NOVA IMS Information Management School

MDSAA

Mestrado em

Data Science and Advanced Analytics

Bias in Artificial Intelligence

Exploring Its Role in Institutional Discrimination and Strategies for Mitigation

Hubert Josef Oberhauser

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Bias in Artificial Intelligence

Exploring Its Role in Institutional Discrimination and Strategies for Mitigation

by

Hubert Josef Oberhauser

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Filipe Montargil, Prof., NOVA Information Management School

November, 2024

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, 30.11.2024

Hubert Josef Oberhauser

Abstract

Artificial Intelligence, especially through machine learning, has become ubiquitous in everyday life, influencing both individual choices and institutional decision-making processes. Despite its potential to bring societal benefits, AI has faced considerable criticism because it perpetuates biases contributing to institutional discrimination. This thesis explores how biases come into AI systems through imbalance in data, algorithmic design, or user interaction, and how they reinforce systemic inequalities in critical domains such as recruitment, justice, and finance. The thesis is based on an extensive literature review, discussing both the technical foundations of AI and the notion of institutional discrimination defined under the EU and German frameworks as systemic inequalities that are buried within institutional practices. It showcases real-world impacts: the biased algorithms, including Amazon's hiring tool, the COMPAS criminal justice system, and discriminatory credit scoring mechanisms, through which AI systems replicate and amplify historical inequities. These challenges are addressed by this research, which assesses the strategies for mitigating AI bias through technical, regulatory, and ethical approaches. Technical interventions include fairness-aware algorithms and improved data practices, while regulatory measures, such as the EU AI Act and GDPR, enforce accountability and transparency. Ethical frameworks, including the OECD Principles on AI and EU Ethics Guidelines for Trustworthy AI, emphasize inclusivity and fairness in AI governance. By integrating these perspectives, actionable strategies to reduce bias and advance equity in AI systems are provided. It reflects a deep understanding of the power of interdisciplinary collaboration in AI for social progress, fairness, and accountability in its deployment processes.

Keywords

Artificial Intelligence; Machine Learning; Bias; Discrimination; Mitigation

Table of Contents

List of Abbreviations and Acronyms

ADS	Antidiskriminierungsstelle des Bundes
ACLU	American Civil Liberties Union
AGG	Allgemeines Gleichbehandlungsgesetz (General Equal Treatment Act)
AI	Artificial intelligence
Art.	Article
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DP	Demographic Parity
D&I	Diversity and Inclusion
EC	Equalized Odds
ECC	Equality of Opportunity
e.g.	for example
EPIC	Electronic Privacy Information Center
et al.	and others
EU	European Union
FADA	Federal Anti-Discrimination Agency
FN	False Negative
FP	False Positive
FTC	Federal Trade Commission
HLEG AI	High-Level Expert Group on AI
ML	Machine Learning
OECD	Organisation for Economic Co-operation and Development
TN	True Negative
TP	True Positive
U. S.	United States
XAI	Explainable Artificial Intelligence

1. Introduction

Artificial Intelligence, largely understood through the paradigm of machine learning within popular discourses, increasingly pervades all aspects of life today, from individual decisions to broader institutional practices. Applications range from automating mundane tasks to informing critical policies in healthcare, education, and justice systems. While considerable benefits are possible through AI, increasing institutional decision-making integration has drawn quite a bit of scrutiny, specifically regarding its capacity to yield discriminatory outcomes.

"RECRUITERS USING AI COULD BE DISCRIMINATING AGAINST OLDER WORKERS"[1]
"Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism"[2]
"Amazon ditched AI recruitment software because it was biased against women"[3]

Recent critiques have highlighted various ways in which AI systems have perpetuated or exacerbated biases, bringing into question their role in institutional discrimination. These critiques raise a seminal question: The extent to which AI engenders institutionalized discrimination and how such effects can be minimized. Any attempt to answer this question must garner a deep understanding of both AI systems and the sociological phenomenon of institutionalized discrimination and the linkages between them. To build this understanding, the thesis employs a comprehensive literature review to synthesize existing research and theoretical insights, forming the basis for the analysis and evaluation of AI's impact on institutional equity.

This thesis begins by exploring the technical foundations of AI, with a focus on machine learning, which relies fundamentally on data and algorithmic decision-making. The analytical focus is on bias-a quality inherently imbricated into the data and models employed by AI, which can lead to discriminatory outcomes. Based on this foundation, the thesis deals with institutional discrimination, defined according to the EU and German frameworks as systemic inequalities consolidated in the structures and practices of institutions. It further explores how statistical discrimination in AI decision-making processes can reinforce these inequities.

---

[1] Vohs, M. (2024). Recruiters using AI could be discriminating against older workers. The HR Director. https://www.thehrdirector.com/business-news/ai/recruiters-using-ai-discriminating-older-workers/

[2] Grant, C. (2019). Algorithms in health care may worsen medical racism. Unclear regulation and a lack of transparency increase the risk that AI and algorithmic tools that exacerbate racial biases will be used in medical settings. American Civil Liberties Union. https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racis

[3] Winick, E. (2018, October 10). Amazon ditched AI recruitment software because it was biased against women. MIT Technology Review. https://www.technologyreview.com/2018/10/10/139858/amazon-ditched-ai-recruitment-software-because-it-was-biased-against-women/

These are underpinned by a theoretical exploration through real-world case studies in finance, recruitment, and the justice system. Each of these areas represents wide reaching into society; thus, any skewed decision-making could have significant effects on people and society as a whole. The case studies will further show specific examples of how algorithmic biases manifest in actual practice, shaping institutional outcomes in ways that may further entrench existing inequalities.

The thesis concludes by systematically assessing strategies to reduce AI bias and its contribution to institutional discrimination. These strategies range from technical interventions, through regulatory measures, to ethical frameworks. Technical approaches are targeted toward increasing algorithmic transparency and fairness through methodologies such as detection of bias, fairness-aware machine learning, and better data curation practices. Regulatory measures consider the role of laws and policies in enforcing accountability and oversight in AI systems. Ethical frameworks highlight values such as fairness, accountability, and inclusivity that should be at the core of designing and deploying AI systems.

Regulatory and ethical frameworks are addressed in this thesis mainly from a European perspective, drawing on key initiatives such as the EU AI Act, the General Data Protection Regulation, and the EU Ethics Guidelines for Trustworthy AI. The exception will be the inclusion of the OECD Principles on AI due to their global scope and relevance. This agrees with the values of openness and international collaboration, ensuring wide and diverse coverage, which extends the region-specific frameworks at the EU level. Merging technical, legal, and ethical standpoints, this study hopes to push forward in giving nuance to the conception of bias in AI and their grand ramification toward institutional equity.

Besides this, the thesis attempts to offer input in producing concrete and operational strategies through which fair, accountable, and just AI systems will be possible. These strategies are important in ensuring that the benefits of AI are equitably distributed throughout society and address systemic inequities that technology can perpetuate or exacerbate.

2. Fundamentals of Artificial Intelligence (AI)

To establish a solid foundation for examining bias in artificial intelligence and its potential role in fostering institutional discrimination, it is crucial to first define the core concepts and principles underlying AI systems. This section focuses on the fundamental elements of AI that directly contribute to the emergence and impact of bias, laying the groundwork for a comprehensive understanding of how these systems may reinforce or perpetuate discriminatory outcomes.

2.1 Defining AI

Artificial Intelligence is an interdisciplinary subfield of computer science, traditionally classified under applied informatics. Its core areas include knowledge representation and reasoning, heuristic search and planning, as well as machine learning. Prominent applications range from natural language processing and image and scene analysis to intelligent robotics and gaming (Russell & Norvig, 2020).

The origins of AI as a formal discipline can be traced to the 1956 Dartmouth Summer Research Project on Artificial Intelligence, where John McCarthy and colleagues proposed the central thesis: "[...] every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1956, p. 2). This event, often regarded as the inception of academic AI research, brought together pioneers such as Marvin Minsky, Nathaniel Rochester, and Claude Shannon to explore the replication of human intelligence in machines.

Despite the absence of a universally accepted definition of AI, McCarthy's idea remains highly influential: designing machines that behave in ways deemed intelligent if performed by humans. Minsky (1986, as cited in Haenlein & Kaplan, 2019, p. 17) echoed this by defining AI as "the science of making machines do things that would require intelligence if done by men." Similarly, contemporary definitions like the Britannica Dictionary's align with this view, defining AI as "an area of computer science that deals with giving machines the ability to seem like they have human intelligence" (Artificial Intelligence, 2024).

