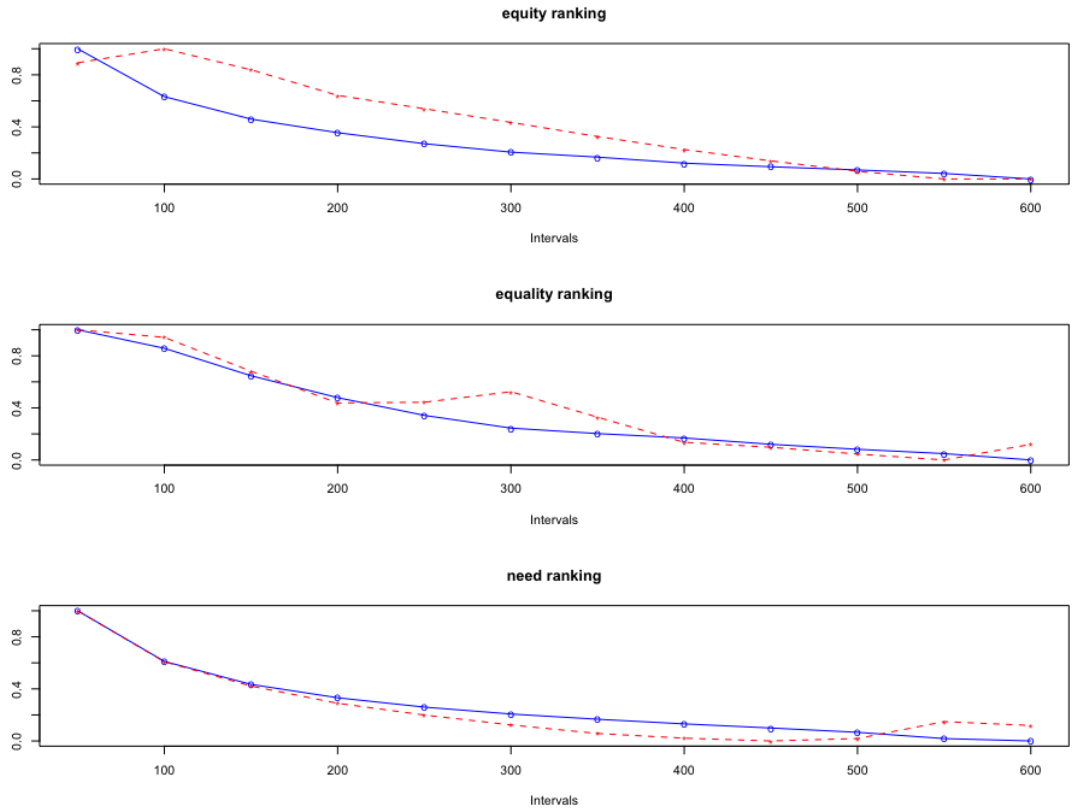
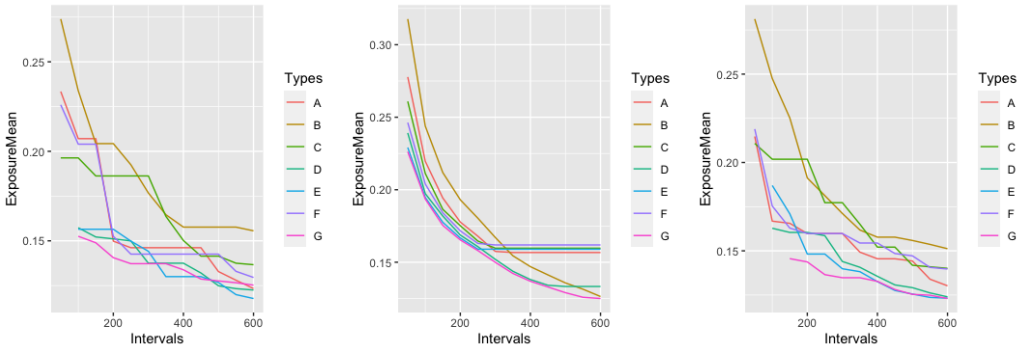


Figure 5.9: Inequality-Utility Trade-off for rankings under fairness constraints. Red line: inequality. Blue line: utility. Y-axis: utility and inequality values ranging from 0 to 1.

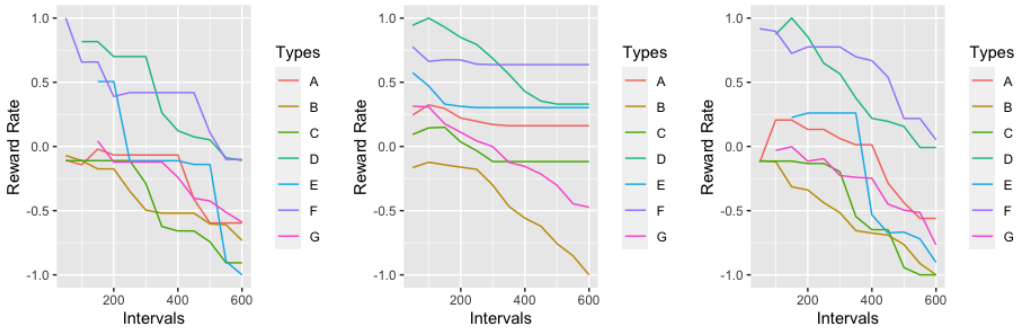


(a) Equity (b) Equality (c) Need

Figure 5.10: Comparison of type-exposure by ranking policy for all top-N-rankings

is inspired by the utilitarian-approach proposed by Fleurbaey 2013, and it aims to redistribute outcome among individuals by considering the contribution that each

effort-type vector give to the overall inequality. In this way, the subgroups outcome overestimation is controlled and balanced (Section 5.6.2). Note that the Equity policy is the sole one that balances this aspect through the realization of a counterfactual outcome; the reward is consequently calculated on the basis of this value. In Equality and Need policies, on the other hand, where no intermediate outcomes are produced, the reward is represented by a reallocation of the positions in the ranking according to the principle of demographic parity (Section 5.5.2). Since our ranking system provides for a reallocation of resources - in our case, the ranking position - we assess the disadvantages and benefits that each fairness constraint entails for types. We measure the Reward Rate that represents the extent and the magnitude of each distributive policy (Figure 5.11). The Reward Rate therefore represents the



(a) Equity (b) Equality (c) Need

Figure 5.11: Comparison of type-reward-rate by ranking policy for all top-N-rankings

loss or gain that each type has suffered as a result of the redistribution process but does not represent a metric of goodness of the policy. In other words, its negative or positive fluctuation is not an indicator of goodness, but rather an indicator of how much the groups receive based on their effort and how much they receive after reallocation. Since the results are not uniform for any of the metrics used along all the top-N-ranking but are subject to variation due to the density of the ranking, we develop a policy simulation scenario to understand which is the best constraint for those parameters.

Simulating a Fair-Distributive Decision-making Process with AFteRS

AFteRS provides rankings based on three principles of distributive justice. The output of the model is subject to fairness constraints that act differently according to the principle to which they are inspired. The fairness criteria that we implement are not better a-priori and there is no predominance of one over the other. This happens for two reasons: first, the system is closely linked to data to the extent that the greater the situation of inequality, the greater the magnitude of redistribution;

second, the policies we implement are not a-priori good for each context because they emphasize diverse dimensions of inequality removal. For example, one policy may emphasize individual fairness by keeping general utility levels lower than average, while another may emphasize the fairness of a given protected group while keeping a high level of general utility. For this reason, we build a policy simulation scenario with the goal of supporting the decision-maker in selecting which principle of distributive justice to apply in order to obtain a fair ranking.

Simulation Scenario and Results. We shape an hypothetical scenario in which a set of students compete to get access to the same university. The decision-maker can adopt a set of policies Θ to decide the best top-N-ranking. In its default version the simulation receives in input the number of individuals desired for the construction of the ranking and performs a max-min optimization of utility, inequality and exposure parameters, returning as output the ranking bound with the policy that performs the best parameters optimization. In the advanced version of the simulation, the decision-maker can instead select which aspect s/he prefer by emphasizing one of the three parameters listed above. The simulation performs a more relaxed version of the max-min optimization function, in order to satisfy the fairness constraints and queries simultaneously. Figure 5.12 summarizes the two simulation settings.

Given that the decision-maker can adopt one and only one ranking at a given

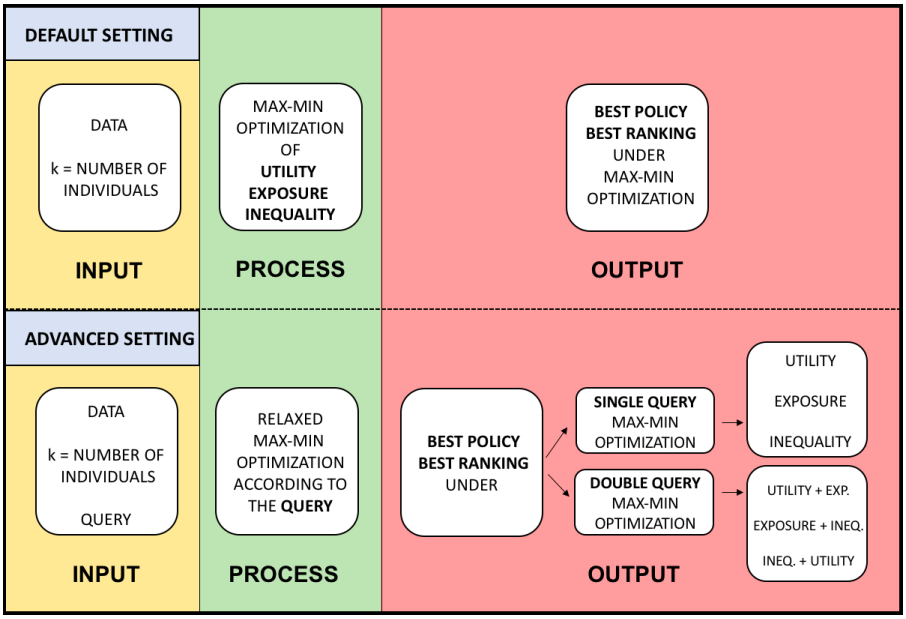


Figure 5.12: Simulations scenarios with AFteRS

time, we perform the experiment with different settings to simulate:

(i) The decision maker's choice to obtain an optimal ranking for all three parameters (default setting);

(ii) The decision maker's choice to maximize only one parameter (single query advanced setting);

(iii) The decision maker's choice to maximize two parameters (double query advanced setting).

Default Setting		**Advanced Setting**				
		Single query		Double query		
Policy	Overall	Utility	Inequality	Exposure	E/I	U/E	U/I
equity	0.53	0.88	0.12	0.34	0.50	0.88	0.66
equality	0.22	0.09	0.88	0.31	0.09	0.03	0.09
need	0.25	0.03	0.00	0.34	0.41	0.09	0.25

Table 5.9: Results of AFteRS simulations

Table 5.9 summarizes the main results of simulating a fair-distributive decision-making process with AFteRS. The simulation is performed for each setting and for all top-N-ranking at intervals of 20. The results shown in the table indicate the percentage of success of every policy for each setting and query type. Equity policy generally exhibits the highest success rate. However, as shown in Figure 5.9, the optimization values are not very far apart for the policies, so although the table displays a clear prevalence of the Equity policy, the difference in goodness from other policies is often very small.

5.7 Discussion, Relations to Related Work and Limitations

The present study was designed to determine the effect of distributive fairness on ranking systems. An interdisciplinary methodology was followed combining machine learning tools and Roemer's Equality of Opportunity theory. The major aim of this work was to design a ranking system based on ranking score re-allocation, in order to compensate individuals for unfair circumstances and to provide more equitable results. A case-study approach in university access was adopted to gain a detailed understanding of distributive justice theory applied to the ranking system domain and to evaluate the effectiveness of our distributive fairness ranking mechanisms. The most interesting finding was that our approach does not suffer fairness constraints and is able to preserve the system's utility by providing the best candidates to attend university on the basis of their actual score. The integration

of distributive justice theory into ranking systems, and more generally into automated decision-making systems, was successfully demonstrated by the evidence emerging from inequality-utility trade-offs, that with respect to previous studies have shown a greater transparency of the design's moral assumptions combined to the results of the ranking system (RQ1, Table 5.1). Our study has introduced an important novelty: fairness has not been merely treated as a debias operation, unlike in previous studies, but has redistributed the ranking score of individuals based on circumstances - features - and individual effort, thus compensating unfairness scenarios (RQ2, Table 5.1). Subsequently, three types of policies have been applied, corresponding to three sub-domains of distributive justice theory. The Equity policy has produced a counterfactual score that indicates the outcome individuals would have gotten if they had not belonged to their type. The Equality policy has acted according to the criterion of demographic parity, assigning equally to each sub-group of individuals classified by circumstances - attributes - the positions in the ranking. Finally, the Need policy has always acted according to the criteria of demographic parity with the difference that the re-distribution of the ranking positions has occurred on the basis of the gender sensitive attribute. The purpose of applying three different policies lies in the fact that, although sharing the same substratum of moral assumptions, different criteria produce different results in the model in terms of both fairness and utility. One interesting finding was that Equity policy offers the best results for numerous rankings, i.e., with more individuals, both in terms of utility and fairness, while on less dense rankings the Equality policy is the best choice. The Need policy, on the other hand, provides lower performance for all rankings in terms of both fairness and utility. These results have two important implications. First, the selection of the distributive justice sub-criterion, and more generally the selection of moral assumptions, is determined by the context. This means that the decision-maker must prefer the policy or criterion that offers the best fairness-utility trade-off, in order to simultaneously preserve the fairness and accuracy of the model. The design of our policy simulation scenario (5.6.2) is based on this assumption. Second, the demographic parity criterion traditionally adopted in supervised learning [14], despite being applied alongside distributive justice criteria, is outperformed in terms of performance by other sub-criteria, especially when based on a single sensitive attribute. Another important novelty, compared to other fairness studies both in the field of ranking and automatic decision-making systems, was the division of individuals into types on the basis of circumstances, which represent the combination of the attributes' realizations. This partitioning has allowed to overcome the traditional approach in the field of fairness that exclusively associates inequality to sensitive attributes, in favor of a broader view in which inequalities and individuals' merits are recognized on the basis of a broader assessment of their attributes. These results have shown how this division is actually more effective in achieving equality and are consistent with the preliminary analysis of individuals' circumstances (RQ3, Table 5.1).

This work is significantly different from those that aim to codify equality of opportunity in computer systems. In our model the outcome of individuals is readjusted through a series of procedures to produce a counterfactual outcome, i.e., the outcome that individuals would show if they had a different set of circumstances. This procedure implements a compensatory mechanism very different from the one used for example by Hardt *et al.* [83]. While in Hardt's definition the compensation is achieved during the post-processing phase, in our model the compensation takes place both in pre-processing and post-processing. The computation of a counterfactual outcome implies that our procedure uses all the individuals features to find similarities or differences in data in terms of Roemerian circumstances. The outcome is therefore redistributed on the basis of the circumstances-effort mechanism proposed by Roemer, by clearly differentiating it from the outcome redistribution methods already existing. In fact, the redistribution in our model is based on the opportunities of individuals consisting of a vector of attributes, rather than a single protected attribute as occurs in several other works (e.g., [83], [178]). In addition, this type of compensation differs from the methods of exposure redistribution in ranking systems because it re-establishes the relevance of the individuals on the basis of the counterfactual outcome. In this way the exposure is indirectly adjusted on the basis of the relevance of the individuals that has already been reassessed through a compensatory mechanism. While redistribution of exposure does not necessarily improves the average conditions and the relevance of a protected group, redistribution in our model improves the relevance of groups on the basis of the circumstances/effort variables and acts indirectly on the group' exposure, which may not necessarily be modified.