The conceptualization of AI hinges on debates around the definition and operationalization of intelligence itself, a topic that remains contentious. Scholars like Gentsch (2019) emphasize that for AI to emulate human intelligence without being human, it must replicate core aspects of human behavior such as "learning, reasoning, problem-solving, perception, and using language" (Copeland, 2024). Furthermore, Gentsch highlights the importance of adaptive responses as a hallmark of human-like behavior. Mainzer (2016, p. 3) broadens this perspective, describing AI as the ability of a system to "independently and efficiently solve problems," with its intelligence gauged by its autonomy, problem complexity, and solution efficiency. Early efforts to quantify intelligence include Alan Turing's renowned Turing Test, which posits that a machine is intelligent if it can simulate human interaction convincingly

enough to prevent an observer from distinguishing between human and machine responses (Haenlein & Kaplan, 2019). While influential, the Turing Test remains a subject of ongoing debate regarding its suitability as a benchmark for intelligence.

An alternative perspective on AI's definition is provided by the High-Level Expert Group on Artificial Intelligence established by the European Commission. Instead of focusing on the term "intelligence," they characterize AI systems based on their capabilities: "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information derived from this data, and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behavior by analyzing how the environment is affected by their previous actions" (High-Level Expert Group on Artificial Intelligence, 2019, p. 6).

AI is frequently divided into two categories: narrow AI and strong AI. Narrow AI refers to systems specialized in specific tasks, operating through pattern recognition and data-driven models (Russell & Norvig, 2020). Many technologies integrated into contemporary devices, such as smartphones and computers, fall into this category and often outperform humans in narrowly defined tasks like image recognition (Committee on Technology National Science and Technology Council, 2016). However, these systems lack general intelligence and remain confined to their designated purposes. Even the most advanced systems capable of solving problems autonomously are ultimately designed and programmed by humans and do not exhibit the breadth of human cognition (Mainzer, 2016).

Strong AI, by contrast, is a hypothetical concept referring to systems that could independently solve complex problems across diverse domains, self-improve, and surpass human intelligence. Currently, such systems do not exist, and there is ongoing debate about whether—and under what circumstances—they might ever become a reality (Committee on Technology National Science and Technology Council, 2016).

With this foundational understanding of AI established, attention now turns to specialized domains such as machine learning. These areas shed light on how AI systems function, including the mechanisms through which data is utilized for decision-making and adaptation. This exploration is essential to understanding where and how bias can infiltrate AI systems, a critical consideration for their development and application.

2.2 Data and Algorithms: The Foundations of AI

At its core, every software system is built on algorithms, which are commonly defined as programmed procedures designed to automate the resolution of complex mathematical or logical problems (Suresh & Guttag, 2019). The functionality of these systems hinges on three key components: input, computation and output. Input consists of the data used as the basis

for decision-making; output represents the processed results. While the exact output cannot always be predicted, its goals and characteristics must be clearly defined to ensure alignment with the intended objectives (Zweig & Lischka, 2018). The algorithm acts as the operational bridge between input and output, and its design critically determines the accuracy, efficiency, and reliability of the results.

Data serves as the foundation for decisions and model development in artificial intelligence, forming the raw material that drives learning and prediction. Data refers to structured information derived from real-world observations, measurements, or transactions and may take the form of numbers, text, images, or other formats. These datasets quantify attributes and behaviors essential for training AI models (Keim & Sattler, 2020). However, raw data is often flawed, containing errors, inconsistencies, or gaps, and must undergo a rigorous preparation process to ensure its usability in AI systems (Lee, 2020).

The preparation of data involves several critical steps designed to enhance its quality and consistency. Data preprocessing addresses issues such as missing values, inconsistencies, and noise, ensuring the dataset's reliability and integrity (Lee, 2020). Transformation standardizes data formats to meet the specific requirements of the AI model (Lee, 2020), while data integration consolidates datasets from diverse sources into a coherent and redundancy-free whole. Another essential step is normalization, which adjusts the scale and structure of data to ensure compatibility with algorithms like neural networks that rely on scaled input. Finally, data reduction minimizes the size of datasets without compromising their essential structure or informational value, improving both computational efficiency and algorithmic performance (García, Luengo & Herrera, 2015). These steps are crucial, as flawed or biased data can undermine the effectiveness and fairness of even the most sophisticated AI systems (García, Luengo & Herrera, 2015).

The importance of data quality in artificial intelligence is well recognized in the literature. Batini et al. (2009) identify six critical dimensions: accuracy, reflecting the degree to which data corresponds to real-world values; consistency, ensuring uniformity and coherence within and across datasets; completeness, guaranteeing that no essential values are missing; timeliness, emphasizing the need for up-to-date data in time-sensitive contexts; credibility, fostering trust in the reliability and authenticity of data sources; and interpretability, reflecting the extent to which data can be easily understood and utilized by its users. Although achieving perfect data quality in large datasets is often impractical, rigorous data preparation processes can significantly enhance the overall standard. This is crucial because even the most advanced algorithms are only as effective as the data they are trained on, and flawed or biased data inevitably leads to unreliable or skewed outcomes (Batini et al., 2009).

Having established the foundational importance of data and algorithms in AI systems, it is clear that their interplay is critical to enabling intelligent decision-making. This relationship is particularly evident in machine learning, where algorithms go beyond predefined instructions and leverage data to identify patterns, adapt, and improve. Understanding this

dynamic provides the basis for exploring the core principles and processes of machine learning, a pivotal subset of AI.

2.3 Machine Learning: Pattern Recognition and Learning Paradigms

Machine learning (ML), a fundamental pillar of artificial intelligence, involves training algorithms to identify patterns within data and learn from them without relying on explicit programming. The ultimate goal is to develop a function capable of accurately processing new inputs and making reliable predictions (Zweig & Lischka, 2018). Over the past decade, the rapid expansion of available data for training algorithms has driven the widespread adoption of machine learning. Its influence has grown so significantly that *AI* is often used synonymously with *machine learning* (Lipton, 2018).

The machine learning process typically unfolds in three key phases. The first phase, training, involves exposing the algorithm to a dataset containing representative examples, enabling it to discern patterns and use these insights to make predictions. In the second phase, validation, a separate dataset is used to evaluate and compare different models, allowing for the selection of the model that delivers the most accurate results (Suresh & Guttag, 2021). Finally, in the testing phase, the chosen model is assessed using an entirely distinct dataset to measure its generalization performance on unseen data (Suresh & Guttag, 2021).

Machine learning is broadly categorized into three learning paradigms, each characterized by distinct approaches to pattern recognition.

Supervised learning involves training algorithms on datasets containing labeled inputs and their corresponding outputs. The objective is to learn a mapping function that can accurately predict outputs for new inputs. During the training process, the algorithm iteratively refines its parameters to minimize discrepancies between its predictions and the actual outputs in the training data (Alpaydin, 2022). This paradigm is widely applied in tasks such as text or image classification, where algorithms categorize data into predefined groups, such as spam versus non-spam emails. Regression models also utilize supervised learning by optimizing parameters to reflect real-world relationships as accurately as possible.

Unsupervised learning, by contrast, does not rely on labeled outputs. Instead, its aim is for the algorithm to autonomously identify patterns, structures, or clusters within the data. A typical example is cluster analysis, where the algorithm groups similar data points based on shared characteristics. This approach is particularly valuable in uncovering hidden structures in large datasets, enabling meaningful data organization and interpretation (Alpaydin, 2022).

Reinforcement learning represents a third paradigm, particularly useful in scenarios where training data is scarce or desired outcomes are not explicitly defined. In this method, the algorithm learns by interacting with its environment, receiving feedback in the form of rewards for desirable outcomes or penalties for undesirable ones. Through this iterative

process, often referred to as a feedback loop, the algorithm refines its strategy to maximize cumulative rewards and develop an optimal solution pathway (Alpaydin, 2022).

As Fischer and Petersen (2018) observe, “understanding how algorithms function not only enhances appreciation of their benefits but also sharpens awareness of their risks” (p. 21, own translation). With the foundational principles of machine learning and the processes by which algorithms learn from data established, it becomes critical to focus on one of the field’s most pressing challenges: bias. Bias is a pervasive issue in AI, capable of distorting outcomes and leading to far-reaching ethical and societal consequences. A comprehensive understanding of bias is therefore essential for assessing the risks of algorithmic decision-making and developing strategies to mitigate its adverse effects.

3. Bias in AI and Its Role in Institutional Discrimination

The principles outlined in the last chapter provide the technical backbone for understanding how artificial intelligence systems operate. However, the very elements that make AI powerful—its reliance on data, algorithms, and adaptive learning—also create vulnerabilities to bias. Flaws in data quality, algorithmic design, and model training processes can introduce or exacerbate inequalities, particularly when deployed in institutional contexts.

For example, biased training datasets may encode historical inequities, while algorithmic optimization processes may prioritize efficiency over fairness. Moreover, the lack of transparency in many AI systems complicates the identification and mitigation of bias. These technical challenges do not exist in isolation but interact with broader societal structures, often reinforcing systemic discrimination.

This chapter builds on this foundation to examine bias in AI more directly, exploring its various forms, its roots in the machine learning lifecycle, and its role in perpetuating institutional discrimination.

3.1 Understanding Bias in AI

The concept of bias is diverse and context-dependent, generally defined as a distorting influence (Olteanu et al., 2019). In psychology, bias refers to attitudes or stereotypes related to factors such as age, gender, or ethnicity, which shape perceptions, decisions, and behaviors, both positively and negatively (Barmeyer & Genkova, 2011). In statistics, bias is understood as systematic errors in data collection or processing that lead to distorted and unrepresentative results (Olteanu et al., 2019). Biases can also be divided into explicit and implicit categories: explicit biases involve conscious attitudes toward specific groups, while implicit biases operate unconsciously, embedded within cognitive structures (Dean & Simpson, 2020). Implicit biases are particularly insidious because of their unconscious nature, subtly influencing decisions and behaviors, often resulting in discriminatory outcomes that disadvantage individuals or groups based on their social identity (Dean & Simpson, 2020).

Building on these definitions, bias in AI systems can represent a pervasive and multifaceted issue: As Crawford (2016) notes, AI systems inherently reflect the perspectives, assumptions, and priorities of their developers. Far from being neutral, algorithms and the systems they underpin inherit both the strengths and flaws of their creators. This often results in the transfer of human biases embedded in the training data into the system's outcomes (Suresh & Guttag, 2021).

In examining bias in machine learning systems specifically, researchers have categorized bias in various ways (Suresh & Guttag, 2021; Olteanu et al., 2019; Mehrabi et al., 2021; Siddique et al., 2024; Barocas & Selbst, 2016). However, most agree that bias originates from three

primary areas: the training data, the design of the algorithms, and the influence of the professionals involved in creating or using these systems. These biases can emerge at different stages of the machine learning lifecycle, spanning from data generation to model deployment (Mehrabi et al., 2021). To understand how these biases manifest and perpetuate inequality, the following chapter will examine the types and sources of bias in detail.

3.1.1 Types and Sources of Bias

According to Mehrabi et al. (2022), bias can be categorized into three primary types: data-driven, algorithmic, and user-driven bias. Data-driven bias arises from the data itself, particularly during collection and representation. This includes measurement bias, where proxies for intended outcomes poorly reflect the underlying reality, such as using arrest records as a proxy for criminality[4], and representation bias, resulting from the underrepresentation or misrepresentation of certain groups in datasets (Mehrabi et al., 2022). Additionally, historical bias, as noted by Suresh and Guttag (2021), often permeates datasets, embedding systemic inequalities that persist even with perfect sampling. This underscores the importance of critically examining the origins of data to mitigate inherited disparities (Suresh & Guttag, 2021).

Algorithmic bias occurs during model design and optimization, where subjective decisions about target variables, features, and objectives can encode or amplify disparities (Suresh & Guttag, 2021). Proxy bias arises when neutral-seeming features act as substitutes for sensitive attributes, such as zip codes functioning as proxies for race in credit scoring, while masking bias obscures discriminatory patterns through selective data handling or feature engineering (Barocas & Selbst, 2016). Aggregation bias occurs when models treat diverse populations as homogeneous, leading to poorer generalization for subgroups (Suresh & Guttag, 2021). Furthermore, postprocessing steps, such as adjustments made after training to achieve desired outputs, may inadvertently reinforce or exacerbate existing biases (Suresh & Guttag, 2021).

User-driven bias arises through interactions with AI systems. Feedback loops, such as those seen in interaction bias, can amplify disparities as users' behaviors, like preferentially clicking on top-ranked search results, reinforce skewed outputs (Mehrabi et al., 2022). These biases are dynamic and evolve with system usage, making their mitigation a persistent and ongoing challenge (Suresh & Guttag, 2021).

3.1.2 Stages Where Bias Arises

Bias can emerge and compound at multiple stages of the machine learning lifecycle. During data generation, biases may originate in the selection of data sources, which often reflect societal prejudices, or during data preparation, where processing and curation methods can distort group representation (Mehrabi et al., 2022; Olteanu et al., 2019). For example,

[4] All examples are discussed in detail in Chapter 4.

historical inequities embedded in datasets can lead to the reinforcement of systemic discrimination, as noted by Barocas and Selbst (2016).

In the model development phase, subjective decisions about defining target variables and selecting features introduce bias. For instance, defining *creditworthiness* based on repayment history may embed structural inequalities, while feature selection bias occurs when features that better explain variations for certain groups are excluded, leading to less accurate predictions for those groups (Barocas & Selbst, 2016). Furthermore, prioritizing accuracy over fairness during optimization can exacerbate disparities, especially for underrepresented groups (Suresh & Guttag, 2021).

During model evaluation, bias emerges when testing datasets fail to represent the deployment context, leading to evaluation bias. For instance, models tested predominantly on urban data may perform poorly when applied in rural areas (Suresh & Guttag, 2021). Additionally, unrepresentative evaluation metrics can produce skewed performance results, complicating fairness assessments. In the deployment phase, deployment bias arises when models are applied in real-world contexts that differ significantly from their training assumptions. This mismatch can amplify existing inequalities, particularly when AI tools disproportionately impact vulnerable populations (Barocas & Selbst, 2016). Biases may also intensify post-deployment due to emergent factors such as demographic shifts or cultural changes, further complicating their mitigation (Suresh & Guttag, 2021).

With an understanding of how bias arises in AI systems, the focus now shifts to its impact, exploring how these biases can perpetuate systemic inequalities within institutional contexts. The next chapter introduces the concept of institutional discrimination and sets the stage for the case studies that follow, illustrating the real-world implications of AI-driven institutional discrimination.

3.2 Understanding Discrimination

Institutions, through their policies, practices, and decision-making frameworks, often play a central role in shaping societal outcomes. While they are designed to promote fairness and efficiency, these structures can also perpetuate systemic inequalities. The integration of AI systems into institutional processes introduces new challenges, as these systems can inadvertently reinforce existing disparities or even create new forms of discrimination.

This chapter examines the relationship between institutional discrimination and AI systems. It begins with an exploration of the concept of discrimination, followed by a detailed discussion of institutional discrimination. Particular attention is given to statistical discrimination, highlighting how bias in AI systems can disproportionately impact marginalized groups and contribute to systemic inequities.

3.2.1 Definition and Conceptual Understanding

The concept of discrimination varies across societies, historical periods, and regions. The term discriminate, derived from the Latin *discriminare*, means *to distinguish*.[5] While distinctions can help simplify complex matters (Beck 2019), the critical question is whether they are justified. In social and political contexts, discrimination in the negative sense refers to the unjustified unequal treatment of equals or the unjustified equal treatment of unequals (Gomolla, 2008). This thesis adopts the EU and German definition, framing discrimination as the unjustified and disadvantageous unequal treatment of individuals based on protected characteristics (European Union Agency for Fundamental Rights & Council of Europe, 2018).

In Germany, the General Equal Treatment Act (Allgemeines Gleichbehandlungsgesetz, AGG) provides a legal framework to combat discrimination. According to §1 AGG, protected characteristics include race or ethnic origin, gender, religion or belief, disability, age, and sexual identity (Allgemeines Gleichbehandlungsgesetz, 2006, last amended in 2022). §3 AGG categorizes discrimination into direct discrimination, where individuals are treated less favorably due to a protected characteristic, and indirect discrimination, where ostensibly neutral practices disproportionately disadvantage specific groups. For instance, part-time employment policies often affect women disproportionately due to traditional gender roles (Hagendorff, 2019). Similarly, applicants with non-German-sounding names may face disadvantages in recruitment due to biases associated with ethnic origin (Hagendorff, 2019). Such measures are not deemed discriminatory if they are objectively justified by a legitimate aim and applied using proportionate means (Allgemeines Gleichbehandlungsgesetz, 2006, last amended in 2022).