While the methodology of AFteRS is generalizable to any automatic task involving the attribution of a score to a group of individuals - as well as the designed metrics (5.6.1), the generalization of the results achieved in the analyzed dataset is subject to certain limitations. For instance, the specificity of the fairness-utility trade-off results is affected by data variation. Furthermore, since the study was limited to a labeled dataset, it was not possible to apply AFteRS on unsupervised tasks. Secondly, the major limitation of this study was the dependence on the specific dataset used. In fact, this type of approach works better when more socio-demographic characteristics of individuals are present in the dataset, since the theoretical assumption underlying the model is to compensate individuals for situations of iniquity through effort and circumstances. As a consequence, the outcome of AFteRS is based on observable assumptions only. There may be unobserved confounders within types, exogenous to the individual, that may influence the individual's effort: for example, differences within types in parental health-status or socio-economic background can have a strong impact on effort towards university applications. As a consequence, the model's outcome is influenced only by the circumstances actually observed in data, which is a further limitation.

Chapter 6

A Decision Support System for Long-term Fairness

6.1 Introduction

As the diffusion of artificial intelligence (AI) and Machine Learning (ML) systems spreads to a wide range of applications, algorithmic fairness has become a prominent open research issue [55], [83], [106]. With the rise of machine learning technologies, such as neural networks and deep learning, more and more government agencies are beginning to consider using these technologies to improve decision-making. As a result, many questions of an ethical nature such as fairness and justice of decisions have been shifted to a sphere of technical formalizations, which while rich cannot fully capture the variety of nuances belonging to the moral sphere [60], [140]. The fundamental part of the ethical-technical analysis of automatic decision-making systems is represented by ensuring non-discriminatory results for the population, above all for minorities and disadvantaged groups [18]. In particular, a crucial aspect is the evaluation of the impacts that these decisions will have on society [5]. Automated decision systems, especially data-driven machine learning systems, have received considerable attention in this respect in recent times [14]. Traditionally, machine learning systems consider the population at a fixed instant of time; however, the decisions of an automated decision system change the population and the way individuals approach institutions in a variety of ways over time [20], [181], [100], [145], [168], [92]. For instance, loan and mortgage decisions can change the profile of applicants over time and lead to a more unequal or fairer distribution of wealth; the use of automated systems to assess recidivism can affect societies' perceptions of institutions and of weaker segments of the population; school-education predictions can change the propensity to pursue higher education and the systems employed to assess educational institutions can determine the distribution of resources. Although these systems may therefore lead to unexpected impacts over time, the study of long-term effects is still an unexplored

field of research.

This study set out to investigate the impact of long-term fairness of automatic decision-making systems on the population by proposing a theoretical model of a Decision Support System. In the spirit of promoting fairer and more effective automated decision systems, the role of individual dynamics in automated decision-making is explored. We consider individual dynamics as a stage of the automated decision-making process. The overall purpose of this work is to illustrate a research strategy that examines the possibility that individual decision-making dynamics in response to a certain policy can affect the effectiveness of a policy itself. Our Decision Support System offer a new way to observe predictive policies by integrating individual dynamics in the model pipeline, which can be used to predict the policy effectiveness. As such, we provide a comprehensive picture - based on the individual response to a certain policy - and provide it as a model for developing and applying theoretical studies on individuals capability to make decisions and on the role of this phase in automated decision-making systems.

In this vein, a Decision Support System that aims to ensure long-term fairness is proposed. Our methodology extends decision theory to automated decision-making systems by introducing a theoretical model to apply fairness to a binary partition of the target population. Specifically, according to our model fairness is achieved in a long-term horizon if both the majority and minority group show an equal amount of fairness in time. We introduce the notion of positive behavior as a baseline of our theoretical model, assuming that in a set of similar policies the best policy is the one that induces equally both majority and minority group to perform a positive behavior, for instance, by improving the qualification profile. As a second constraint, the best policy is the one that induces the majority of individuals to perform a positive behavior. In order to offer a best understanding of our theoretical model, we set a simulation scenario of a university selection process, in which an institutional decision-maker has to select in a set of policies, the policy to be adopted in order to maximize the long-term selection.

6.2 Research Questions

The following research questions are currently partially answered by the scientific community, and hence our contribution aims to explored them here:

RQ1 How do decisions resulting from an automated decision making process affect the underlying population?

RQ2 Do the fairness constraints keep their validity for as long as they act?

RQ3 How do individual dynamics in the long run affect system decisions?

6.3 Related Work

A large and growing body of literature has investigated statistical discrimination and unfairness in machine learning domain. However, prior work examines fairness constraints in a static way [14]. Studies dealing with changes in the population belong for the majority to economic and game theory studies (e.g., [89], [105], [90]). Recently, a growing amount of literature have been focusing on the changing in individual dynamics caused by algorithmic decisions, which have been found strictly related [36], [73]. Starting from this assumption, a strand of works has started to study the effects of imposing some fairness constraints on groups' population. For instance, [105] and [120] propose two-stage models at one-step impact to study the effects of fairness constraints on the underlying population. Kannan *et al.* [99] study which rules have to be applied in college admissions and hiring in order to achieve Equality of Opportunity in a one step-model. However, only in recent years have studies directly started to address how to ensure fairness in a long-term horizon. D'Amour *et al.* [48] consider effects of fairness constraints through a simulation study of evolving system's dynamics, proposing a framework to fairness-focused simulation studies. Zhang *et al.* consider the evolution of qualification profile and impact of long-term fairness through Markov decision problem setting. A significant analysis and discussion on the subject was presented by Mouzannar *at al.* [136] and Liu *et al.* citeLiu:2020, where the qualification profiles at each time step are assumed to be known by the decision-maker, that receives an increase amount of utility based on the goodness of predictions (aka, fairness). Furthermore, there is a relatively small body of literature that is concerned with long-term impacts of decisions on qualification profiles by analyzing system dynamics [90], [136], [121], [204]. The current study constitutes an intersection of the aforementioned works, by investigating the impacts and effectiveness of policies in the long run through the study of individual dynamics in response to system decisions.

6.4 Problem Formulation

6.4.1 Society: Groups, Positive Behavior

We consider the case in which the total population is divided into two groups A and B, where g_A and $g_B : 1 - g_A$ constitutes the fractions of the population [1]. We assume $\theta \in \Theta$ is the set of the individual's attributes that determine the maximization of the decision-maker's utility. For instance, in a job recruitment

[1]Binary partitions of the population are also explored in other similar work to reduce the complexity the model (e.g, [120] and [136]). It refers to privileged and unprivileged subgroups of the population.

scenario $g_A(\theta)$ and $g_B(\theta)$ represent the average qualifications of the candidates per group. In this work we assume there already exists a set of fair policies $\phi \in \Phi$ previously selected that maximizes the decision-maker's utility, such that:

$$U_I = \sum g_A(\theta) \cdot g_B(\theta) \tag{6.1}$$

We study the case in which the decision maker has to select in a set of policies, which at time t_0 produce a similar effect, the policy to be adopted in order to maximize the long-term selection. To make the policy selection model effective, we additionally include the behavior of individuals in groups. The addition of this term serves to study individual dynamics over time and in relation to group membership, and to determine how they affect the long-term selection model. We therefore assume there exists a function of individual dynamics $\lambda : \Phi \to \mathbb{R}$ that leads individuals to assume a certain type of behavior $\lambda \in \Lambda$ in response to an institution's decision, i.e. the policy ϕ. To give an example of behavior in response to an institution's decision, consider the case of the university selection process. Depending on the results of the admission test, individuals may decide to try the test again the following year. The long-term decision maker's utility is thus determined as well by individual dynamics, which is a factor that generally does not receive particular attention in fairness and computer science ground. It is described by the following equation:

$$U_I = \sum \lambda_A g_A(\theta) \cdot \lambda_B g_B(\theta) \tag{6.2}$$

We assume that the decision-maker's utility is given by changes in the population in terms of qualification profiles and behavior in response to a policy. This constitutes a more consistent analysis of the real evolutionary conditions.

6.4.2 Individual Dynamics

Novel to this work, individual dynamics are a crucial part of the policy selection process. In fact, they determine the success or failure of a policy over time. An individual dynamic is the decision-making process - or problem - that individuals carry out, in our case in response to a policy [67]. It is defined by the sixfold $P(X, \Omega, F, f, D, \Pi)$, which are respectively the alternatives, the scenarios, the impacts, the utility function, the deciders and the function of the preferences [4]. Given the deciders' preferences among a set of possible impacts, the aim is to identify an $x*$ solution or a subset of $X*$ solutions among the possible alternatives X that the deciders consider satisfactory. In other words, the aim is to identify what kind of behavior is being promoted by a certain policy.

Alternatives X

X is the set of possible alternatives x, or admissible solutions, representing the events falling under the deciders' control, i.e. all the possible deciders' choices. An

example of possible alternative to the university selection ground is to apply again the following year if the first time the application failed. To avoid confusion, we refer to the institutional decision-maker as decision-maker, while the terms deciders, actors or individuals are used when referring to individual dynamics. The set of the alternatives is assumed to be described by a finite vector of n real numbers, such that $X \subseteq \mathbb{R}^n$. Representing the alternatives as a finite set is convenient, but partially restricting since it doesn't allow to consider the alternatives as a set of infinite possible choices. Since a large number of problems in machine-learning decision-making systems are coded as binary problems, for instance applying or not applying, we decide to restrict the set of alternatives to the finite case.

Scenarios Ω

Ω is the set of possible scenario or outcomes ω, representing the events beyond the deciders' control but having a substantial effect on the system. The set of the scenarios is assumed to be described by a finite vector of n real numbers, such that $\Omega \subseteq \mathbb{R}^n$. The scenarios embed the uncertainty on the system behavior in the temporal horizon affecting the decider. In the case of university selection, one source of uncertainty may be not knowing the qualifications of candidates who will want to apply in the future. As for alternatives, this formalization excludes scenarios with infinite dimensions.

Impacts F and Utility Function f

F is the set of possible impacts, representing all the relevant information necessary in order to make a decision. For instance, if a candidate after a first failed university application has to choose whether to reapply, s/he can decide whether to retry the application at the same university (alternative) knowing that it will depend on how many candidates apply each year (scenario), but what matters for the decision is the outcome of the application. That constitutes the impact and depends on both the alternatives and the scenarios. The set of the impacts is assumed to be described by a finite vector of n real numbers, such that $F \subseteq \mathbb{R}^n$. Every impact depend on both the alternatives x and the scenarios ω and is described by the utility function $f(x, \omega)$.

Deciders D and Preference Function Π

D is the set of deciders whereas $\Pi : D \to 2^{F \times F}$ is the function the associates to every deciders d a subset Π_d of ordered pairs of impacts. The function Π serves to establish an oriented relation among the pairs of impact, such that for every decider there exist a vector P_d describing the decider' preferences: $\forall d \in D \exists P_d$. For instance, given two possible alternatives for a candidate d - applying or not applying - leading to two distinct impacts f and f', where $(f, f') \in \Pi_d$, if the

decider prefers the first impact over the second the preference is represented by the following relationship: $f \preceq f'$. If the second impact is preferred, the representation becomes the following: $f' \preceq f$.

Assumption of Rationality In this work it is assumed that individuals act in conditions of rationality. This means that deciders are assumed to evaluate the best possible strategy to maximize their personal utility. Although there are theories pertaining to the economic field that identify the rationality of individuals as limited by various cognitive and contextual factors, it is beyond the scope of this work to address this debate. The assumption of rationality we assume perfectly fits the current definitions in decision theory, we point out that there exist other theories on this topic.

6.4.3 Policy selection

Dominance of Solutions

The selection of a policy is conditioned by changes in the population and the individuals' behavior. Specifically, we set that policy eligibility criteria require population to is induced to perform a positive behavior. To provide a better understanding of what we mean by positive behavior, we introduce an example that will be discussed in more detail in Section 6.5. In the case of university selection, a rejected candidate in the first round has a number of possible alternatives. In this case, the positive behavior is represented by a positive change in the candidate's qualification profile, which corresponds to the candidate's decision to improve their exam scores in order to pass the university selection the following year. We establish that the best policy is that which positively affects the population by inducing an improvement in the candidate's qualification profile at time $t + 1$. Given two behaviors γ_1 and γ_2, where γ_1 is the positive behavior, which in this example corresponds to an improvement in the qualification profile, we set the γ_1 solution as dominant on γ_2. This means that the policy to select is the one where γ_1 solution is dominant in the target population. The behaviors in the population are named as $\gamma \in \Gamma$, but a closer look reveals that they are closely related to the deciders' alternatives. For instance, an alternative x_1 is represented by the choice to re-apply by improving the qualification profile - positive behavior - while a second alternative x_2 is represented by re-applying without improvement - other type of behavior -. As we will see in Section 6.5, the deciders' alternatives do not necessarily fall in these two cases solely. The reason why we assign a different notation to behavior is that the qualification profile improvement - or any positive behavior - is the sole discriminating factor in policy selection. As a result, the behavior is limited to the binary form: the positive behavior γ_1 and any other behavior γ_2 that does not induce an improvement.

Maximizing Dominance

The criterion of the behavior dominance γ_1 is applied to groups g_A and g_B. This means that in order a policy is selected, it must induce a positive behavior in both groups.

$$max \int_0^\infty \gamma_1(g_A) + \gamma_1(g_B) \tag{6.3}$$

Through Equation 6.3 the policy that induces the maximum number of individuals to adopt a positive behavior is selected. The evolution of qualification profiles is evaluated in the total time span and not in each single time variation. The reason is that it is not possible to assume a linear trend of the behaviors. In addition, the trend over time is evaluated by the sum of the profiles and not by the speed of convergence of the curves. In this latter case it would be required to assume the functions as monotonically increasing, which would force the variation of behavior in groups by altering the individual dynamics.

Minimization of Dominance between Groups

The condition of dominance maximization does not guarantee by itself that dominance is similar in both groups. We therefore set an additional condition to the dominance that guarantees its minimization in both groups. Without this condition it is impossible to assess whether the contribution to the dominance of the groups is equal.

$$min \mid \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \mid \tag{6.4}$$

Equation 6.4 corresponds to selecting the policy that minimizes the dominance difference between the two groups, i.e. it chooses the policy where both groups have a similar rate of qualification profile improvement. This specification prevents the condition in which a policy is evaluated as better since it shows a higher dominance rate, although only one group performs a positive behavior. Ideally, a policy should have an equal rate of improvement in both groups:

$$\mid \int_0^\infty \gamma_1(g_A) - \gamma_1(g_B) \mid = 0, \tag{6.5}$$

but in reality it is hard for this to happen.

6.5 Example of Application

We study the case of university selection process in which the decision maker has to select in a set of policies, the policy to be adopted in order to maximize the long-term selection, i.e. to select the best candidates. As we define in Section 6.4, fairness constraints are shaped in the form of inducing a positive behavior in the population, that in this case means inducing an improvement of candidates qualification at time $t + 1$.

6.5.1 Data

For the purpose of this study, the following key data were synthesized from Goodman *et al.* [80]. The source contains a report on retaking SAT statistics. The data generated consist of 5 attributes, *GPA score, SAT score, GRE score, age and sex*. Each row represents an observation, i.e. a candidate. Data synthesis was implemented as follows: i) previous biases in the data were removed so that both privileged and unprivileged groups were equally represented for negative and positive scores; ii) statistics regarding scores inherent in qualifications were preserved, thus the synthesized data retained the same statistical properties regarding variables concerning qualification.