Importantly, §3 AGG emphasizes that discrimination does not require intent or malicious purpose. The defining factor is the adverse impact on the individual due to unequal treatment, underscoring the consequences of the action rather than its motivation (Antidiskriminierungsstelle des Bundes n.d.). This framework ensures that both overt and subtle forms of discrimination are addressed, promoting equal treatment across all areas of life.

However, discrimination does not always stem from individual actions or prejudices. In many cases, it is embedded within societal structures and institutional frameworks. This type of systemic inequity, known as institutional discrimination, operates independently of individual intent and plays a significant role in perpetuating disadvantage across social groups.

[5] The terms "discriminate" and "distinguish" differ in both meaning and connotation in English: while "distinguish" is typically used in a neutral or positive sense to indicate the act of identifying differences, "discriminate" often carries a negative implication, referring to unfair treatment based on specific attributes or characteristics (Cambridge Dictionary, n.d.). For further details, see Cambridge Dictionary: Discriminate (https://dictionary.cambridge.org/dictionary/english/discriminate) and Cambridge Dictionary: Distinguish (https://dictionary.cambridge.org/dictionary/english/distinguish).

3.2.2 Institutional Discrimination

Hasse and Schmidt (2012) define institutional discrimination as the systematic and persistent disadvantage of social groups, deeply embedded in formal rights, organizational frameworks, programs, and routines in domains such as employment and labor, education, the legal system, finance, and healthcare (Jost et al., 2009). This form of discrimination reinforces existing inequalities, extending its impact beyond isolated cases (Hasse & Schmidt, 2012).

This concept is rooted in Durkheim's understanding of institutions. Durkheim (1980) viewed institutions as recurring societal phenomena that emerge independently of individual actions and are grounded in collective social patterns. He described them as external forces that shape individual behavior and remain statistically observable and predictable within a society. These structures evolve over time, often becoming stable and persistent features of social life. Organizations, as a distinct type of institution, are shaped by external influences such as societal values and cultural constructs. These forces impose normative and structural frameworks that guide their operations, often formalized through formal rules and procedures (Gomolla, 2017).

Unlike individual discrimination, institutional discrimination is not necessarily intentional. Well-meaning actors may unintentionally perpetuate it through ingrained organizational cultures and professional routines (Hasse & Schmidt, 2012). As the underlying mechanisms are often opaque, institutional discrimination is difficult to recognize and prove - both for those affected and for those involved (Gomolla, 2017). This complexity arises from the fact that discriminatory processes are deeply embedded in the everyday culture of organizations and the professional routines of their staff (Gomolla, 2008). The sociologist Stuart Hall (2001) emphasized that such dynamics are often reinforced informally through habitual practices and daily procedures, becoming so normalized that they are no longer consciously acknowledged. Organizations are particularly prone to institutional discrimination, as formal rules and procedures often leave room for unjustified unequal treatment (Hasse & Schmidt, 2012).

The integration of AI systems into organizational processes further complicates institutional discrimination by embedding and amplifying biases present in the datasets they process. Operating within critical institutions like healthcare, education and legal systems, AI systems frequently reinforce structural inequalities rather than alleviating them (Hagendorff, 2019a). For instance, algorithms may use irrelevant personal characteristics, such as race or gender, as proxies for decision-making criteria, reflecting and perpetuating existing inequities (Suresh & Guttag, 2019). This dynamic is exemplified by statistical discrimination, where algorithms rely on group-level proxies to optimize decision-making.

3.2.3 Statistical discrimination in AI

Statistical discrimination is a key mechanism through which AI systems perpetuate institutional inequities. It occurs when algorithms rely on group-level data or statistical

correlations rather than individual-specific information to optimize decisions (Guzman et al., 2021). This heuristic approach often arises when individualized data is unavailable or costly to acquire, leading algorithms to use observable features, such as race, gender, or socioeconomic status, as proxies for desired attributes like creditworthiness, productivity, or risk.[6] While this approach may enhance decision-making efficiency, it frequently embeds and amplifies existing biases from the training environment. Guzman et al. (2021) demonstrate that the extent of statistical discrimination depends heavily on biases present in training data. Algorithms trained in contexts where specific features correlate with behaviors (e.g., cooperation or defection) are more likely to adopt discriminatory patterns.

Statistical discrimination does not require intent or malice. Instead, it reflects the optimization processes of algorithms designed to maximize outcomes based on available data. When these algorithms operate within institutional frameworks shaped by systemic inequalities, they exacerbate disparities, especially when decisions impact access to resources, opportunities, or rights (Guzman et al., 2021). Addressing statistical discrimination requires bias-aware training practices and fairness-oriented interventions that account for the societal impact of AI systems.

The following sections focus on specific institutional domains where bias in AI systems has led to discriminatory outcomes. Through detailed case studies, these sections translate the theoretical understanding of bias and institutional discrimination into practical contexts, highlighting their real-world implications.

[6] Further explanation can be found in Chapter 4.

4. Applications of AI Bias leading to Discrimination in Institutional Contexts

This chapter examines how AI bias manifests in three selected domains: finance, recruitment, and criminal justice. Through these case studies, it highlights the mechanisms through which biased algorithms contribute to institutional discrimination, perpetuating existing inequities and creating new barriers for marginalized groups.

Beyond these immediate impacts, the chapter also explores the broader, long-term consequences of institutional discrimination driven by AI. It considers how such biases can erode public trust and deepen systemic inequalities. By understanding both the specific instances and the overarching implications of AI-driven discrimination, this chapter aims to provide a comprehensive view of the societal risks posed by unmitigated bias in institutional systems.

4.1 Analysis of Specific Institutional Areas

To better understand the concrete effects of AI bias, this section focuses on its application in three different institutional domains. By examining finance, recruitment, and criminal justice, the specific mechanisms through which biased algorithms operate and the resulting discriminatory outcomes are highlighted. Each case offers unique insights into the ways in which institutional contexts influence the impact of AI bias.

4.1.1 AI Bias in Recruitment

Building on the discussion of bias and discrimination in the previous chapter, this section delves into how bias in artificial intelligence systems contributes to discriminatory outcomes in institutional contexts, with a focus on employment and labor. In recruitment processes, AI-enabled tools are often promoted for their potential to enhance efficiency and objectivity. However, these systems are inherently vulnerable to replicating and amplifying existing societal biases. By operationalizing patterns derived from historical and often inequitable data, such systems can unintentionally perpetuate discrimination, leading to exclusionary practices that disproportionately disadvantage marginalized groups.

Chen (2023) identifies several critical factors contributing to bias in AI recruitment systems. As already discussed, Algorithmic discrimination often stems from biased training data and subjective decisions made during algorithm design. Skewed datasets frequently underrepresent historically marginalized groups, such as women and ethnic minorities, resulting in hiring decisions that reflect and reinforce societal inequities. Similar findings are echoed by Fabris et al. (2023), who emphasize that skewed datasets and algorithmic design choices systematically exclude underrepresented groups in hiring. AI-based recruitment systems also exhibit discrimination across multiple dimensions, including gender, race, color, and personality traits. For instance, facial recognition systems have been shown to misclassify or disadvantage candidates with darker skin tones, a phenomenon also highlighted by Wilson and Caliskan (2024), who demonstrated that language models used in

resume screening favor White-associated names in 85.1% of cases. Furthermore, many organizations perceive AI systems as inherently neutral, but Mujtaba and Mahapatra (2024) argue that these tools often replicate historical inequities under the guise of objectivity. The impacts of such biases are far-reaching: discriminatory practices harm individuals, reduce organizational diversity, and undermine the economic benefits of AI-driven recruitment (Chen, 2023; Li, Li, & Lu, 2023). Compounding this issue, the lack of transparency in algorithmic processes creates an "algorithmic black box," leaving candidates unable to contest biased decisions or understand the rationale behind them (European Union Agency for Fundamental Rights, 2022; Chen, 2023).

One well-documented example is Amazon's discontinued AI-powered hiring tool. Launched in 2014, the system utilized machine learning algorithms to evaluate job applicants' resumes. By 2015, Amazon discovered that the tool exhibited significant gender bias against women (Dastin, 2018). This bias stemmed from the training data, which consisted of resumes submitted to Amazon over a ten-year period, most of which came from male applicants due to the male-dominated tech industry (Simonite, 2018). Consequently, the AI system learned to prefer male candidates and penalized resumes containing terms like "women's," such as "captain of women's chess club," and downgraded graduates from all-women's colleges (Dastin, 2018; Simonite, 2018). Despite efforts to correct the bias, Amazon ultimately abandoned the tool, citing concerns that the system would continue to produce discriminatory patterns (Dastin, 2018).