Stochasticity Reduction

For simplicity of study, we assume that candidates can only improve the SAT score. In this way, the policy selection is based on the ability of the policy to induce a positive behavior in the candidates so that they improve their SAT score. In order to reduce the stochasticity of individual dynamics, each candidate is assigned an average improvement score based on SAT retaking statistics [80]. In our application this value serves to establish the maximum score that each candidate could obtain if s/he decided to retake the SAT and to reduce the randomness that inevitably arises when unknown individual dynamics are encoded.

6.5.2 Policy

In this case, policies are represented by different classification algorithms that aim to predict students future performances on the basis of their past qualifications: Gradient Boosting Machine, Generalized Linear Model, k-Nearest Neighbour, Naive Bayes Classifier, Support Vector Machine.

6.5.3 Alternatives

Candidates are assumed to have a set of finite and discrete alternatives $x \in X$:

x_1 : applying with qualification;

x_2 : applying without qualification;

x_3 : do not apply.

Since the choice to improve the qualification profile and not to apply is an alternative that would not be directly observable in the model, only the choice of not applying is considered as an alternative. In addition, taking the option not applying with the qualification as a possible alternative, implies assuming that the

policy has an effect on a population that will no longer be observed by the model itself. Assumption as well as hardly susceptible to validation, also not consistent to reality. To facilitate the reader, henceforth we refer to the specific alternatives with the acronyms $x_1 = Q$, $x_2 = NQ$, $x_3 = N$, such that $X = [Q, NQ, N]$

6.5.4 Scenarios

Scenarios are assumed to be finite. In this case, the utility function f can be represented with an evaluation matrix, in which the rows are associated with alternatives x and the columns with the scenarios ω) (Table 6.1). In the case in

$f(x, \omega)$	ω_1	ω_2	ω_3
x_1	...	...	...
x_2	...	...	...
x_3	...	...	...

Table 6.1: Example of evaluation matrix with three alternatives and three scenarios

which the state of nature is influenced by the decision variables, the probabilities of the scenarios do not constitute anymore a vector of absolute values, but a matrix of values conditioned by the selected alternative.

Scenarios Types

In the case of selection university process we assume individual dynamics are computed with a set of three possible scenarios: optimistic, pessimistic and agnostic scenario. Scenarios represent the probability that the future set of applicants is constituted by individuals showing respectively an average SAT score equal to the mean SAT score of non accepted applicants, i.e., the worst candidates; an average SAT score equal to the mean SAT score of accepted applicants, i.e., the best candidates; an average SAT score equal to the mean SAT score of the overall previous set of applicants (time $t - 1$). As a consequence, the state of nature, i.e., the set of next generation applicants (time $t + 1$), is influenced by the decision variables, i.e., the qualification profile of the set of applicants at time $t + 1$. Hence, the probability of the optimistic, pessimistic and agnostic scenarios is constituted by a matrix of values $\pi(\omega|x)$ conditioned by the alternatives $x \in X$ (Section 6.5.3). Table 6.2 provides an example of the scenario probability matrix conditioned by the alternatives *Q, NQ, N* (Section 6.5.3), i.e., applying with improvement of SAT score, applying without improvement of SAT score, not applying. In Table 6.2,

$\pi(\omega\|x)$	Q	NQ	N
optimistic	0.75 | 0.86	0.75 | 0.76	0.75 | 0.76
pessimistic	0.85 | 0.86	0.85 | 0.76	0.85 | 0.76
agnostic	0.8 | 0.86	0.8 | 0.76	0.8 | 0.76

Table 6.2: Example of scenario probability matrix conditioned by alternatives Q, NQ, N

a scenario probability matrix for a candidate c having SAT score of 0.76 is computed[2]. The columns represent the alternatives of the candidate; in this case, the candidate could improve his/her SAT score by 0.1 points, so that if s/he decided to improve his/her qualification profile, the SAT score would be 0.86 (column Q). In the other two cases NQ and N the SAT score doesn't change as the candidate do not improve the qualification profile. The rows represent the possible scenario. In the *optimistic* scenario, candidates shown an average SAT score of 0.75, i.e., the mean SAT score of the worst candidates at the time the candidate c is rejected; in the *pessimistic* scenario, candidates shown an average SAT score of 0.85, i.e., the mean SAT score of the best candidates at the time the candidate c is rejected; in the *agnostic* scenario, candidates shown an average SAT score of 0.8, i.e., the mean SAT score of the overall set of candidates at the time the candidate c is rejected.

6.5.5 Impacts F and Utility Function f

The evaluation matrix is therefore constituted by the expected outcome derived from the scenario probability matrix (Table 6.2). The values in Table 6.3 represent

$f(x,\omega)$	Q	NQ	N
optimistic	0.9	1	-1
pessimistic	0.9	0	0
agnostic	0.9	0	0

Table 6.3: Example of evaluation matrix for candidate c (Section 6.5.4)

the expected outcome derived from the comparison of the mean of the candidates for each alternative in each scenario - expressed in a binary form, 1 positive outcome, i.e., passing the university selection, otherwise 0 - plus the cost of the decision. For the alternative Q, the utility of $f(Q,\omega)$ is computed by subtracting to the expected outcome 1 - the candidate in each if the three scenario would have a favorable outcome by comparing the improved SAT score with the average SAT score of

[2]The original SAT score attribute is ranged from 600 to 2400 points. To compute individual dynamics the score is normalized between 0 and 1.

applicants in each scenario - the cost of the SAT score improvement 0.1. The utility function $f(NQ, optimistic)$ is 1, as if the candidate c decides to apply without improving the SAT score, s/he would obtain a favorable outcome without extra costs. The value -1 of the utility function $f(N, optimistic)$ expresses the expected losses in case of wrong decision. In fact, in this scenario if the candidate c chose to not apply s/he would be taking the wrong decision as his/her average SAT score exceeds the average SAT score of the overall candidates (Table 6.2. The remainder of the utility functions assume value 0 since the candidate average SAT score is lower than the mean of the set of candidates in the respective scenarios. As can be noticed, utility functions can assume specific values based on the configurations of the alternatives x and the scenarios ω and the costs associated with the decision.

6.5.6 Preferences and Laplace Criterion

Once the evaluation matrix is derived the candidate preferences are optimized. This means that for each alternative the possible outcomes conditional on the scenarios and the costs of the decision are assessed. Since candidates do not have knowledge on the true state of nature, i.e., they do not know the real set of future candidates, but calculate their utilities based on a belief about the future state, the scenarios are considered equiprobables. For preferences optimization the Laplace criterion is applied (Equation 6.6):

$$max_{x \in X} Laplace(x) = max_{x \in X} \frac{\sum_{\omega \in \Omega} f(x, \omega)}{|\Omega|} \tag{6.6}$$

Lacking information on the scenarios' likelihood, for each candidate the Laplace criterion combines their impacts by applying the same weight to all of them. The

$f(x, \omega)$	Q	NQ	N
$optimistic$	0.9	1	-1
$pessimistic$	0.9	0	0
$agnostic$	0.9	0	0
$Laplace(x)$	0.9	0.33	-0.33

Table 6.4: Example of Laplace optimization on the evaluation matrix for candidate c

resulting preferences order is $Q \prec NQ \prec N$, meaning that candidate c maximizes his/her preferences through the choice of alternative $x_1 = Q$.

6.5.7 Time Treatment

Non-evolving Preferences

The several choices that constitute the preference relation are assumed to be time-independent, i.e., established once and for all. Although a decision often consists of several elementary choices carried out at different instants, employing a highly complex preference model would imply a high degree of complexity, which is undesirable since the purpose of the model is to support decisions. We therefore fix that the decider does not change his or her preferences over time to simplify preference modeling. This assumption implies that a candidate might decide to apply indefinitely, or continue to qualify over time. This variable's modeling might be different since preferences can change over time, e.g. a candidate might become discouraged after an unknown number of applications, but setting it up would involve making strong assumptions about the deciders' true propensities. Therefore it would not be possible if not through a parametrization according to an unknown function.

Data at time t+1

At each time t of the model, the data is updated. The new dataset is composed of two parts. One part is synthesized from the original data, meaning that a dataset with the same statistical characteristics at time t = 0 is reprocessed. The second part of the new dataset is the output of the model at time t - 1; this means that a portion of the data derived from applying individual dynamics is added to the synthesized data. The new dataset constitutes the input of the model at time t + 1 and contains the individuals who at time t - 1 chose the alternatives $x_1 = Q$ and $x_2 = NQ$.

Knowledge on the State of Nature

At each time instant t, the decider does not update the knowledge about the state of nature but uses the knowledge about the previous state of nature to optimize its outcome probabilistically. The reason why this knowledge is not used as a variable to determine the strategy is caused by the fact that otherwise the model would be deterministic. This is because the next state of nature is constituted by a set of data (scenario w_3). If the decider used this information to evaluate the strategy, s/he would not evaluate the other scenarios (w_1 and w_2), using it as deterministic knowledge even though s/he has no real knowledge of future events, i.e., s/he does not know what the future candidates will actually be.

Time and Policy Selection

As we define in Section 6.4.3, policies are evaluated through Dominance maximization (Equation 6.3) and Dominance minimization between groups (Equation 6.4). The Equations domain is adapted according to the needs of the study:

$$max \int_0^{10} \gamma_1(g_A) + \gamma_1(g_B) \tag{6.7}$$

$$min \mid \int_0^{10} \gamma_1(g_A) - \gamma_1(g_B) \mid \tag{6.8}$$

For this case we assume that the long-term fairness observation has a duration of 10 years (t = [0, 10]). In fact, it would not be efficient for this type of projection to have a study of longer duration, first because the forecasting tools as well as the available data may be subject to change over time; secondly, because we believe that a decision support system must be considered over a reasonable period of time to allow for its effectiveness, although a study of longer than 10 years is feasible.

6.5.8 System's Pipeline

Figure 6.1 schematically shows and summarizes the system pipeline. During the *preparation* phase, the five models (aka, *policies*) are trained with a starting synthetic dataset (aka, *original data*), tested on a portion of it and then evaluated for learning performance and audited for fairness. During *time t + 1* phase, the system receives as input from the *preparation* phase the learning models and the *original data*, which are synthesized to create a dataset with similar statistical properties (aka, *new data*). At each time step, the system *computes average improvement* (Section 6.5.1), makes predictions and figures out individual dynamics for each candidates. Once these operations are completed, the system updates the data. The *new data* at time *t + 1* will consist for a part of the data that have statistically similar characteristics to the *original data*, and for the remainder of the candidates that at time t chose the alternatives $x_1 = Q$ and $x_2 = NQ$, i.e., that decided to reapply.

6.6 Results

6.6.1 Classifiers Model Performance Evaluation

For the purposes of this study, five classification models were trained to represent the policies available to the decision maker. The models were trained on a portion of the original data (70%), and then used on new synthetic data with statistical characteristics similar to the original data. Models were validated via 10-fold cross-validation on the remaining portion of the original data (30%). Good classification

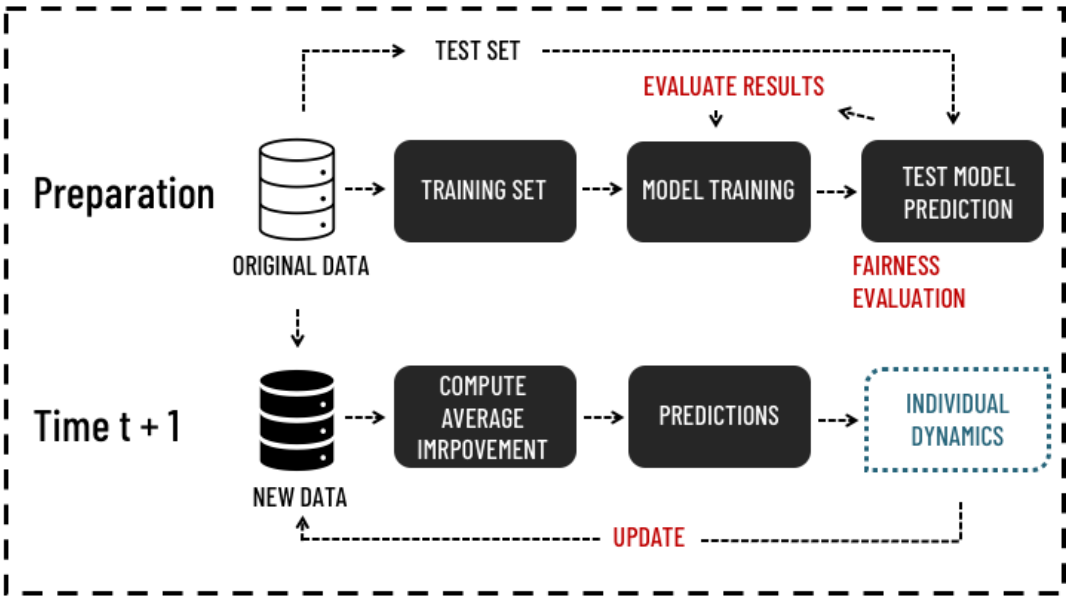


Figure 6.1: Overall System pipeline

performance was achieved for all cases. Figure 6.2 shows the confusion matrices and key performance metrics for each of the five models, and Figure 6.3 displays the ROC curves. As can be seen from the validation, the models show quite similar levels of performance. Therefore, at this step it is not useful or possible to pick or discard models a priori.

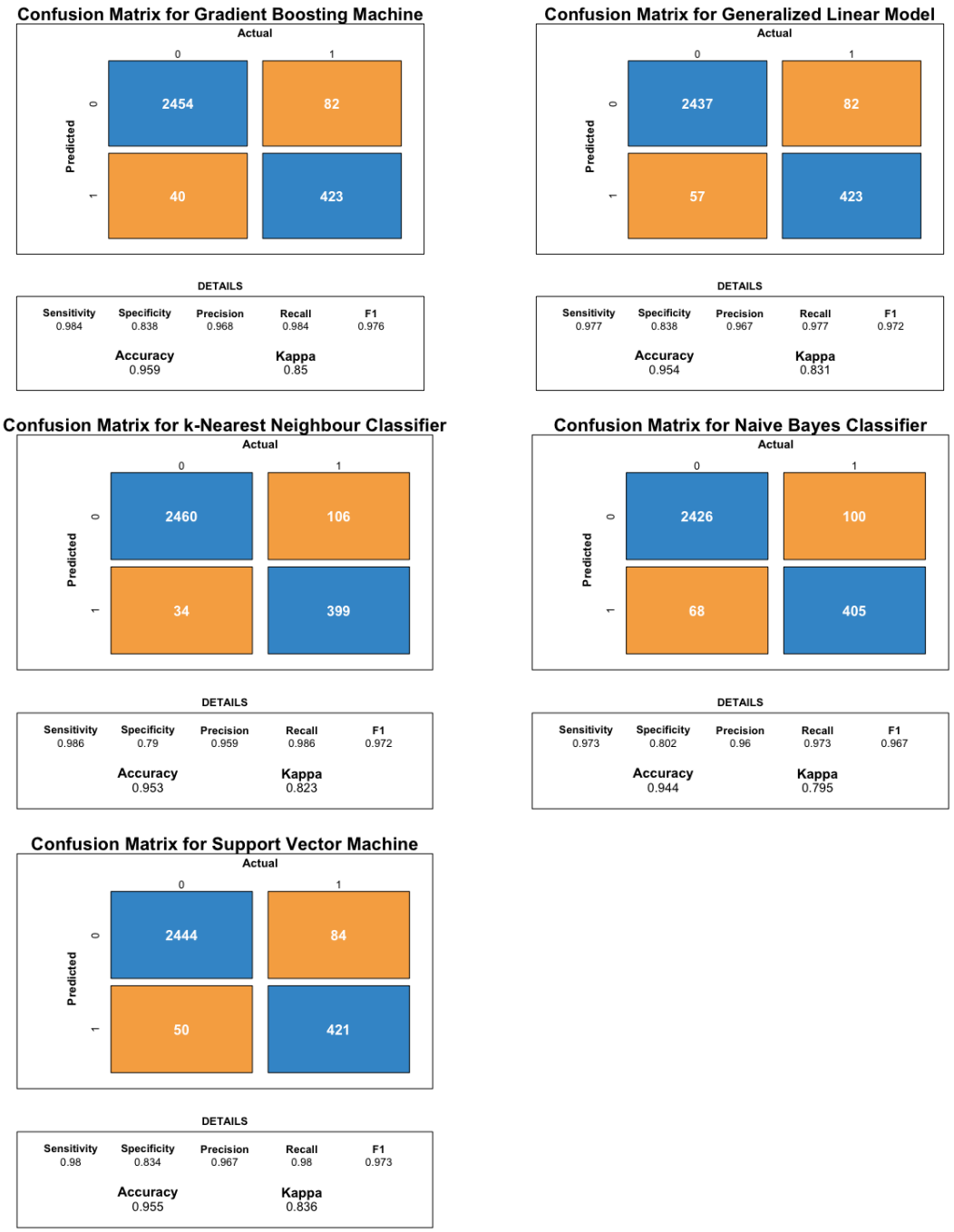


Figure 6.2: Confusion Matrices and Metrics

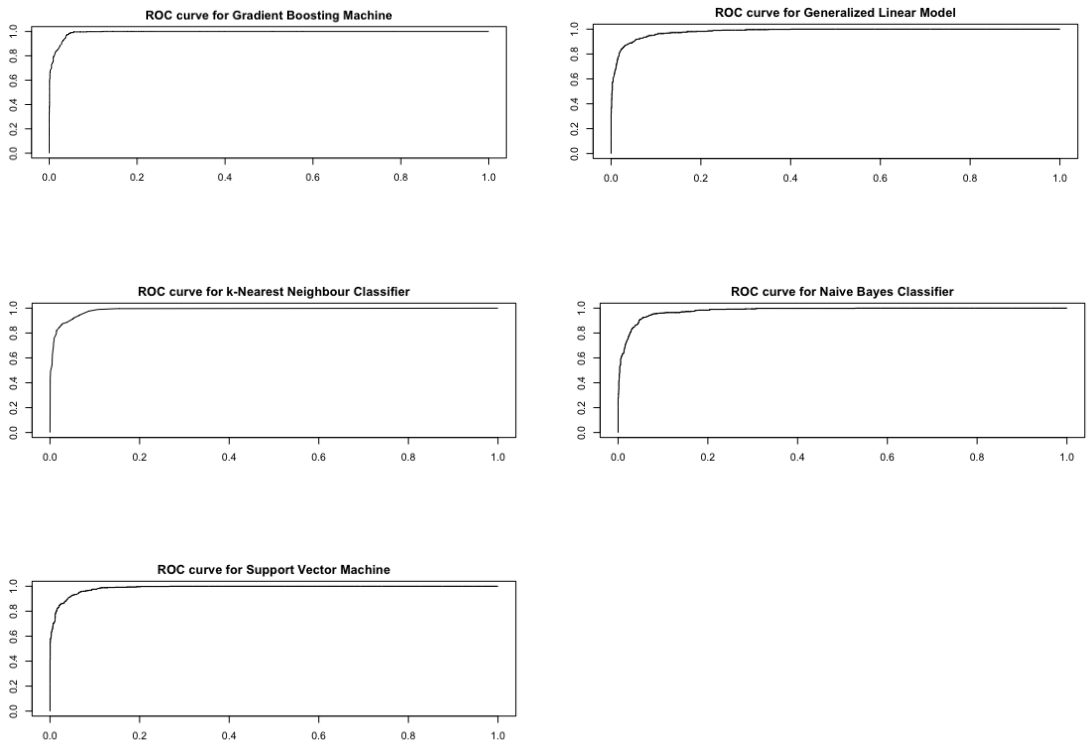


Figure 6.3: ROC curves. x axis: False Positive Rates. y axis: True Positive Rates

6.6.2 Fairness Evaluation

Once validated, the models undergo a second control phase represented by the fairness assessment. This step is necessary to ensure that the models do not generate discriminatory results for subgroups. In our case study, the protected attribute is represented by the binary variable *sex*; the fairness assessment is therefore performed in order to verify that each of the models presents an equal or at least similar degree of performance for each of the levels of the protected attribute. Table 6.5 displays the fairness results for the five models. The results are shown according to relative performance with respect to the *male* group. This means that the *sex* = *male* attribute was considered as the base group for computing the metrics. In case of parity, the metrics are equal to 1 in both groups, which means that the fairness in each group is the same as the base group. Parity greater than one indicates that the fairness in the observed group is relatively higher, while lower parity implies a lower level of fairness. Observing a large variation in fairness performance indicates that the model is not performing as well for the different sensitive groups. As can be seen from the Table 6.5, the overall fairness results across the five models show

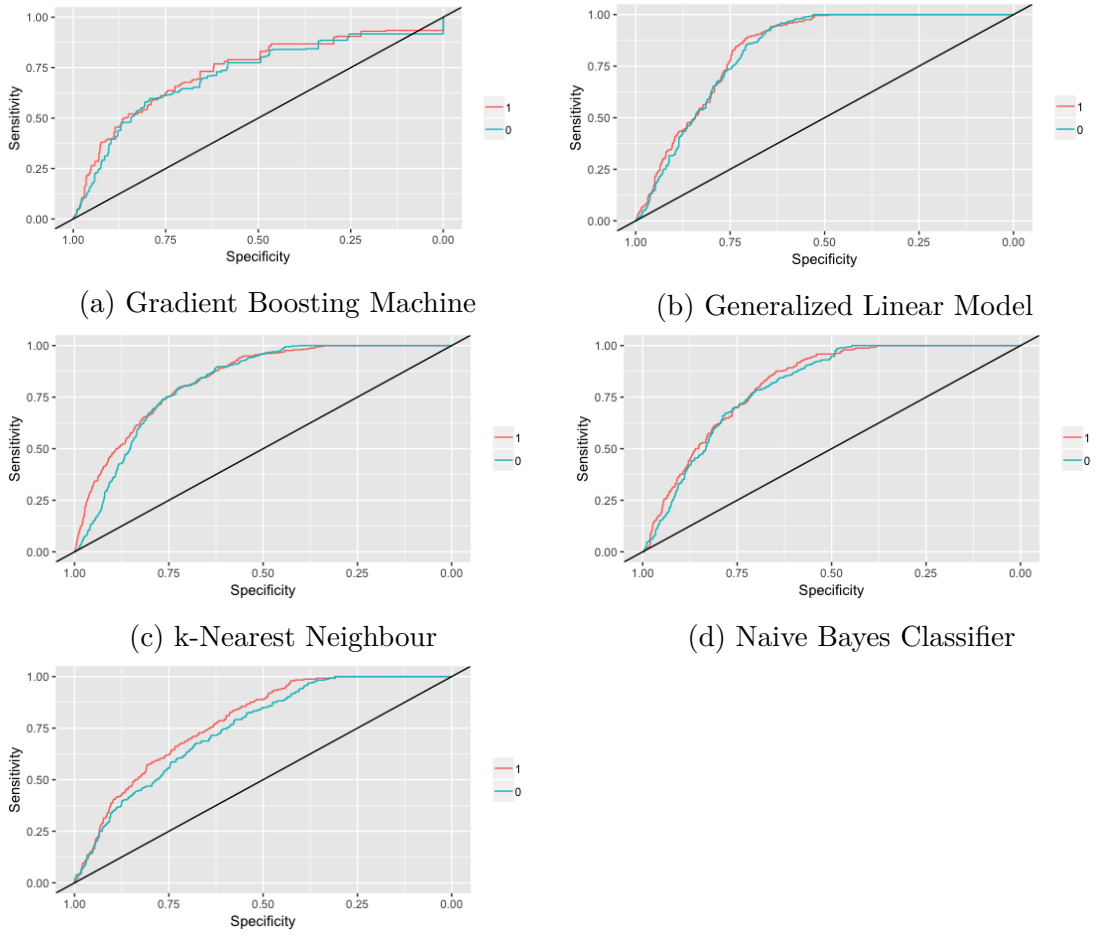
	Male (0)	Female (1)
Gradient Boosting Machine		
Sensitivity	1.00	1.00
Accuracy Parity	1.00	1.07
Predictive Rate Parity	1.00	1.07
k-Nearest Neighbour		
Sensitivity	0.00	0.00
Accuracy Parity	1.00	0.76
Predictive Rate Parity	1.00	0.78
Naive Bayes Classifier		
Sensitivity	1.00	1.00
Accuracy Parity	1.00	1.05
Predictive Rate Parity	1.00	1.02
Support Vector Machine		
Sensitivity	1.00	1.00
Accuracy Parity	1.00	1.06
Predictive Rate Parity	1.00	1.03
Generalized Linear Model		
Sensitivity	1.00	1.00
Accuracy Parity	1.00	1.04
Predictive Rate Parity	1.00	1.07

Table 6.5: Fairness results for main classification metrics

good performance. In general, there is a tendency for the *female* group to perform slightly better, but the difference should not be considered significant. In contrast, the k-Nearest Neighbour model shows a slightly lower level of fairness for the *female* group for both metrics, but the difference is not substantial in any case. Further details on fairness in the different models are illustrated in Figure 6.4, which shows the ROC curves for both groups. The results confirm what has already been highlighted by the metrics in Table 6.5, i.e., that fairness parity across groups is met for all five models.

6.6.3 Policy Selection and Individual Dynamics

As shown in section 6.4.3, the policy selection is conditioned by individual dynamics. In this case, the best policy is the one that induces the greatest number of individuals to perform a positive behavior, i.e., to improve their qualification in the SAT score, and that simultaneously minimizes the difference in qualification profiles across groups. This means that the best policy, in our case the classification algorithm, is the one that induces an almost similar improvement in qualification



(a) Gradient Boosting Machine

(b) Generalized Linear Model

(c) k-Nearest Neighbour

(d) Naive Bayes Classifier

(e) Support Vector Machine

Figure 6.4: ROC curves for groups. 0: Male Group; 1: Female Group

profiles in both *male* and *female* levels of the gender sensitive attribute. The first set of analyses examined the impact of long-term policies in the overall improvement of qualification profiles. Figure 6.5 compares the results obtained from the analysis of the profile qualification evolution in time $t = [0, 10]$ for each policy. It is apparent from this figure that a linear evolution of overall qualification profile in time doesn't exist. This result is in part due to the lack of information on actual possible improvements in individual qualification profiles, which as noted in Sections 6.5.1 and 6.5.4 are partially replaced by a stochastic process. Despite this, Figure 6.5 highlights a significant positive trend in increasing overall qualification profiles, an indication that fair classifiers can induce improvements in qualification profiles. The second set of analyses examined the impact of long-term policies in the improvement of qualification profiles in each group. Figure 6.6 visually displays the evolution in time of the qualification profiles for each policy. Since at

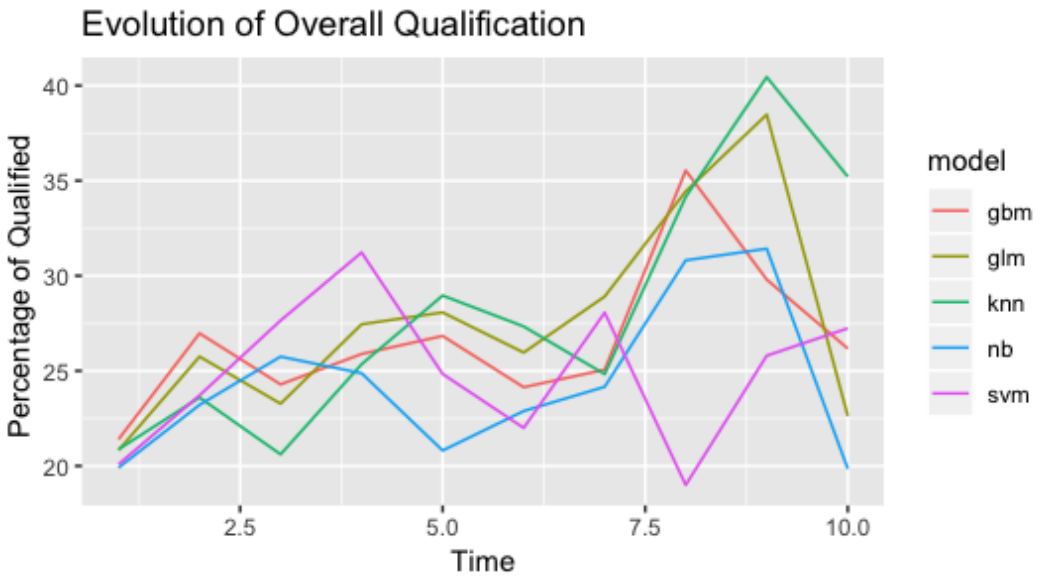


Figure 6.5: Dynamics of profile qualification evolution in time

each time instant t in the model the numerosity of the dataset differs from that at the previous time t, the amount of individuals choosing the $x_1 = Q$ alternative (Section 6.5.3) is shown as a percentage (Figure 6.5 and Figure 6.6). The results

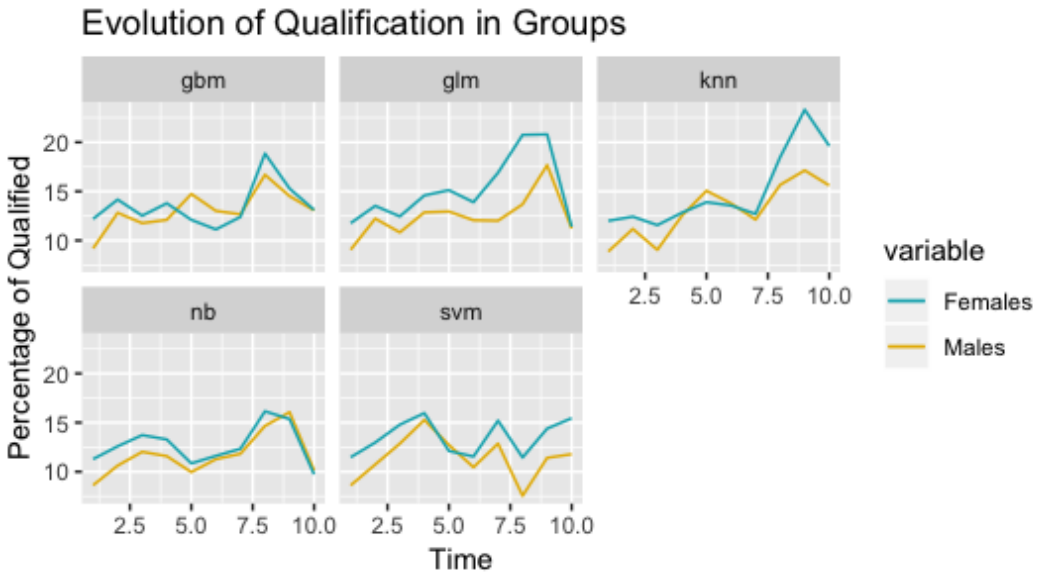


Figure 6.6: Dynamics of profile qualification evolution in time per Groups

in the figure confirm the positive trend shown in Figure 6.5 for overall improvement in qualification profiles and specify that the positive trend is accepted for

both groups. As a final step, the fairness criteria we establish in Section 6.4.3 are implemented. Table 6.6 illustrates fairness criteria developed in Equations 6.3 and 6.4; the column 1 provides the results of Dominance Maximization and the column 2 provides the results of Dominance Minimization between groups. The best

	$\int_0^{10} \gamma_1(g_A) + \gamma_1(g_B)$	$\int_0^{10} \gamma_1(g_A) - \gamma_1(g_B)$
Gradient Boosting Machine	242.31	3.48
k-Nearest Neighbour	253.33	15.83
Naive Bayes Classifier	223.82	9.02
Support Vector Machine	225.95	17.83
Generalized Linear Model	254.06	25.19

Table 6.6: Policy evaluation. First column: Dominance Maximization (Equation 6.3). Second column: Dominance Minimization between Groups (Equation 6.4).

results for the first fairness constraint are produced under the Generalized Linear Model. This means that this policy leads the greatest number of individuals to improve their SAT score comparing to the other policies. However, for the second fairness constraint the observed policy produces the worst results, showing a substantial difference in the evolution of qualification profiles between the two groups in comparison with the other policies. This means that this policy induces positive behavior predominantly in one of the two groups. The best results for the second fairness constraint are produced under the Gradient Boosting Machine. This result leads to two important considerations: first, for this case study both constraints cannot be simultaneously satisfied; second, in case of constraints' incompatibilities the policy that satisfies the Dominance Minimization constraint between the groups must be chosen. While in fact the first constraint is necessary to establish no involution of the qualification profiles, the second constraint is essential for achieving equal fairness in both groups, which is why a fair modeling is employed. Together these results provide important insights into long-term fairness, highlighting the study of individual dynamics contributes positively to the achievement of an enduring fair policy.