Chang (2023) provides further insights into the broader implications of Amazon's tool, revealing how gender biases in AI recruitment can exacerbate systemic inequities. The study highlights that men applying for technical roles at Amazon were twice as likely to be hired as women (40% vs. 20%), reflecting how the algorithm disproportionately disadvantaged female candidates (Chang, 2023). Furthermore, Chang demonstrates that the algorithm's reliance on stereotypically *male* keywords and experiences not only reduced the diversity of hires but also influenced salary offers, with women often receiving lower compensation packages compared to their male counterparts (Chang, 2023). These findings show how algorithmic tools can institutionalize gender inequities under the guise of objectivity.

Another case involves HireVue, a company offering AI-driven assessments of job applicants through video interviews. Their technology analyzes facial expressions, tone of voice, and word choice to predict a candidate's employability. Critics argue that such systems can inadvertently discriminate against individuals who do not conform to specific speech patterns or facial expressions, potentially disadvantaging non-native speakers, people with disabilities, or those from different cultural backgrounds (EPIC, 2019). In 2019, the Electronic Privacy Information Center (EPIC) filed a complaint with the Federal Trade Commission (FTC), asserting that HireVue's AI assessments were unproven and biased, potentially violating candidates' rights (EPIC, 2019). Similarly, the American Civil Liberties Union (ACLU) raised concerns about HireVue’s lack of transparency, warning that opaque scoring systems could

perpetuate discrimination without candidates' knowledge or recourse (ACLU, 2020). Following significant backlash in 2021, HireVue reduced its reliance on facial analysis and shifted toward text-based assessments (Simonite, 2021).

LinkedIn's AI-based recruitment algorithms have similarly come under scrutiny. Studies have shown that the platform's algorithms were more likely to recommend higher-paying job opportunities to men than to women, reflecting existing gender disparities in the labor market (Chen, 2023). This bias was attributed to the algorithm's reliance on historical hiring patterns, which disproportionately favored men. Moreover, the system's feedback loop exacerbated these disparities: as more men applied for and were hired into high-paying roles, the algorithm increasingly recommended such roles to male candidates, further marginalizing women (Chen, 2023).

A further example involves the perpetuation of racial bias in AI systems. Research conducted by Caliskan, Bryson, and Narayanan (2017) revealed that widely used AI language models encode biases that associate certain races or genders with negative stereotypes. When these models are integrated into recruitment tools—for instance, for parsing and ranking resumes—they can reinforce discriminatory patterns by favoring candidates who align with biased profiles (Caliskan et al., 2017).

Chen (2023) also explores regional biases in AI recruitment systems. A multinational company's AI-driven recruitment algorithm was found to systematically deprioritize candidates from rural or economically disadvantaged areas. This occurred because the algorithm was trained on historical hiring data that predominantly favored candidates from urban regions with a higher concentration of successful applicants. As a result, rural applicants were unfairly penalized, perpetuating existing regional inequalities in hiring (Chen, 2023).

These cases demonstrate that institutional discrimination in AI recruitment is not merely a technical issue but a systemic one, deeply intertwined with historical inequities and societal norms. The opacity of many AI-driven systems exacerbates the problem, making it challenging for candidates or regulators to detect and address unfair practices. As these examples illustrate, biases embedded in training data and algorithmic design can have profound and lasting impacts on the diversity and fairness of hiring practices.

4.1.2 AI Bias in Justice

The case of the COMPAS algorithm provides a striking example of how biased AI systems can contribute to institutional discrimination, particularly within the criminal justice system. COMPAS, which stands for *Correctional Offender Management Profiling for Alternative Sanctions*, is a predictive tool developed in 1998 to assess the likelihood of recidivism among offenders. Since its introduction, COMPAS has been widely implemented in the United States and has been used to evaluate more than one million individuals. The algorithm plays a

critical role in judicial decision-making, including sentencing and probation evaluations (Dressel & Farid, 2018).

The functionality of COMPAS relies on analyzing data across 137 variables, encompassing personal details such as age, gender, and place of residence, as well as social and biographical information (Dressel & Farid, 2018). For instance, the algorithm factors in whether an individual's parents were ever incarcerated, the prevalence of illegal drug use among their acquaintances, and whether they engaged in fights during school. These inputs are synthesized into a risk score ranging from 1 to 10, with scores above 5 indicating a medium-to-high likelihood of recidivism. While this system ostensibly aims to provide an evidence-based assessment of risk, it has drawn significant criticism for its lack of transparency and the discriminatory outcomes it produces (Rudin, Wang, & Coker, 2020).

A landmark investigation by ProPublica examined the outputs of COMPAS for 7,000 individuals arrested in Broward County, Florida, between 2013 and 2014 (Angwin, Larson, Mattu, & Kirchner, 2016). Although race is not explicitly included as an input variable, the study revealed substantial disparities in the algorithm's error rates for Black and white defendants. The findings indicated that Black defendants were nearly twice as likely as white defendants to be incorrectly classified as high risk for future criminal activity. Specifically, 45% of Black defendants who did not reoffend were labeled as high risk, compared to 23% of white defendants. Conversely, white defendants who reoffended were more often incorrectly classified as low risk. These findings suggest that COMPAS reproduces and exacerbates systemic racial biases embedded in the underlying data, treating Black defendants more harshly than their white counterparts without justifiable cause (Angwin et al., 2016).

However, subsequent research by Barenstein (2019) revisited ProPublica's dataset and identified a significant data processing error that inflated the recidivism rate by over 24%. This error occurred because ProPublica failed to consistently apply a two-year observation window for recidivists and non-recidivists. While non-recidivists with insufficient observation time were excluded, recidivists within the same window were retained, leading to an overrepresentation of recidivists in the dataset. As a result, while ProPublica reported a recidivism rate of 45.1%, Barenstein's correction showed that the actual rate was closer to 36.2%. Despite this correction, the racial disparities highlighted by ProPublica, such as the higher false positive rate (FPR) for Black defendants, remain valid and unaffected by the data processing error. Barenstein's findings underscore the importance of rigor in dataset construction when assessing algorithmic fairness.

Bao et al. (2021) further examined the COMPAS dataset and reinforced ProPublica's core findings by highlighting systemic biases in both the data and the socio-technical context of COMPAS. They pointed out that the outcome variable often relies on "re-arrest" data rather than "re-offense," introducing measurement bias influenced by over-policing in predominantly Black neighborhoods. Covariates like criminal history and neighborhood crime rates, although not explicitly tied to race, act as proxies for racial disparities in law

enforcement practices, perpetuating inequities. Bao et al. also criticized fairness metrics such as Equalized Odds or Demographic Parity, arguing that they fail to address broader societal inequities embedded in the criminal justice system. Moreover, discretionary decision-making by judges further complicates the fairness of COMPAS, as human interpretations and overrides of its predictions can amplify disparities. Together, these findings validate ProPublica's concerns about the systemic injustices underlying COMPAS, independent of Barenstein's critique of their dataset.

The algorithm has faced further criticism for its opacity and methodological shortcomings. The proprietary nature of COMPAS's risk models, developed by the private company Equivant (formerly Northpointe), prevents independent scrutiny and limits understanding of how its predictions are generated (Rudin et al., 2020). This lack of transparency has fueled concerns over the reliability and fairness of its assessments. Additionally, the complexity of the model, which incorporates numerous variables, increases the likelihood of data collection errors that could influence outcomes. These flaws underscore the broader issue of bias in AI systems, particularly when such systems are applied to sensitive domains like criminal justice (Dressel & Farid, 2018; Rudin et al., 2020).

Northpointe has contested the critiques of ProPublica, arguing that their analysis fails to consider standard fairness measures and highlighting that COMPAS assessments are based on a broad dataset encompassing diverse factors (Dieterich, Mendoza, & Brennan, 2016). Nevertheless, studies such as ProPublica's have raised significant ethical and practical questions about the deployment of algorithms like COMPAS. These concerns include the system's tendency to reinforce existing social inequities and the challenges of balancing accuracy with fairness across demographic groups (Angwin et al., 2016; Rudin et al., 2020).

The case of COMPAS has sparked extensive debate in the Anglophone world regarding the ethical implications of algorithmic decision-making in the justice system. In contrast, similar tools have not been widely adopted in the German-speaking context, partly due to the critical issues identified in the U.S. experience (Räz, 2022).

This example highlights the pressing need for greater transparency in algorithmic decision-making, particularly in high-stakes contexts such as criminal justice. It demonstrates how biases embedded in training data and opaque scoring mechanisms can disproportionately impact marginalized communities, perpetuating systemic inequalities.