6.7 Discussion, Relations to Related Work and Limitations

The present study was designed to determine the effectiveness of a set of policies in the long run. Specifically, a Decision Support System to determine the effects of automated decision systems on the underlying population and achieving long-term fairness was supplied. An interdisciplinary methodology was followed combining machine learning models and Decision Theory. The major aim of this work was to

propose a theoretical formalization based on the evolution of individual dynamics, in order to assess the validity and the effectiveness of some fairness constraints on a long-time horizon. Two fairness constraints have been proposed to assess fairness in the long run. A case-study approach in university access was adopted to test the Decision Support System. The most relevant finding on the case-study was that our system is efficient in analyzing the long-term effects of policies by providing the evolution of qualification profiles for both groups in which the underlying population was partitioned (aka, males and females). Our theoretical formulation allows us to study how automated system decisions affect the population and groups; results on the case study indicate that learning models (aka, policy) although showing similar performance have different influences on groups in a non-one-step model (RQ1, Section 6.2). In fact, the qualification profile shows substantial differences across groups over the time period analyzed. These findings suggest important considerations. Over time, the qualification profile is being modified based on system decisions and on individual dynamics. This means that a time-dependent analysis highlights fairness is not consistent over time, revealing critical issues in applying static fairness constraints. In fact, if these same constraints were evaluated in a static model, the results would be different, lacking the insight to distinguish the goodness of models and criteria in the long term. In our case study, fairness at time $t = 0$ is almost similar for the five policies, thus making the preference of one model over another indistinguishable; in the long run, however, policies act differently on groups, in some cases clearly markedly different impacts on the population. As a result, fairness constraints do not necessarily keep their validity for as long as they act (RQ2, Section 6.2). As part of this result, individual dynamics assume a key role. In fact, they act as an individual component in a model that aims to achieve group fairness by shaping the individuals' response to implemented policies. In this sense, individual dynamics largely affect system outcomes, showing that the individual discernment component, albeit rationally modeled in our case-study, affects system decisions and the validity of fairness constraints in the long run (RQ3, Section 6.2).

Our study is close to some of the prior work in the existing literature with some key differences. For instance, Liu *et al.* [121] assume that individuals can observe current policy to make strategic decisions and change the qualification profile in order to improve their outcome. Although this study provides important insights, it lacks a study of equity constraints and how they are affected by individual dynamics. In contrast, in our model the study of individual dynamics underlies fairness constraints. Moreover, in our system individuals are not aware of the policy and the state of nature but act strategically based on the probability of receiving a positive outcome. In Mouzannar *et al.* [136] it is assumed that the evolution dynamics of the qualification profiles are the same in both groups and that the evolution is qualified by the policy and qualification profiles at the previous state of the system. Our system assumes that individual dynamics are agnostic to group membership;

in fact, we find it more realistic to assume that decision-making dynamics are more likely to be differentiated on an individual basis than on a group basis.

By imposing a selection constraint among several policies and selecting the best one in terms of fairness for both subgroups, our system is classifier type agnostic, i.e., it is applicable to both classifiers previously made fair and classifiers not subject to fairness constraints. In addition, the system can also be extended to non-binary partitions of the population. Although our theoretical formulation is fully generalizable, this aspect presents some limitations for the applicative setting. In fact, the results of the case study are case and data sensitive and cannot be extended in their specificity to other cases. Moreover, model validation is time sensitive and thus a study over a longer period of time could vary the fairness results obtained for our case study. However, we believe that only in part these limitations can be considered as such. In this domain, in fact, it is necessary to consider fairness as a case-sensitive constraint related to the study context.

Chapter 7

Conclusions and Future Directions

It is now widely acknowledged that algorithms reproduce and reinforce human prejudices that have historically led to discriminatory practices, especially against disadvantaged groups. Evidence of such discrimination has been collected and reported in numerous studies. As a result, deciding which standards of fairness and values should be embodied by algorithms poses significant ethical and political challenges to those responsible. Some solutions have been introduced to mitigate the impact of automated decisions systems that focus on metrics that measure algorithms equality or differing notions of fairness, such as gender equality. Algorithms should also explicitly encode certain values such as equity criteria. As a result, fairness requires a more broader planning with respect to the standard engineering process that meets some mathematical and statistical properties. A large number of fairness measurements in the literature are due to these efforts, although it may be mathematically impossible to simultaneously achieve different fairness measures except in limited special cases. As the focus has shifted from purely technical requirements to a multi-layered problem, choosing a fairness metric often involves deciding which models should have which conditions, and which conditions are to be borrowed from moral and political philosophy. Several recent studies have drawn attention to the importance of ethical considerations in measuring fairness in machine learning systems. These studies show that fairness should be seen as a trade-off process, with the priorities of the system as the backdrop. Indeed, since the beginning of the first machine learning fairness studies, the biggest challenge has been to define what fairness means.

In this manuscript, the issue of fairness and bias in automated data-driven decision systems has been addressed. First, a general overview of the context, functioning, and problem issues has been provided; second, analyses and case studies on fairness and bias in ADMs in specific application contexts have been supplied. The development of this thesis and the case studies was particularly inspired by the Research Problems highlighted in Section 1.1 and the advocated need to treat fairness in these systems as a multifaceted problem. Indeed, the study of fairness in

ADMs requires the use of specific programming, i.e., that metrics, constraints, and models be designed for the specific context. Obviously, because engineering human aspects means including in the modeling a complexity that is often not controllable and not necessarily known by the programmer, the development of ethical systems is still partly limited. As a result, fairness in automated systems stands in contrast to the generalization tendency of automated models, which aim to learn a pattern and reuse it, since they are guided by optimization principles. These considerations have driven the above manuscript and represent one of the probably most important challenges for ethical programming.

Main Contributions Firstly, we have studied and analyzed the role of rational actors in mainstream AI definition, which are currently driving the development of most AI systems. We underscore the importance of socially responsible AI actors by stating the limitations of the current definition of AI based on rational actors choosing their planned actions and generating expected benefits from the environment in which they operate. We give examples of the problem of distortions in the data used by AI, and we observe that measuring distortions in data sets is sufficient to identify potential risks of discrimination when data is used for rational AI agents. Using these examples, we present general ethical principles and discuss other open questions related to rational AI actors. This study adds a new perspective to the analysis of the common definitions of AI that have emerged from the work of many experts in the field of artificial intelligence (AI) research and development.

Secondly, we have developed a data annotation system that serves as a diagnostic framework containing immediate information about the data appropriateness, in order to more accurately assess the quality of the available data used in training models. The data annotation system follows a Bayesian statistical inference that aims to warn of the risk of discriminatory results of a given data set. In particular, the method aims to deepen the statistical knowledge related to the information contained in the available data, and to promote awareness of the sampling practices used to create the training set, highlighting that the probability of a discriminatory result is strongly influenced by the structure of the available data. This research is grounded in evidence showing that the process of data collection and the way data are recorded have a strong relation with ethics and transparency of data-driven systems. As a consequence, practices related to data collection, data transparency and data explainability become even more relevant and urgent. In fact, although the process of rigorous data collection and analysis is fundamental to the design of the model, this step is still largely overlooked by the machine learning community. The data annotation system has been tested on three different data sets that are well known in the fair machine learning community. In particular, the system test was particularly effective on the COMPAS dataset, used to predict the risk of recidivism in the American justice system. The dataset has been taken as a standard model by the fair machine learning community due to the presence of a high rate

of bias. Results have highlighted that in the COMPAS dataset the reoffending is related to ethnicity, and that success or failure are determined by the membership to a specific ethnic group. In general, the data annotation system has brought to light the risk of future bias in diverse magnitudes for all dataset tested. Furthermore, in the case of the COMPAS dataset it anticipated the underestimation of recidivism for the Caucasian ethnic group and the overestimation of recidivism for the Black ethnic group proven in recent studies.

Thirdly, a decision-making model to mitigate potential discriminatory effects of ranking systems has been implemented. We have proposed AFteRS, an Automated Fair-Distributive Ranking System, that has the objective of determining the best top-N-ranking in a set of candidates while simultaneously satisfying fairness constraints and preserving the general utility of the system. The approach takes inspiration from Roemer's Equality of Opportunity theory and from the distributive fairness notion that have been adopted as the basis for defining fairness and inequality. The ranking system implements three fairness criteria, each one based on a different dimension of the distributive justice theory, namely equity, equality, and need. Each fairness criterion provides diverse ranking results as well as different effects on individuals and groups of individuals. The system has been tested in an hypothetical scenario of a university selection process in which the decision-maker determines which students are suitable on the basis of their personal qualifications and achievements, so as to maximize the institution's utility. In such a context, we have examined the expected outcome for groups of individuals in the ranking system before and after the application of our distributive fairness approach, and we have explored the trade-off between the three different fairness policies in relation to the obtained rankings. Furthermore, a set of metrics to evaluate fairness in combination with traditional valuation metrics of ranking systems has been proposed. Results do not show an absolute predominance of one fairness criterion over another one, and that it is possible to achieve fairness constraints with a minimal impact on the general utility of the system.

Fourth and last, we have contributed to the literature in the long-term fairness domain by proposing a Decision Support System. We have studied the case in which the decision maker has to select in a set of policies, which at time $t = 0$ produce a similar effect, the policy to be adopted in order to maximize the long-term selection. To make the policy selection model effective, the behavior of individuals in groups has been additionally included. The addition of this term serves to study individual dynamics over time and in relation to group membership, and to determine how they affect the long-term selection model. Therefore, it has been assumed that there exists a function of individual dynamics that leads individuals to assume a certain type of behavior in response to an institution's decision, i.e. the policy. The long-term decision maker's utility is thus determined as well by individual dynamics, which is a factor that generally does not receive particular attention in fairness and computer science ground. The most relevant finding was that our system is

efficient in analyzing the long-term effects of policies by providing the evolution of qualification profiles for both groups in which the underlying population was partitioned (i.e. males and females). Our theoretical formulation allows us to study how algorithmic system decisions affect the population and the groups. Moreover, the results indicate that different learning models (i.e. policies) although showing similar performance have different influences on groups in a non-one-step model.

Challenges and New Trajectories for Future Research Data scientists and AI scholars agree that fairness and transparency are principles that need to be addressed as a matter of urgency. Unfortunately the road to fairness in machine learning and AI is littered with obstacles that are not all easy to overcome. The challenge for scholars and scientists to address this task requires defining certain statistical fairness qualities, properly normalizing various data sets, optimizing and testing algorithms to ensure fair (or at least fairer) results. This means keeping an eye on data types such as age, gender, age group, ethnicity, educational level and other relevant factors. With the list of course, some techniques will find value in the data when faced with a mix of demographic data, but not all [134]. As a consequence, several questions still remain to be answered. In the light of the issues raised by this manuscript, some insights and new trajectories for future research have been drawn:

NT1. Cross-Disciplinary Validation: due to the high interdisciplinary nature, validation is one of the most tricky aspects of research in this field of studies. In general, it has been observed that the combination of validation methods from different disciplines is the most appropriate choice for the evaluation of fair engineering systems. Despite the premises, this domain is still far from having a unified hybrid approach to validation. However, recently tools and programming libraries have been made open-source to verify and validate fairness and bias in machine learning systems [17, 71, 202, 108]. These tools are certainly a good start, but the road to fairness requires that cross-disciplinary validation methods be systematically incorporated into model pipelines;

NT2. Multi-High-Interpretability: in this specific domain, interpretability and explainability are two properties that are often confusingly interchanged [137]. While explainability is defined as the "*knowledge of what one node represents and how important it is to the model's performance*" [97], the interpretability is related to the model's ability to create a meaningful definition around the discovered relationships. It means that "*the cause and effect can be determined*" [97]. Models that show discriminatory results often exhibit a strong lack of the latter property, misinterpreting the causal relationship between two or more variables. Although this property is desirable for all machine learning and AI models, we invoke a greater focus on data-driven automated decision-making systems and the need to set relatively higher thresholds since

these systems often impact people's daily lives. In addition, an extra level of interpretability based on ethical criteria needs to be established in order for systems to be able to discern a discriminatory outcome;

NT3. Systematic Ground Encoding: as pointed out in several places in the manuscript, fairness is not only a technical problem but a multifaceted one that concerns the codification of moral principles. Automated decision-making systems should therefore systematically integrate ethical principles that should not be based on the perception of programmers but on a systematic modeling of fairness theories [82]. Otherwise, the problem continues to be buffered without a solid moral theoretical basis.

There are limits to how much we can and should trust automated decisions in human decision-making. All decision-makers should be aware of the issues involved and the decisions and assumptions developed by ML and AI scientists. To avoid perpetuating harmful prejudices against marginalized communities, algorithms designed to make decisions for others have a responsibility to identify bias points and rethink the standards used to determine when human decisions are fair and when they reflect problematic prejudices. Business and organizational leaders must ensure that their use of AI systems meets ethical standards, and they have a duty to promote research and standards that reduce prejudice in AI. Successful management of these problems will encourage the development of more efficient, effective and human-friendly machines.

Bibliography

[1] GDPR (EU) 2016/679. *General Data Protection Regulation*. URL: https://gdpr-info.eu/.

[2] J. Stacy Adams. "Inequity In Social Exchange". In: *Advances in Experimental Social Psychology*. Ed. by Leonard Berkowitz. Vol. 2. Academic Press, Jan. 1965, pp. 267–299. DOI: 10.1016/S0065-2601(08)60108-2. URL: http://www.sciencedirect.com/science/article/pii/S0065260108601082.

[3] Alekh Agarwal et al. *A Reductions Approach to Fair Classification*. 2018. eprint: arXiv:1803.02453[cs].

[4] George A. Akerlof and Janet L. Yellen. "Rational Models of Irrational Behavior". In: *The American Economic Review* 77.2 (1987), pp. 137–142. ISSN: 00028282. URL: http://www.jstor.org/stable/1805441.

[5] Aws Albarghouthi and Samuel Vinitsky. "Fairness-Aware Programming". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 211–219. ISBN: 9781450361255. DOI: 10.1145/3287560.3287588. URL: https://doi.org/10.1145/3287560.3287588.

[6] Marco Almada. "Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. ICAIL '19. Montreal, QC, Canada: Association for Computing Machinery, 2019, pp. 2–11. ISBN: 9781450367547. DOI: 10.1145/3322640.3326699. URL: https://doi.org/10.1145/3322640.3326699.

[7] Maureen L. Ambrose and Marshall Schminke. *The role of overall justice judgments in organizational justice research: a test of mediation*. en. Library Catalog: www.ncbi.nlm.nih.gov. 2009. DOI: 10.1037/a0013203. URL: https://www.ncbi.nlm.nih.gov/pubmed/19271803 (visited on 04/29/2020).

[8] Julia Angwin and Terry Jr. Parris. *Facebook Lets Advertisers Exclude Users by Race*. ProPublica. Oct. 2016. URL: https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race.

[9] Richard J. Arneson. "Equality and Equal Opportunity for Welfare". In: *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 56.1 (1989), pp. 77–93. ISSN: 00318116, 15730883. URL: http://www.jstor.org/stable/4320032.

[10] David Arnott and Graham Pervan. "A Critical Analysis of Decision Support Systems Research". In: *Journal of Information Technology* 20.2 (2005), pp. 67–87. DOI: 10.1057/palgrave.jit.2000035. eprint: https://doi.org/10.1057/palgrave.jit.2000035. URL: https://doi.org/10.1057/palgrave.jit.2000035.

[11] A. Asudeh, Z. Jin, and H. V. Jagadish. "Assessing and Remedying Coverage for a Given Dataset". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. New Jersey, US: IEEE, 2019, pp. 554–565.

[12] Abolfazl Asudeh et al. "Designing Fair Ranking Schemes". In: *Proceedings of the 2019 International Conference on Management of Data*. SIGMOD '19. Amsterdam, Netherlands: Association for Computing Machinery, 2019, pp. 1259–1276. ISBN: 9781450356435. DOI: 10.1145/3299869.3300079. URL: https://doi.org/10.1145/3299869.3300079.

[13] S. Barocas and A.D Selbst. "Big data's disparate impact". In: *California Law Reviews* 104.3 (2016), pp. 671–732.

[14] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2018.

[15] P.R. Bartlett et al. "Consumer Lending Discrimination in the FinTech Era". In: *UC Berkeley Public Law Research Paper* (2017). DOI: http://dx.doi.org/10.2139/ssrn.3063448.

[16] Marion Baylé. *Ethical dilemmas of AI: fairness, transparency, collaboration, trust, accountability morality*. 2019. URL: https://uxdesign.cc/ethical-dilemmas-of-ai-fairness-transparency-human-machine-collaboration-trust-accountability-1fe9fc0ffff3.

[17] Rachel K. E. Bellamy et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943.

[18] Ruha Benjamin. "Assessing risk, automating racism". In: *Science* 366.6464 (2019), pp. 421–422. ISSN: 0036-8075. DOI: 10.1126/science.aaz3873. eprint: https://science.sciencemag.org/content/366/6464/421.full.pdf. URL: https://science.sciencemag.org/content/366/6464/421.

[19] E. Beretta et al. "The Invisible Power of Fairness. How Machine Learning Shapes Democracy". In: *Advances in Artificial Intelligence, Proceedings of 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019*. Ed. by Marie-Jean Meurs and Frank Rudzicz. Vol. 11489. Kingston, ON, Canada: Springer, Cham, 2019, pp. 238–250. DOI: 10.1007/978$-$3$-$030$-$18305$-$9{_}19.

[20] Richard Berk et al. *Fairness in Criminal Justice Risk Assessments: The State of the Art*. Mar. 2018. DOI: 10.1177/0049124118782533. URL: https://doi.org/10.1177/0049124118782533.

[21] M. Berliant and W. Thomson. "On the fair division of a heterogeneous commodity". In: *Journal of Mathematical Economics* 21 (1992), pp. 201–216.

[22] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. "Equity of Attention: Amortizing Individual Fairness in Rankings". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: Association for Computing Machinery, 2018, pp. 405–414. ISBN: 9781450356572. DOI: 10.1145/3209978.3210063. URL: https://doi.org/10.1145/3209978.3210063.

[23] Reuben Binns. "Fairness in Machine Learning: Lessons from Political Philosophy". In: *Proceedings of Machine Learning Research*. Vol. 81. New York, NY, USA: Sorelle A. Friedler, Christo Wilsonf, 2018, pp. 149–159.

[24] Sarah Bird et al. "Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: Association for Computing Machinery, 2019, pp. 1297–1298. ISBN: 9781450366755. DOI: 10.1145/3308560.3320086. URL: https://doi.org/10.1145/3308560.3320086.

[25] NY Times Editorial Board. *The Race-Based Mortgage Penalty*. 2018. URL: https://www.nytimes.com/2018/03/07/opinion/mortage-minority-income.html.

[26] Carlo E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Florence, Italy: Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze, 1936.

[27] Owen Bowcott and Alex Hern. "Facebook and Cambridge Analytica face class action lawsuit". In: *The Guardian* (Apr. 2018). URL: https://www.theguardian.com/news/2018/apr/10/cambridge-analytica-and-facebook-face-class-action-lawsuit.

[28] Barry Brian. *Theories of Justice*. Oakland, California, USA: Berkeley: University of California Press, 1991.

[29] Maja Brkan. "AI-Supported Decision-Making under the General Data Protection Regulation". In: *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*. ICAIL '17. London, United Kingdom: Association for Computing Machinery, 2017, pp. 3–8. ISBN: 9781450348911. DOI: 10.1145/3086512.3086513. URL: https://doi.org/10.1145/3086512.3086513.

[30] Paolo Brunori and Guido Neidhöfer. "The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach". In: *SERIES Working Papers, N.01/2020* 1 (2020). DOI: 10.2139/ssrn.3520652. URL: https://ssrn.com/abstract=3520652.

[31] Robin Burke. *Multisided Fairness for Recommendation*. 2017. eprint: arXiv:1707.00093[cs]. URL: http://arxiv.org/abs/1707.00093.

[32] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. "Balanced Neighborhoods for Multi-sided Fairness in Recommendation". In: *Conference on Fairness, Accountability and Transparency in Proceedings of Machine Learning Research*. New York, NY: Proceedings of Machine Learning Research, 2018, pp. 202–214. URL: http://proceedings.mlr.press/v81/burke18a.html.

[33] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186. ISSN: 0036-8075. DOI: 10.1126/science.aal4230. eprint: https://science.sciencemag.org/content/356/6334/183.full.pdf. URL: https://science.sciencemag.org/content/356/6334/183.

[34] A. Campolo et al. *AI Now 2017 Report*. Tech. rep. New York, NY: AI Now Institute, 2017. URL: https://ainowinstitute.org/AI_Now_2017_Report.pdf.

[35] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. *Ranking with Fairness Constraints*. 2018. eprint: arXiv:1704.06840.

[36] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. RecSys '18. Vancouver, British Columbia, Canada: Association for Computing Machinery, 2018, pp. 224–232. ISBN: 9781450359016. DOI: 10.1145/3240323.3240370.

[37] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. "Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 2334–2346. ISBN: 9781450346559. DOI: 10.1145/3025453.3026044. URL: https://doi.org/10.1145/3025453.3026044.

[38] Daniele Checchi and Vito Peragine. "Inequality of opportunity in Italy". In: *Journal of Economic Inequality* 8.4 (2010), pp. 429–450. DOI: 10.1007/s10888-009-9118-3.

[39] Le Chen et al. "Investigating the Impact of Gender on Rank in Resume Search Engines". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–14. ISBN: 9781450356206. DOI: 10.1145/3173574.3174225. URL: https://doi.org/10.1145/3173574.3174225.

[40] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big Data* 5.2 (2017), pp. 153–163.

[41] G. A. Cohen. "On the Currency of Egalitarian Justice". In: *Ethics* 99.4 (1989), pp. 906–944. ISSN: 00141704, 1539297X. URL: http://www.jstor.org/stable/2381239.

[42] Ashley Colley et al. "The Geography of PokéMon GO: Beneficial and Problematic Effects on Places and Movement". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 1179–1192. ISBN: 9781450346559. DOI: 10.1145/3025453.3025495. URL: https://doi.org/10.1145/3025453.3025495.

[43] Jason A. Colquitt and Jessica B. Rodell. "Measuring Justice and Fairness". In: *The Oxford Handbook of Justice in the Workplace* 94.2 (July 2015), pp. 491–500. DOI: 10.1093/oxfordhb/9780199981410.013.8. URL: https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199981410.001.0001/oxfordhb-9780199981410-e-8.

[44] Sam Corbett-Davies et al. "Algorithmic decision making and the cost of fairness". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*. 2017.

[45] Paulo Cortez and Alice Silva. "Using Data Mining to Predict Secondary School Student Performance". In: *Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008)*. Porto, Portugal: A. Brito and J. Teixeira Eds., Apr. 2008, pp. 5–12. ISBN: 9789077381397. URL: https://archive.ics.uci.edu/ml/datasets/Student+Performance.

[46] Kate Crawford. "Artificial Intelligence's White Guy Problem". In: *New York Times* (June 2016). URL: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

[47] D. Dahiwade, G. Patle, and E. Meshram. *Designing Disease Prediction Model Using Machine Learning Approach*. 2019.

[48] Alexander DÁmour et al. "Fairness is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 525–534. ISBN: 9781450369367. DOI: 10.1145/3351095.3372878.

[49] Chris DeBrusk. *The Risk of Machine-Learning Bias (and How to Prevent It)*. 2018. URL: https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/.

[50] Center for Democracy and Technology. *AI Machine Learning*. 2020. URL: https://cdt.org/ai-machine-learning/.

[51] W. Dieterich, C. Mendoza, and T Brennan. *Compas risk scales: demonstrating accuracy equity and predictive parity*. Tech. rep. 2016.

[52] Agid – Agenzia per l'Italia Digitale. *Libro Bianco sull'Intelligenza Artificiale al servizio del cittadino*. 2017. URL: https://ia.italia.it/assets/librobianco.pdf.

[53] Finale Doshi-Velez et al. "Accountability of AI Under the Law: The Role of Explanation". In: *Berkman Center Research Publication Forthcoming. Harvard Public Law Working Paper* 18.07 (2017).

[54] Yanqing Duan, John S. Edwards, and Yogesh K. Dwivedi. "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda". In: *International Journal of Information Management* 48 (2019), pp. 63–71. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.01.021. URL: http://www.sciencedirect.com/science/article/pii/S0268401219300581.

[55] Cynthia Dwork et al. "Fairness through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts: Association for Computing Machinery, 2012, pp. 214–226. ISBN: 9781450311151. DOI: 10.1145/2090236.2090255. URL: https://doi.org/10.1145/2090236.2090255.

[56] Ronald Dworkin. "What is Equality? Part 2: Equality of Resources". In: *Philosophy and Public Affairs* 10 (1981), pp. 283–345.

[57] Benjamin Edelman, Michael Luca, and Dan Svirsky. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment". In: *American Economic Journal: Applied Economics* 9.2 (Apr. 2017), pp. 1–22. DOI: 10.1257/app.20160213. URL: https://www.aeaweb.org/articles?id=10.1257/app.20160213.

[58] Bora Edizel et al. "FaiRecSys: mitigating algorithmic bias in recommender systems". In: *International Journal of Data Science and Analytics* 9.2 (2020), pp. 197–213.

[59] Shady Elbassuoni, Sihem Amer-Yahia, and Ahmad Ghizzawi. "Fairness of Scoring in Online Job Marketplaces". In: *ACM/IMS Trans. Data Sci.* 1.4 (Nov. 2020). ISSN: 2691-1922. DOI: 10.1145/3402883. URL: https://doi.org/10.1145/3402883.

[60] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. USA: St. Martin's Press, Inc., 2018. ISBN: 1250074312.

[61] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. "Fairness in Relational Domains". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New Orleans, LA, USA: ACMP ress, 2018, pp. 108–114. DOI: 10.1145/3278721.3278733. URL: https://dl.acm.org/citation.cfm?id=3278733.

[62] Golnoosh Farnadi et al. *A Fairness-aware Hybrid Recommender System*. 2018. eprint: arXiv:1809.09030[cs,stat]. URL: http://arxiv.org/abs/1809.09030.

[63] Elaine Fehrman, Vincent Egan, and Evgeny M. Mirkes. *UCI Machine Learning Repository*. 2015. URL: http://archive.ics.uci.edu/ml.

[64] Elaine Fehrman et al. *The Five Factor Model of Personality and Evaluation of Drug Consumption Risk*. 2017. DOI: 10.1007/978-3-319-55723-6_18. URL: https://doi.org/10.1007/978-3-319-55723-6_18.

[65] Francisco H. G. Ferreira and Jérémie Gignoux. "The measurement of inequality of opportunity: theory and an application to Latin America". In: *Review of Income and Wealth* 57.4 (2011), pp. 622–657. DOI: 10.1111/j.1475-4991.2011.00467.x. URL: http://dx.doi.org/10.1111/j.1475-4991.2011.00467.x.

[66] Michèle Finck. "Automated Decision-Making and Administrative Law". In: *Oxford Handbook of Comparative Administrative Law* 19.10 (2019). Ed. by P. Cane et al. Forthcoming. URL: https://ssrn.com/abstract=3433684.

[67] Bruno de Finetti. "Probabilism: A Critical Essay on the Theory of Probability and on the Value of Science". In: *Erkenntnis (1975-)* 31.2/3 (1989), pp. 169–223. ISSN: 01650106, 15728420. URL: http://www.jstor.org/stable/20012237.

[68] Marc Fleurbaey and Vito Peragine. "Ex Ante Versus Ex Post Equality of Opportunity". In: *Economica* 80.317 (2013), pp. 118–130. DOI: 10.1111/j.1468-0335.2012.00941.x. URL: https://doi.org/10.1111/j.1468-0335.2012.00941.x.

[69] L. Floridi and M. Taddeo. "What is data ethics?" In: *The Royal Society* 374.2083 (2016).

[70] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. *On the (im)possibility of fairness*. 2016. eprint: arXiv:1609.07236[cs,stat]. URL: http://arxiv.org/abs/1609.07236.

[71] Sorelle A. Friedler et al. "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 329–338. ISBN: 9781450361255. DOI: 10.1145/3287560.3287589. URL: https://doi.org/10.1145/3287560.3287589.

[72] Daniel J. Fuchs. "The Dangers of Human-Like Bias in Machine-Learning Algorithms". In: *Missouri S&T's* 2.1 (Aug. 2018). URL: https://scholarsmine.mst.edu/peer2peer/vol2/iss1/1.

[73] Andreas Fuster et al. *Predictably Unequal? The Effects of Machine Learning on Credit Markets*. 2020. URL: https://ssrn.com/abstract=3072038%20or%20http://dx.doi.org/10.2139/ssrn.3072038.

[74] Joseph L. Gastwirth. "The Estimation of the Lorenz Curve and Gini Index". In: *The Review of Economics and Statistics* 54 (1972), pp. 306–316.

[75] Timnit Gebru et al. *Datasheets for Datasets*. 2018. eprint: arXiv:1803.09010.