4.1.3 AI Bias in Finance

The integration of AI into the financial sector, particularly in credit lending and risk assessment, has amplified concerns about institutional discrimination. While AI systems have the potential to analyze vast datasets and identify complex patterns, they are inherently shaped by the biases present in their training data. These biases can lead to discriminatory practices that disproportionately harm certain demographic groups, especially in areas such

as credit scoring, lending decisions, and the use of unconventional data sources (Sadok et al., 2022; Noble, 2018).

A notable example of bias in AI-driven financial services is the perpetuation of racial disparities through credit assessment algorithms. Historical practices like redlining, which systematically excluded racial minorities from financial services, have left a lasting imprint on the datasets used to train these algorithms. Noble (2018) highlights how modern credit algorithms frequently replicate these historical inequities, disproportionately flagging neighborhoods with high concentrations of minority residents as high-risk. This practice results in higher interest rates, more frequent loan denials, and reduced credit scores for affected groups (Bartlett et al., 2022). Bartlett et al. (2022) estimate that such algorithmic underwriting models cost Black and Hispanic borrowers an additional $765 million annually in interest payments—an outcome directly tied to systemic bias in data and algorithm design.

Gender disparities also emerge in AI-driven financial tools. A high-profile case involved allegations against Apple and Goldman Sachs, where female applicants for a credit card were systematically assigned lower credit limits than their male counterparts, despite comparable financial profiles (CNBC, 2023). Such outcomes strongly suggest that the algorithms used perpetuated gender biases rooted in historical data reflecting long-standing inequities in women's financial access (Bartlett et al., 2022).

Bartlett et al. (2022) note that such disparities demonstrate how AI systems, when built on biased datasets, can unintentionally replicate societal prejudices.

Another dimension of bias arises from the use of alternative data in financial decision-making. AI systems increasingly rely on unconventional sources, such as smartphone usage, social media activity, and other personal data, to assess creditworthiness. The World Bank (2021) warns that such practices often penalize marginalized groups due to their differing social and financial networks. For instance, individuals from economically disadvantaged backgrounds are disproportionately affected by algorithms that penalize users based on associations with financially unstable contacts. While such practices may appear neutral, they disproportionately harm marginalized communities whose networks are constrained by structural inequalities. These decisions, as Sadok et al. (2022) note, reflect systemic biases deeply embedded in the training data.

The complexity of AI systems, particularly their black box-nature, compounds these issues. Sadok et al. (2022) explain that advanced machine learning models, such as those employing ensemble methods like boosting, often obscure the relationship between input variables and outcomes. This lack of transparency not only complicates accountability but also reinforces perceptions of unfairness among affected groups. Similarly, Penedo and Kramcsák (2023) emphasize that the opacity of many financial AI systems allows biases to remain undetected, institutionalizing discrimination in critical processes like credit approvals.

Ratzan and Rahman (2024) echo these concerns, noting that it is often unclear whether biased outcomes stem from the algorithm, the training data, or both.

Institutional harm caused by biased AI systems is not confined to credit decisions or the United States. The Dutch child welfare scandal serves as a stark international example of how algorithmic systems can lead to large-scale discrimination. In this case, an algorithm used to detect welfare fraud disproportionately targeted low-income and minority families, resulting in significant financial and emotional harm. The World Bank (2021) identifies this incident as a cautionary tale, illustrating how biased AI systems, when deployed without adequate oversight, can institutionalize systemic inequalities.

In conclusion, the integration of AI into the financial sector has revealed significant vulnerabilities to bias and discrimination, particularly against historically marginalized groups. By operationalizing inequities embedded in training data and obscuring decision-making processes, AI systems risk perpetuating societal disparities under the guise of objectivity. These issues demand critical examination as they raise profound ethical and societal questions about the implications of automated decision-making in finance.

4.2 Long-Term Implications of Institutional Discrimination

4.2.1 Reduced Trust in Institutions

The erosion of trust in institutions represents a significant long-term consequence of institutional discrimination resulting from bias in AI. Discriminatory outcomes in AI applications, such as digital redlining in lending practices or biased hiring algorithms, perpetuate systemic inequality and marginalization, fundamentally undermining the public's confidence in the institutions deploying these systems (Shams et al., 2023). For example, when marginalized groups consistently experience negative impacts due to biased AI systems, they are likely to perceive institutions as untrustworthy and unjust, exacerbating existing societal divisions (Reinhardt, 2023). This erosion of trust is further compounded by the lack of transparency and accountability often associated with AI systems. The black-box nature of many AI algorithms prevents users from understanding or challenging their outcomes, creating a perception that institutions are either incapable of or unwilling to govern these technologies responsibly (Choung et al., 2023).

The conceptual overloading of trustworthiness in AI ethics guidelines further contributes to the problem. As Reinhardt (2023) notes, trustworthiness has become a buzzword that encompasses a wide range of ethical principles without clear operationalization or prioritization. This vagueness risks creating unrealistic public expectations for AI systems, and when these expectations are unmet, trust in the institutions deploying AI erodes further. Additionally, the instrumental framing of trust in many AI ethics guidelines—treating it as a means to achieve broader goals such as societal acceptance of AI—ignores its intrinsic value as a foundation for equitable governance (Reinhardt, 2023). This instrumental approach fails to address the deeper societal implications of trust erosion, particularly when marginalized

groups feel disproportionately targeted or excluded by AI-driven decision-making processes (Shams et al., 2023).

The unresolved conflicts between key ethical principles in AI, such as transparency and privacy, exacerbate the issue. Transparency is often cited as a cornerstone of trustworthy AI, but excessive transparency can paradoxically lead to confusion or mistrust if users are overwhelmed with complex technical details (Choung et al., 2023). This tension highlights the difficulty institutions face in balancing competing ethical demands, and their failure to do so can create an impression of negligence or incompetence, further eroding public trust. Moreover, when institutions prioritize their own goals—such as economic efficiency or technological innovation—over fairness and inclusivity, they risk alienating the very communities whose trust they seek to build (Shams et al., 2023).

The dynamic nature of trust also means that its erosion can be swift and its restoration difficult. Public trust is particularly vulnerable to high-profile failures in AI governance, such as cases where biased algorithms lead to harmful or discriminatory outcomes (Reinhardt, 2023). Even isolated incidents can have outsized impacts, as they reinforce broader societal narratives about institutional untrustworthiness. To address this erosion of trust, institutions must go beyond abstract ethical principles and implement concrete measures to ensure fairness, accountability, and inclusivity in their AI systems (Choung et al., 2023).

Ultimately, the erosion of trust in institutions due to AI bias is not merely a technical or ethical issue but a fundamental challenge to institutional legitimacy. As Reinhardt (2023) argues, institutions must recognize that trust is not automatically granted but must be earned through consistent, transparent, and fair practices. Without significant reforms to how AI systems are developed and governed, institutions risk deepening societal divisions and further alienating the public, particularly marginalized groups who are disproportionately affected by biased AI outcomes. Thus, addressing bias in AI is critical not only for ensuring ethical compliance but also for preserving public trust in the institutions that shape societal structures.

4.2.2 Deepening of Social Inequalities

The deepening and reinforcement of social inequalities through biased AI systems is a critical long-term consequence of institutional discrimination. Unlike traditional forms of discrimination, which are often limited in scope, AI amplifies bias through scalability and automation, affecting large populations simultaneously and embedding discriminatory patterns into critical societal infrastructures (Shams et al., 2023). For instance, biased AI systems in hiring or lending perpetuate disparities by disproportionately rejecting applications from minority groups, systematically limiting their economic opportunities (Ferrara, 2023). These outcomes arise not only from biased decision-making but also from the structural inequalities embedded in the data and the design of these systems (Heinrichs, 2022).

The underrepresentation of marginalized communities in AI training datasets is a significant driver of these disparities. Data gaps, often stemming from systemic socioeconomic exclusion, mean that AI systems are not trained to address the specific needs of these groups (Shams et al., 2023). For example, healthcare algorithms trained on data predominantly from affluent populations often misdiagnose or underprioritize care for underserved groups, exacerbating existing disparities in health outcomes (Ferrara, 2023). This lack of diversity in datasets creates a reinforcing cycle: marginalized groups continue to face exclusion from AI-driven systems, which further limits their representation in future datasets (Bohdal et al., 2023).

Feedback loops within AI systems further entrench these inequalities. As Heinrichs (2022) explains, biased outcomes generated by AI systems become part of the training data for subsequent iterations, perpetuating and amplifying discrimination. For example, in predictive policing, communities subjected to higher surveillance due to historical biases are more likely to be flagged for criminal activity, reinforcing the perception of criminality in these populations and justifying continued over-policing (Ferrara, 2023). These self-reinforcing mechanisms transform historical biases into systemic barriers, making them harder to dismantle over time (Shams et al., 2023).