[76] R. Stuart Geiger et al. "Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?" In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 325–336. ISBN: 9781450369367. DOI: 10.1145/3351095.3372862. URL: https://doi.org/10.1145/3351095.3372862.

[77] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-Aware Ranking in Search Recommendation Systems with Application to LinkedIn Talent Search". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2221–2231. ISBN: 9781450362016. DOI: 10.1145/3292500.3330691. URL: https://doi.org/10.1145/3292500.3330691.

[78] Yolanda Gil et al. "Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance". In: *Earth and Space Science* 3.10 (2016), pp. 388–415. DOI: 10.1002/2015EA000136. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015EA000136. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015EA000136.

[79] Corrado Gini. "Methods of Measuring the Concentration of Wealth". In: *The Economic Journal* 31.121 (1921), pp. 124–126.

[80] Joshua Goodman, Oded Gurantz, and Jonathan Smith. *Take Two! SAT Retaking and College Enrollment Gaps*. Working Paper 24945. National Bureau of Economic Research, Aug. 2018. DOI: 10.3386/w24945.

[81] Nina Grgic-Hlaca et al. "Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning". In: *32nd AAAI 2018*. New Orleans, LA, USA, 2018, pp. 51–60.

[82] Krishna P. Gummadi and Hoda Heidari. "Economic Theories of Distributive Justice for Fair Machine Learning". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. San Francisco, USA: ACM, 2019, pp. 1301–1302. ISBN: 978-1-4503-6675-5. DOI: 10.1145/3308560.3320101.

[83] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *Advances in Neural Information Processing Systems*. Barcelona, Spain, 2016.

[84] Jeff Harry Thornburg Larson, Surya Mattu, and Lauren Kirchner. *Machine Bias*. ProPublica. May 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[85] Hoda Heidari et al. "A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity". In: *FAT/ML, Proceedings of the Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA: ACM Press, 2019, pp. 181–190. DOI: 10.1145/3287560.3287584. URL: https://dl.acm.org/citation.cfm?id=3287584.

[86] Natali Helberger, Kari Karppinen, and Lucia D'Acunto. "Exposure diversity as a design principle for recommender systems". In: *Information, Communication & Society* 21.2 (2016). ISSN: 10.1080/1369118X.2016.1271900.

[87] Sarah Holland et al. "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards". In: *CoRR* abs/1805.03677 (2018). arXiv: 1805.03677. URL: http://arxiv.org/abs/1805.03677.

[88] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. "Unbiased Recursive Partitioning: A Conditional Inference Framework". In: *Journal of Computational and Graphical Statistics* 15.3 (2006), pp. 651–674.

[89] Lily Hu and Yiling Chen. "A Short-Term Intervention for Long-Term Fairness in the Labor Market". In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee, 2018, pp. 1389–1398. ISBN: 9781450356398. DOI: 10.1145/3178876.3186044.

[90] Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. "The Disparate Effects of Strategic Manipulation". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 259–268. ISBN: 9781450361255. DOI: 10.1145/3287560.3287597. URL: https://doi.org/10.1145/3287560.3287597.

[91] IEEE – Advancing Technology for Humanity –. *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (AI/AS)*. Tech. rep. The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, 2016. URL: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf.

[92] K. Indira and M. K. Kavithadevi. "Efficient Machine Learning Model for Movie Recommender Systems Using Multi-Cloud Environment". In: *obile Networks and Applications* 24.6 (2019), pp. 1872–1882. DOI: 10.1007/s11036-019-01387-4.

[93] Information and Communications Technology Council. *Artificial Intelligence in Canada. Where do we stand?* Tech. rep. Information and Communications Technology Council, 2016. URL: https://www.ictc-ctic.ca/wp-content/uploads/2015/06/AI-White-paper-final-English1.pdf.

[94] Inria. *Artificial Intelligence Current challenges and Inria's engagement*. Tech. rep. Inria, 2016. URL: https://www.inria.fr/en/news/news-from-inria/artificial-intelligence-current-challenges-and-inria-s-engagement.

[95] Kalervo Järvelin and Jaana Kekäläinen. "Cumulated Gain-Based Evaluation of IR Techniques". In: *ACM Trans. Inf. Syst.* 20.4 (Oct. 2002), pp. 422–446. ISSN: 1046-8188. DOI: 10.1145/582415.582418. URL: https://doi.org/10.1145/582415.582418.

[96] Eun Seo Jo and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 306–316. ISBN: 9781450369367. DOI: 10.1145/3351095.3372829. URL: https://doi.org/10.1145/3351095.3372829.

[97] Jonathan Johnson. *Interpretability vs Explainability: The Black Box of Machine Learning*. 2020. URL: https://www.bmc.com/blogs/machine-learning-interpretability-vs-explainability/%5C#:%5C~%5C:text=Interpretability%5C%20has%5C%20to%5C%20do%5C%20with,Nets%5C%2C%5C%20to%5C%20justify%5C%20the%5C%20results.%5C&text=Why%5C%20a%5C%20model%5C%20might%5C%20need%5C%20to%5C%20be%5C%20interpretable%5C%20and%5C%2For%5C%20explainable.

[98] Matthew Joseph et al. "Rawlsian Fairness for Machine Learning". In: *ArXiv* abs/1610.09559 (2016).

[99] Sampath Kannan, Aaron Roth, and Juba Ziani. "Downstream Effects of Affirmative Action". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 240–248. ISBN: 9781450361255. DOI: 10.1145/3287560.3287578. URL: https://doi.org/10.1145/3287560.3287578.

[100] Sumitkumar Kanoje, Debajyoti Mukhopadhyay, and Sheetal Girase. "User Profiling for University Recommender System Using Automatic Information Retrieval". In: *Procedia Computer Science*. Vol. 78. 1st International Conference on Information Security & Privacy 2015. 2016, pp. 5–12. DOI: https://doi.org/10.1016/j.procs.2016.02.002.

[101] N. S. A. Karim, F. A. Ammar, and R. Aziz. "Ethical Software: Integrating Code of Ethics into Software Development Life Cycle". In: *2017 International Conference on Computer and Applications (ICCA)*. Sept. 2017, pp. 290–298. DOI: 10.1109/COMAPP.2017.8079763.

[102] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, pp. 3819–3828. ISBN: 9781450331456. DOI: 10.1145/2702123.2702520. URL: https://doi.org/10.1145/2702123.2702520.

[103] D. Kaye. *Report of the Special Rapporteur on the promotion and the protection of the right to freedom of opinion and expression: Note by the Secretary-General*. United Nations, General Assembly. 2018. URL: https://freedex.org/wp-content/blogs.dir/2015/files/2018/10/AI-and-FOE-GA.pdf.

[104] Nicolas Kayser-Bril. *Google apologizes after its Vision AI produced racist results*. AlgorithmWatch. Apr. 2020. URL: https://algorithmwatch.org/en/story/google-vision-racism/.

[105] Moein Khajehnejad et al. *Optimal Decision Making Under Strategic Behavior*. 2019. URL: http://arxiv.org/abs/1905.09239.

[106] Jon Kleinberg. "Inherent Trade-Offs in Algorithmic Fairness". In: *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*. SIGMETRICS '18. Irvine, CA, USA: ACM Press, 2018, pp. 40–40. DOI: 10.1145/3219617.3219634. URL: http://doi.acm.org/10.1145/3219617.3219634.

[107] Ronny Kohavi and Barry Becker. *UCI Machine Learning Repository*. 1996. URL: http://archive.ics.uci.edu/ml.

[108] Nikita Kozodoi and Tibor V. Varga. *fairness: Algorithmic Fairness Metrics. R package version 1.2.0*. 2020. URL: https://github.com/kozodoi/fairness.

[109] Joe Kukura. *Facebook (Finally) Removes Racial Ad Targeting*. SFist. Aug. 2020. URL: https://sfist.com/2020/08/31/facebook-finally-removes-racial-ad-targeting/.

[110] Matt Kusner et al. "Counterfactual Fairness". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4069–4079. ISBN: 9781510860964.

[111] Ioan Doré Landau and Vlad Landau. *From data driven decision making (DDDM) to automated data driven model based decision making (MBDM)*. 2017. URL: https://hal.archives-ouvertes.fr/hal-01527766/document.

[112] J. Larson et al. *Compas Analysis*. 2016. URL: https://github.com/propublica/compas-analysis.

[113] James Larus and Chris Hankin. "Regulating Automated Decision Making". In: *Commun. ACM* 61.8 (July 2018), p. 5. ISSN: 0001-0782. DOI: 10.1145/3231715. URL: https://doi.org/10.1145/3231715.

[114] Alexandre Leblanc. "On estimating distribution functions using Bernstein polynomials". In: *Annals of the Institute of Statistical Mathematics* 64 (2012), pp. 919–943. DOI: 10.1007/s10463-011-0339-4.

[115] Jurek Leonhardt, Avishek Anand, and Megha Khosla. "User Fairness in Recommender Systems". In: *Companion Proceedings of the The Web Conference 2018*. Lyon, France, 2018, pp. 101–102. DOI: 10.1145/3184558.3186949. URL: http://arxiv.org/abs/1807.06349.

[116] Bruno Lepri et al. *The Tyranny of Data? The bright and dark sides of data-driven decision-making for social good*. Springer, Cham, 2017, pp. 3–24.

[117] Paolo Li Donni, Juan Gabriel Rodríguez, and Pedro Rosa Dias. "Empirical definition of social types in the analysis of inequality of opportunity: a latent classes approach". In: *Social Choice and Welfare* 44.3 (2015), pp. 673–701. DOI: 10.1007/s00355-014-0851-6. URL: https://doi.org/10.1007/s00355-014-0851-6.

[118] M. Lichman. *Default of credit card clients Data Set*. UCI Machine Learning Repository, University of California, School of Information and Computer Science (distributor). 2013. URL: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

[119] Patrick Lin. "AI Decisions, Risk, and Ethics: Beyond Value Alignment". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '18. New Orleans, LA, USA: Association for Computing Machinery, 2018, p. 2. ISBN: 9781450360128. DOI: 10.1145/3278721.3278806. URL: https://doi.org/10.1145/3278721.3278806.

[120] Lydia T. Liu et al. "Delayed Impact of Fair Machine Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, July 2018, pp. 3150–3158.

[121] Lydia T. Liu et al. "The Disparate Equilibria of Algorithmic Decision Making When Individuals Invest Rationally". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 381–391. ISBN: 9781450369367. DOI: 10.1145/3351095.3372861. URL: https://doi.org/10.1145/3351095.3372861.

[122] Max O. Lorenz. "Methods of Measuring the Concentration of Wealth". In: *Publications of the American Statistical Association* 9.70 (1905), pp. 209–219.

[123] K. Lum and W.S Isaac. "To predict and serve?" In: *Significance* 13.5 (2016), pp. 14–19. DOI: 10.1111/j.1740-9713.2016.00960.x. URL: https://doi.org/10.1111/j.1740-9713.2016.00960.x.

[124] Patrizia Luongo. "Chapter 2 The Implication of Partial Observability of Circumstances on the Measurement of IOp". In: *Gabriel Rodríguez, J. (Ed.) Inequality of Opportunity: Theory and Measurement (Research on Economic Inequality)* 19 (2011), pp. 23–49.

[125] David Madras, Toni Pitassi, and Richard Zemel. "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio et al. Curran Associates, Inc., 2018, pp. 6147–6157.

[126] Gianclaudio Malgieri. "Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations". In: *Computer Law Security Review* 35.5 (2019), p. 105327. ISSN: 0267-3649. DOI: https://doi.org/10.1016/j.clsr.2019.05.002. URL: http://www.sciencedirect.com/science/article/pii/S0267364918303753.

[127] Vidushi Marda and Shivangi Narayan. "Data in New Delhi's Predictive Policing System". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 317–324. ISBN: 9781450369367. DOI: 10.1145/3351095.3372865. URL: https://doi.org/10.1145/3351095.3372865.

[128] Ramon Margalef. *Perspectives in Ecological Theory*. Chicago: Chicago University Press, 1968.

[129] Daniel McDuff, Roger Cheng, and Ashish Kapoor. *Identifying Bias in AI using Simulation*. 2018. eprint: arXiv:1810.00471.

[130] Andrew McNamara, Justin Smith, and Emerson Murphy-Hill. "Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?" In: *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2018. Lake Buena Vista, FL, USA: Association for Computing Machinery, 2018, pp. 729–733. ISBN: 9781450355735. DOI: 10.1145/3236024.3264833. URL: https://doi.org/10.1145/3236024.3264833.

[131] Ninareh Mehrabi et al. *A Survey on Bias and Fairness in Machine Learning*. 2019. eprint: https://arxiv.org/abs/1908.09635.

[132] Marcelo Mendoza and Nicolás Torres. "Evaluating content novelty in recommender systems". In: *Journal of Intelligent Information Systems* 54 (2019), pp. 297–316. DOI: 10.1007/s10844-019-00548-x.

[133] J. Metcalf, E.F. Keller, and D. Boyd. *Perspectives on Big Data, Ethics, and Society*. Tech. rep. The Council for Big Data, Ethics, and Society, 2016. URL: https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/.

[134] Florian Meyer. *Everything AI?* 2020. URL: https://techxplore.com/news/2020-09-ai.html.

[135] Steven Mintz. *Ethical AI is Built On Transparency, Accountability and Trust*. 2020. URL: https://www.corporatecomplianceinsights.com/ethical-use-artificial-intelligence/.

[136] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. "From Fair Decision Making To Social Equality". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 359–368. ISBN: 9781450361255. DOI: 10.1145/3287560.3287599.

[137] Rym Nassih and Abdelaziz Berrado. "State of the Art of Fairness, Interpretability and Explainability in Machine Learning: Case of PRIM". In: *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*. New York, NY, USA: Association for Computing Machinery, 2020. ISBN: 9781450377331. URL: https://doi.org/10.1145/3419604.3419776.

[138] Central Banking Newsdesk. *Machine learning algorithms could increase ethnic bias – research*. 2020. URL: https://www.centralbanking.com/central-banks/economics/7693006/machine-learning-algorithms-could-increase-ethnic-bias-research.

[139] X. Ning and G. Karypis. "SLIM: Sparse Linear Methods for Top-N Recommender Systems". In: *2011 IEEE 11th International Conference on Data Mining*. 2011, pp. 497–506. DOI: 10.1109/ICDM.2011.134.

[140] S. U. Noble. *Algorithms of oppression: How search engines reinforce racism*. New York, NY, USA: NYU Press, 2018.

[141] Christine Nothelfer, Michael Gleicher, and Steven Franconeri. "Redundant encoding strengthens segmentation and grouping in visual displays of data". In: *Journal of experimental psychology. Human perception and performance* 43.9 (2017), pp. 1667–1676. DOI: 10.1037/xhp0000314. URL: https://visualthinking.psych.northwestern.edu/projects/papers/Nothelfer,%5C%20Gleicher,%5C%20&%5C%20Franconeri%5C%20-%5C%20Redundant%5C%20Encoding.pdf.

[142] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group, 2016.

[143] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. ISSN: 0036-8075. DOI: 10.1126/science.aax2342. eprint: https://science.sciencemag.org/content/366/6464/447.full.pdf. URL: https://science.sciencemag.org/content/366/6464/447.

[144] Information Commissioner's Office. *Rights related to automated decision making including profiling*. 2020. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/rights-related-to-automated-decision-making-including-profiling/.

[145] Oladapo Oyebode and Rita Orji. *A hybrid recommender system for product sales in a banking environment*. 2020. DOI: 10.1007/s42786-019-00014-w.

[146] Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. USA: Harvard University Press, 2015. ISBN: 0674368274.