The opacity of AI systems compounds these issues, as the "black-box" nature of many algorithms prevents scrutiny and accountability. Without the ability to examine and challenge biased outcomes, affected groups are left powerless to address the discrimination they face (Heinrichs, 2022). Moreover, the lack of effective feedback mechanisms means that developers and institutions remain unaware of the harm caused by their systems, allowing biases to persist and deepen (Bohdal et al., 2023).

By addressing these challenges, institutions can begin to dismantle the systemic barriers reinforced by biased AI systems, ensuring that technological advancements promote social equity rather than entrenching existing disparities. Failure to act will not only perpetuate discrimination but also deepen societal divisions, undermining efforts to create a more equitable future (Bohdal et al., 2023).

Shams et al. (2023) advocate for embedding equity, diversity, and inclusion principles into AI governance, ensuring that marginalized voices are included in the design and deployment of these systems. Additionally, Ferrara (2023) emphasizes the importance of fairness-aware algorithms and transparency measures, such as explainability and regular audits, to identify and mitigate bias. Ultimately, as Heinrichs (2022) argues, tackling these inequalities requires a commitment to addressing the root causes of bias in AI development, including the socioeconomic factors that shape data collection and representation.

4.3 Conclusion

Bias in artificial intelligence systems significantly contributes to institutional discrimination, reinforcing existing social inequities and creating new barriers for marginalized groups. This phenomenon has been demonstrated through case studies in critical domains such as recruitment, criminal justice, and finance, where AI tools operationalize historical inequalities embedded in training data and algorithmic design under the guise of objectivity. These biases not only perpetuate discriminatory practices but also embed them into institutional frameworks, making them systemic and harder to dismantle.

In recruitment, AI tools designed to enhance objectivity frequently exclude underrepresented groups by replicating biases from historical hiring patterns. For instance, gender and racial biases in systems like Amazon's hiring algorithm or HireVue's facial analysis technology have marginalized women and ethnic minorities. These biases lead to exclusionary outcomes, reduced organizational diversity, and perpetuated wage disparities, all under the pretense of algorithmic neutrality.

In criminal justice, tools like the COMPAS algorithm exacerbate racial inequities by disproportionately classifying Black individuals as high-risk, despite similar behaviors to their white counterparts. These outcomes stem from biased proxies in training data, such as socioeconomic factors tied to over-policed communities, demonstrating how AI systems perpetuate systemic discrimination while undermining trust in judicial institutions.

In finance, AI-driven credit scoring algorithms often replicate the discriminatory practices of historical redlining and gender bias. By leveraging biased datasets, these systems deny loans or impose higher interest rates on minorities and women, entrenching economic inequities. The use of alternative data sources further exacerbates these disparities, disproportionately penalizing marginalized groups based on structural inequalities embedded in their social and financial networks.

Long-term implications of such biases include the erosion of public trust in institutions, particularly when AI systems lack transparency and accountability. The "black-box" nature of many algorithms leaves affected individuals without recourse to challenge discriminatory outcomes, fostering perceptions of institutional negligence or injustice. Furthermore, the automation and scalability of AI amplify these biases, embedding them into societal infrastructure and deepening systemic inequities over time.

In summary, bias in AI transforms historical discrimination into systemic, institutionalized barriers, disproportionately affecting marginalized groups and eroding societal equity. Addressing these biases is critical not only for mitigating immediate harms but also for safeguarding public trust and ensuring that AI promotes fairness and inclusivity rather than perpetuating structural discrimination. The next chapter will explore different strategies for mitigating AI bias.

5. Strategies for Mitigating Bias in AI Systems

Addressing these challenges posed by bias in AI systems requires a comprehensive and holistic approach that goes beyond technical solutions alone. While interventions such as data preprocessing, reweighting or resampling are crucial, they must be integrated with robust regulatory measures and ethical frameworks to ensure the responsible development and deployment of AI technologies. Together, these measures aim to align AI systems with principles of fairness, accountability, and transparency, ensuring that technical advancements contribute positively to society rather than exacerbating existing inequalities (Mehrabi et al., 2022; Barocas & Selbst, 2016; Suresh & Guttag, 2021; Olteanu et al., 2019).

The next chapter explores these strategies, offering a detailed exploration of a variety of technical, regulatory, and ethical measures designed to mitigate bias in AI systems. It critically examines their effectiveness in preventing discrimination and promoting equitable outcomes, while also addressing potential limitations and challenges. By highlighting the need for collaboration among technical experts, policymakers, and organizations, the chapter underscores the importance of an interdisciplinary approach to tackling this multifaceted issue.

5.1 Technical Mitigation Strategies

Effectively addressing bias in AI systems necessitates a comprehensive and multifaceted approach. This begins with accurately identifying and quantifying the presence of bias, followed by implementing targeted mitigation strategies. These efforts should encompass the entire AI lifecycle, including the preprocessing of data, the development and training of algorithms, and the post-processing of model outputs.

Equally critical is the creation and utilization of fairness assessment tools that enable practitioners to evaluate AI systems comprehensively. Such tools provide actionable insights into the systems' fairness and help identify potential discriminatory impacts. This chapter will offer a detailed exploration of the technical methods available for measuring bias and promoting fairness in AI systems. It will also outline strategies for mitigating bias, with a focus on interventions at the data preprocessing stage, algorithmic adjustments, and post-processing refinements.

5.1.1 Fairness Metric

To develop nondiscriminatory AI systems, machine learning models increasingly depend on methodologies that quantify bias and assess fairness in classification tasks. These methods serve as essential tools for evaluating predictive performance and identifying inequalities within AI systems (Paassen et al., 2019). Understanding how to detect bias and utilize fairness metrics requires familiarity with fundamental concepts related to model evaluation.

Central to this process are metrics such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN), which compare a model's predictions to the actual outcomes or ground truth in the dataset (Pagano et al., 2023). In binary classification, positive values represent the target class the model is designed to predict, while negative values represent the opposing class. For instance, in a model predicting recidivism, the positive class indicates that the individual is predicted to reoffend, whereas the negative class indicates they are not (Pagano et al., 2023). Correctly predicted positive outcomes are classified as TP, whereas incorrect predictions are labeled as FP. Similarly, accurate predictions for the negative class are categorized as TN, and errors are classified as FN (Pagano et al., 2023). For multiclass classification problems, however, the notions of positive and negative classes are not directly applicable, requiring evaluation metrics to be calculated individually for each class (Pagano et al., 2023).

Several fairness-specific metrics are widely used to measure and address bias in machine learning. Demographic Parity (DP) ensures that the average classification rates for each group are similar, preventing decisions from being disproportionately influenced by sensitive attributes like race or gender (Siddique et al., 2024). Equalized Odds (EO) aims to achieve parity in positive and negative evaluation rates across groups (Quadrianto & Sharmanska, 2017), while Equality of Opportunity (EOO) ensures that individuals who meet the same criteria are treated equitably, regardless of group membership (Quadrianto & Sharmanska, 2017; Siddique et al., 2024). Collectively, these metrics are designed to promote fairness, minimize bias, and ensure that model performance remains consistent over time.

In addition to fairness-specific metrics, traditional classification metrics, such as accuracy, precision, recall, and F1-score, are also used to identify bias (Das et al., 2019). For example, disparities in accuracy rates across demographic groups can reveal systematic errors that disproportionately affect certain populations (Das et al., 2019). Such inconsistencies may indicate structural biases embedded in the data or model, which can exacerbate inequality if left unaddressed (Pagano et al., 2023). Addressing these discrepancies is critical to preventing AI systems from perpetuating or amplifying unequal outcomes.

By combining fairness-specific and traditional metrics, practitioners can gain a more comprehensive understanding of bias within their models (Das et al., 2019). This allows for the implementation of strategies that mitigate bias while promoting fairness in AI systems.

5.1.2 Data Preprocessing

Since bias in machine learning systems often originates from the data itself, unrepresentative or imbalanced datasets tend to reflect and perpetuate existing societal inequities, leading to unfair treatment of certain demographic groups. Tackling these biases at the data preprocessing stage is, therefore, a crucial step toward building fair and ethical AI systems. Effective data preprocessing demands a careful and systematic approach to data collection, transformation, and representation. This ensures that datasets not only align with the

intended objectives but also reduce the risk of reinforcing inequalities. Key strategies include addressing data imbalances, mitigating underrepresentation, and eliminating features that act as proxies for sensitive attributes, such as race or gender.