[147] G. Phillips-Wren and L. Jain. "Artificial Intelligence for Decision Making". In: *In: Gabrys B., Howlett R.J., Jain L.C. (eds) Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science*. Vol. 4252. KES 2006. Berlin, Heidelberg: Springer, 2006. DOI: 10.1007/11893004_69. URL: https://doi.org/10.1007/11893004_69.

[148] Jill Quadagno. "Theories of the Welfare State". In: *Annual Review of Sociology* 13.1 (1987), pp. 109–128. DOI: 10.1146/annurev.so.13.080187.000545.

[149] Di Cosmo R. and Zacchiroli S. "Software Heritage: Why and How to Preserve Software Source Code Inproceedings". In: *Proceedings of 14th International Conference on Digital Preservation*. iPRES 2017. Kyoto, Japan, 2017. URL: https://hal.archives-ouvertes.fr/hal-01590958.

[150] Manish Raghavan et al. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 469–481. ISBN: 9781450369367. DOI: 10.1145/3351095.3372828. URL: https://doi.org/10.1145/3351095.3372828.

[151] Foyzur Rahman et al. "Sample Size vs. Bias in Defect Prediction". In: *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*. ESEC/FSE 2013. Saint Petersburg, Russia: Association for Computing Machinery, 2013, pp. 147–157. ISBN: 9781450322379. DOI: 10.1145/2491411.2491418. URL: https://doi.org/10.1145/2491411.2491418.

[152] Inioluwa Deborah Raji et al. "Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 145–151. ISBN: 9781450371100. DOI: 10.1145/3375627.3375820. URL: https://doi.org/10.1145/3375627.3375820.

[153] Xavier Ramos and Dirk Van de Gaer. *Empirical approaches to inequality of opportunity: principles, measures, and evidence*. Available at SSRN: https://ssrn.com/abstract=2096802. 2012. eprint: ZADiscussionPaperNo.6672.

[154] F. Raso et al. *Artificial Intelligence and Human Rights: Opportunities Risks*. Berkman Klein Center Research. 2018. URL: http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439.

[155] John Rawls. *A theory of justice*. Harvard, US: Harvard University Press, 1971.

[156] John Rawls. *Justice as Fairness: A Restatement*. Cambridge, US: Harvard University Press, 2001.

[157] John Rawls. *The idea of public reason*. Cambridge, US: The MIT Press, 1997.

[158] L. Rice and D. Swesnik. "Discriminatory Effects of Credit Scoring on Communities of Color". In: *Suffolk University Law Review* 46 (2013), pp. 935–966.

[159] John E. Roemer. "A Pragmatic Theory of Responsibility for the Egalitarian Planner". In: *Philosophy and Public Affairs* 22.2 (1993), pp. 146–166.

[160] John E. Roemer and Alain Trannoy. "Equality of Opportunity". In: *Handbook of Income Distribution* 2.2 (2015), pp. 217–300.

[161] Sheldon M Ross. *Stochastic processes*. Wiley series in probability and statistics: Probability and statistics. New Jersey, US: Wiley, 1996. ISBN: 9780471120629. URL: https://books.google.de/books?id=ImUPAQAAMAAJ.

[162] S. Roy. *Investments in cognitive and AI will reach $19.1b in 2018*. 2018. URL: https://techwireasia.com/2018/03/investments-cognitive-ai-will-reach-19-1bn-2018/.

[163] Minna Ruckenstein and Julia Velkova. *Automating Society*. 2019. URL: https://algorithmwatch.org/en/automating-society-finland/.

[164] Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia: Pearson Education Limited, 2020. URL: http://aima.cs.berkeley.edu/.

[165] Diana Sancho. "Automated Decision-Making under Article 22 GDPR: Towards a More Substantial Regime for Solely Automated Decision-Making". In: *Algorithms and Law*. Ed. by Martin Ebers and SusanaEditors Navas. Cambridge University Press, 2020, pp. 136–156. DOI: 10.1017/9781108347846.005.

[166] Vinod Saratchandran. *6 Ways Artificial Intelligence Is Driving Decision Making*. 2019. URL: https://www.fingent.com/blog/6-ways-artificial-intelligence-is-driving-decision-making/.

[167] Teresa Scantamburlo, Andrew Charlesworth, and Nello Scantamburlo. "Machine Decisions and Human Consequences". In: *Algorithmic Regulation*. Ed. by Karen Yeung and Martin Lodge. Oxford: Oxford Scholarship Online, 2019. Chap. 3. ISBN: 9780198838494. DOI: 10.1093/oso/9780198838494.003.0003.

[168] Markus Schedl et al. “Current challenges and visions in music recommender systems research”. In: *International Journal of Multimedia Information Retrieval* 7.2 (2018), pp. 95–116. DOI: 10.1007/s13735-018-0154-2.

[169] Silvia Schiaffino and Analıa Amandi. “Intelligent User Profiling”. In: *Artificial Intelligence: An International Perspective*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 193–216. ISBN: 3642032257.

[170] National Science and Technology Council – NSTC –. *Preparing for the Future of Artificial Intelligence*. Tech. rep. Executive Office of the President National Science and Technology Council Committee on Technology, 2016. URL: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

[171] Andrew D. Selbst et al. “Fairness and Abstraction in Sociotechnical Systems”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* ’19. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 59–68. ISBN: 9781450361255. DOI: 10.1145/3287560.3287598. URL: https://doi.org/10.1145/3287560.3287598.

[172] Amartya Sen. “Rights and Capabilities”. In: *Resources, Values and Development*. Cambridge, US: Harvard University Press, 1997, pp. 307–324.

[173] Claude E. Shannon. “A mathematical theory of communication”. In: *Bell Syst. Tech. J.* 27.3 (1948), pp. 379–423.

[174] J.P. Shim et al. “Past, present, and future of decision support technology”. In: *Decision Support Systems* 33.2 (2002). Decision Support System: Directions for the Nest Decade, pp. 111–126. ISSN: 0167-9236. DOI: https://doi.org/10.1016/S0167-9236(01)00139-7. URL: http://www.sciencedirect.com/science/article/pii/S0167923601001397.

[175] Anastasia Siapka. “The Ethical and Legal Challenges of Artificial Intelligence: The EU response to biased and discriminatory AI”. In: *SSRN* (Dec. 2018). DOI: http://dx.doi.org/10.2139/ssrn.3408773.

[176] C. Simoiu, S. Corbett-Davies, and S. Goel. “The problem of infra-marginality in outcome tests for discrimination”. In: *The Annals of Applied Statistics* 11 (2017), pp. 1193–1216.

[177] Edward H. Simpson. “Measurement of Diversity”. In: *Nature* 163.4148 (1949), pp. 688–688. DOI: 10.1038/163688a0.

[178] A. Singh and T. Joachims. *Equality of Opportunity in Rankings*. 2017. URL: https://www.k4all.org/wp-content/uploads/2017/09/WPOC2017_paper_9.pdf.

[179] Ashudeep Singh and Thorsten Joachims. "Fairness of Exposure in Rankings". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, pp. 2219–2228. ISBN: 9781450355520. DOI: 10.1145/3219819.3220088. URL: https://doi.org/10.1145/3219819.3220088.

[180] Ashudeep Singh and Thorsten Joachims. "Policy Learning for Fairness in Ranking". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. New York, US: Curran Associates, Inc., 2019, pp. 5426–5436. URL: http://papers.nips.cc/paper/8782-policy-learning-for-fairness-in-ranking.pdf.

[181] Z. Siting et al. "Job recommender systems: A survey". In: *2012 7th International Conference on Computer Science Education (ICCSE)*. 2012, pp. 920–924.

[182] Lin Song. "Two-Sided Price Discrimination by Media Platforms". In: *Marketing Science* 39.2 (2020), pp. 317–338. DOI: 10.1287/mksc.2019.1211. eprint: https://doi.org/10.1287/mksc.2019.1211. URL: https://doi.org/10.1287/mksc.2019.1211.

[183] Toll Speicher et al. "Potential for Discrimination in Online Targeted Advertising". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, Feb. 2018, pp. 5–19. URL: http://proceedings.mlr.press/v81/speicher18a.html.

[184] Byron Spice. *Questioning the Fairness of Targeting Ads Online*. 2017. URL: https://www.cmu.edu/news/stories/archives/2015/july/online-ads-research.html.

[185] Megha Srivastava, Hoda Heidari, and Andreas Krause. *Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning*. 2019. eprint: arXiv:1902.04783[cs].

[186] P. Stone et al. *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence*. Tech. rep. Stanford, CA: Stanford University, 2016. URL: http://ai100.stanford.edu/2016-report.

[187] Christian Strasser Helmut ans Weber. "On the asymptotic theory of permutation statistics". In: *Mathematical Methods of Statistics* 2 (1999), pp. 220–250. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.7071.

[188] Credit Suisse. *Algorithmic bias: a new fintech challenge*. 2017. URL: https://qz.com/1121150/algorithmic-bias-a-new-fintech-challenge/.

[189] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. "Evolution and impact of bias in human and machine learning algorithm interaction". In: *PLOS ONE* 15.8 (Aug. 2020), pp. 1–39. DOI: 10.1371/journal.pone.0235502. URL: https://doi.org/10.1371/journal.pone.0235502.

[190] Harini Suresh and John V. Guttag. "A Framework for Understanding Unintended Consequences of Machine Learning". In: *arXiv preprint arXiv:1901.10002* (2019). URL: https://arxiv.org/abs/1901.10002.

[191] Latanya Sweeney. "Discrimination in Online Ad Delivery". In: *Queue - Storage* 11.3 (2013), 10:10–10:29. DOI: 10.1145/2460276.2460278. URL: http://doi.acm.org/10.1145/2460276.2460278.

[192] Henri Theil. *Economic Forecasts and Policy*. Contributions to economic analysis. North-Holland Publishing Company, 1961.

[193] Nava Tintarev. "Presenting Diversity Aware Recommendations: Making Challenging News Acceptable". In: *FATREC Workshop on Responsible Recommendation at RecSys*. Como, Italy, 2017.

[194] Tatiana Tommasi et al. *A Deeper Look at Dataset Bias*. Swiss: Csurka G. (eds) Domain Adaptation in Computer Vision Applications. Advances in Computer Vision and Pattern Recognition, Springer, Cham, 2017, pp. 37–55. ISBN: 978-3-319-58347-1. DOI: 10.1007/978-3-319-58347-1_2. URL: https://doi.org/10.1007/978-3-319-58347-1_2.

[195] Virginia Tsintzou, Evaggelia Pitoura, and Panayiotis Tsaparas. *Bias Disparity in Recommendation Systems*. 2018. eprint: {arXiv}:1811.01461[cs]. URL: http://arxiv.org/abs/1811.01461.

[196] Nicol Turner Lee, Paul Resnick, and Genie Barton. *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. 2019. URL: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

[197] H. Varian. "Equity, envy, and efficiency". In: *Journal of Economic Theory* 9 (1974), pp. 63–91.

[198] Michael Veale and Reuben Binns. "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data". In: *Big Data & Society* 4.2 (2017). DOI: 10.1177/2053951717743530. URL: https://doi.org/10.1177/2053951717743530.

[199] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *Proceedings of the International Workshop on Software Fairness*. FairWare '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450357463. DOI: 10.1145/3194770.3194776. URL: https://doi.org/10.1145/3194770.3194776.

[200] Yan Wang and Xuelei Sherry Ni. "Predicting Class-Imbalanced Business Risk Using Resampling, Regularization, and Model Emsembling Algorithms". In: *International Journal of Managing Information Technology (IJMIT)* 11.1 (2017). URL: https://ssrn.com/abstract=3366806.

[201] Darrell M. West and John R. Allen. *How artificial intelligence is transforming the world*. 2018. URL: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/.

[202] J. Wexler et al. "The What-If Tool: Interactive Probing of Machine Learning Models". In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), pp. 56–65. DOI: 10.1109/TVCG.2019.2934619.

[203] Betsy A. Williams, Catherine F. Brooks, and Yotam Shmargad. "How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications". In: *Journal of Information Policy* 8 (2018), pp. 78–115. ISSN: 23815892, 21583897. URL: https://www.jstor.org/stable/10.5325/jinfopoli.8.2018.0078.

[204] J. Williams and J. Z. Kolter. *Dynamic modeling and equilibria in fair decision making*. 2019. URL: arXiv%20preprint%20arXiv:1911.06837.

[205] Lin Xiao et al. "Fairness-Aware Group Recommendation with Pareto-Efficiency". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. RecSys '17. Como, Italy: ACM, 2017. DOI: 10.1145/3109859.3109887.

[206] Ke Yang and Julia Stoyanovich. "Measuring Fairness in Ranked Outputs". In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. SSDBM '17. Chicago, IL, USA: Association for Computing Machinery, 2017. ISBN: 9781450352826. DOI: 10.1145/3085504.3085526. URL: https://doi.org/10.1145/3085504.3085526.

[207] T. Yang. "Choice and Fraud in Racial Identification: The Dilemma of Policing Race in Affirmative Action, the Census, and a Color-Blind Society". In: *Michigan Journal of Race and Law* 11.2 (2006).

[208] Sirui Yao and Bert Huang. "Beyond Parity: Fairness Objectives for Collaborative Filtering". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 2921–2930. URL: http://papers.nips.cc/paper/6885-beyond-parity-fairness-objectives-for-collaborative-filtering.pdf.

[209] Samuel Yeom, Anupam Datta, and Matt Fredrikson. "Hunting for Discriminatory Proxies in Linear Regression Models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018, pp. 4568–4578. URL: https://proceedings.neurips.cc/paper/2018/file/6cd9313ed34ef58bad3fdd504355e72c-Paper.pdf.

[210] Muhammad Bilal Zafar et al. "Fairness Beyond Disparate Treatment Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 1171–1180. ISBN: 9781450349130. DOI: 10.1145/3038912.3052660. URL: https://doi.org/10.1145/3038912.3052660.

[211] Muhammad Bilal Zafar et al. "Fairness Constraints: A Mechanism for Fair Classification". In: *20th International Conference on Artificial Intelligence and Statistics, AISTATS*. Fort Lauderdale, Florida, USA, 2017.

[212] Muhammad Bilal Zafar et al. "From Parity to Preference-Based Notions of Fairness in Classification". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 228–238. ISBN: 9781510860964.

[213] Meike Zehlike et al. "FA*IR: A Fair Top-k Ranking Algorithm". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17. Singapore, Singapore: Association for Computing Machinery, 2017, pp. 1569–1578. ISBN: 9781450349185. DOI: 10.1145/3132847.3132938. URL: https://doi.org/10.1145/3132847.3132938.

[214] Achim Zeileis, David Meyer, and Hornik Kurt. "Residual-based Shadings for Visualizing (Conditional) Independence". In: *Report 20* (2005).

[215] Guan Zhong. "Efficient and robust density estimation using Bernstein type polynomials". In: *Journal of Nonparametric Statistics* 28.2 (2016), pp. 250–271. DOI: 10.1080/10485252.2016.1163349.

[216] Yao Zhou et al. "Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction". In: *Briefings Bioinform* 18.5 (2017), pp. 744–753. DOI: 10.1093/bib/bbw064. URL: https://doi.org/10.1093/bib/bbw064.

[217] Donald W. Zimmerman, Bruno D. Zumbo, and Richard H. Williams. "Bias in estimation and hypothesis testing of correlation". In: *Psicológica* 24.1 (2017), pp. 133–158.

[218] Indre Zliobaite. *On the relation between accuracy and fairness in binary classification*. 2015. eprint: arXiv:1505.05723[cs].

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.