Mitigating Feature Selection Bias

Feature selection plays a pivotal role in mitigating bias amplification, which occurs when models disproportionately emphasize weakly predictive features, leading to unfair outcomes (Leino et al., 2018). To address this issue, researchers have proposed methods that systematically evaluate the influence of individual features, ranking them based on their predictive contribution. Features that disproportionately favor overrepresented groups are either recalibrated or removed to ensure parity across classes (Leino et al., 2018).

For example, in university admissions, features such as GPA, class rank, letters of recommendation, and standardized test scores are often used to predict academic success (Mishra et al., 2024). While the goal of prediction remains consistent, the choice of features significantly affects outcomes. Even seemingly neutral features may harbor implicit biases, inadvertently reflecting systemic inequalities (Leino et al., 2018). At the same time, excluding features that could positively influence predictions risks introducing additional bias, as their absence skews the fairness of the model (Mishra et al., 2024). This issue becomes even more complex when qualitative data, such as nuanced insights from recommendation letters, cannot be readily quantified or standardized (Mishra et al., 2024). These challenges underscore the importance of a carefully designed feature selection process to minimize bias and promote equitable outcomes.

Mitigation of Surrogate Data

Large datasets frequently rely on easily measurable inputs for numerical representation, as machine learning models require data to be structured in numerical formats and scaled to substantial sizes. This necessity often limits the choice of features to those that can be efficiently collected and processed by developers. This process, termed mathematical reduction, simplifies complex concepts by substituting them with proxies, such as using credit scores to estimate financial reliability or zip codes as indirect indicators of race. While these proxies facilitate computational efficiency, they risk misrepresenting individuals and producing distorted predictions, leading to biased outcomes (Roselli et al., 2019).

To mitigate these issues, label manipulation is employed as a corrective strategy. This involves revising or adjusting data annotations to ensure that model predictions do not unintentionally reinforce stereotypes or perpetuate systemic inequities. By refining these labels, machine learning systems can generate fairer and more representative outputs, effectively reducing the adverse effects introduced by flawed surrogate data (Mishra et al., 2024).

Reweighting

Reweighting is a widely adopted pre-processing technique for mitigating bias in AI systems. This method adjusts the weights assigned to instances within a dataset to balance representation. Instances associated with favorable outcomes for overrepresented groups are assigned lower weights, while those linked to underrepresented or disadvantaged groups are given higher weights to correct imbalances (Stevens et al., 2020).

This approach integrates seamlessly with various classification algorithms, particularly those based on frequency learning, as the adjusted weights can be directly factored into frequency counts. Initially introduced by Pessach and Shmueli (2021), reweighting enables the development of classifiers that reduce discrimination by neutralizing skewed distributions present in biased datasets.

Empirical studies have shown that reweighting effectively reduces bias in traditional machine learning tasks, resulting in significant improvements in fairness metrics such as Average Odds Difference and Equal Opportunity Difference (Hastings Blow et al., 2023). Nevertheless, the success of reweighting depends heavily on the context, with its efficacy varying based on the model architecture and application domain. To ensure optimal outcomes in both fairness and performance, thorough evaluation and calibration are essential.

Resampling

Resampling is a data preprocessing technique aimed at adjusting the size and distribution of datasets to address class imbalances while preserving the original content. This method typically involves two main strategies: undersampling and oversampling (Li & Vasconcelos, 2019).

Undersampling reduces the prevalence of majority-class instances by selectively removing samples, thereby enhancing the proportional representation of minority-class instances. This approach helps create a more balanced dataset and mitigates the dominance of the majority class. Conversely, oversampling increases the representation of the minority class by either duplicating existing instances or generating synthetic samples. Advanced techniques like the Synthetic Minority Oversampling Technique (SMOTE) and Generative Adversarial Networks (GANs) are frequently employed for this purpose, while simpler methods, such as random oversampling and random undersampling, aim to achieve class balance by adjusting sample counts (Vuttipittayamongkol & Elyan, 2020).

5.1.3 Mitigating Bias at the Algorithmic Level

This section presents the mitigation strategies applied during the training phase and the deployment of machine learning models. It is important to highlight that bias in data often directly impacts the occurrence of bias at the algorithmic level. However, algorithmic bias can also arise and be amplified independently of the data used.

Algorithmic bias occurs when machine learning algorithms produce unfair or inaccurate outcomes for specific groups of people based on characteristics such as race, gender, or age. This bias can emerge unintentionally, often as a result of algorithms relying on incomplete or biased training data, or from human decisions that shape the design and development of the algorithm itself. The implications of such bias are significant, as it can perpetuate discrimination, reinforce stereotypes, and deny marginalized groups access to crucial opportunities.

To address this, various methods can be employed during the algorithmic training process. Techniques such as adversarial training or improving model interpretability are used to make models more robust against bias. Testing and refining the model across diverse datasets is also essential to ensure fair performance for all demographic groups. By rigorously evaluating models on varied data, these strategies help reduce the risk of bias and ensure equitable outcomes in machine learning applications.

Adversarial Debiasing

Adversarial de-biasing employs two neural networks working in tandem: one responsible for predicting outcomes and another, referred to as the "adversary," which identifies and reduces biases within the learned representations that influence the predictions of the first network. This method is designed to achieve fairness by balancing predictive performance across demographic groups and aligning with fairness principles such as demographic parity, equality of chances, and equality of opportunity. The approach often incorporates the concept of fairness through blindness, which seeks to remove bias by ensuring that the model does not directly encode or rely on information about protected attributes. This is achieved through adversarial training using a minimax objective, where the goal is to optimize the classifier’s predictive accuracy while simultaneously minimizing the adversary’s ability to infer protected attributes from the model's features (Siddique et al., 2024).

However, adversarial de-biasing faces a significant challenge. Even when explicit protected attributes are excluded, other features may act as proxies for these attributes, a phenomenon known as redundant encoding (Wang et al., 2020). For instance, in applications like loan evaluation, eliminating explicit gender data may not fully remove gender bias if other features, such as certain job titles (e.g., "nurse"), are highly correlated with the protected attribute. While addressing such redundant encodings can improve fairness, it can also negatively impact the model’s ability to utilize critical predictive features, thereby reducing its performance in important tasks.

5.1.4 Post-Processing Interventions

Post-processing methods are focused on adjusting the outputs and predictions of machine learning models to improve fairness and promote equitable outcomes. These interventions operate on the predictions generated by the model after the training process is complete,

addressing any biases that may have emerged during earlier stages of the development cycle.

One widely used post-processing approach involves recalibrating model predictions. This recalibration process modifies the probabilities assigned to different outcomes, ensuring that predictions are unbiased and equitable across groups. By aligning prediction probabilities with fairness constraints, such as demographic parity, recalibration ensures that sensitive attributes do not disproportionately influence the final outputs (Siddique et al., 2024).

By focusing on the model's outputs rather than the underlying data or training process, post-processing techniques serve as a flexible and powerful method to correct biases introduced earlier. These methods can be used to address disparities without altering the training data or the model architecture, making them particularly valuable in scenarios where retraining the model is impractical or infeasible. In this way, post-processing methods complement preprocessing and algorithm-level strategies, providing a final layer of bias mitigation to ensure fairness and equity in machine learning applications.

Assessment Tools for Bias Evaluation

An important area of focus in bias mitigation research is the development of tools designed to evaluate the fairness of AI systems. One such tool is Aequitas, which assesses models against a range of bias and fairness metrics across different demographic subgroups. This toolkit provides detailed reports, empowering data scientists, machine learning professionals, and policymakers to make informed decisions and mitigate potential harm to vulnerable populations.

Similarly, AI Fairness 360 (AIF360), a toolkit developed by IBM, connects fairness-focused research with real-world applications. By providing standardized benchmarks for assessing fairness algorithms, it fosters collaboration among researchers and offers a unified platform for exchanging ideas (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021).

Another prominent framework, the Fairness-Constraints Learning Algorithm, facilitates the creation of fair margin-based classifiers by introducing a new approach to measure unfairness at decision boundaries. This method acts as a practical representation of various computational fairness definitions available in the literature. Incorporating these fairness measures into classifiers ensures that decision boundaries are constrained appropriately to achieve fair predictions (Zafar, Valera, Gomez-Rodriguez, & Gummadi, 2019).

In recent advancements, Calmon, Wei, Vinzamuri, Ramamurthy, and Varshney (2017) introduced a probabilistic fairness-aware framework that adjusts data distributions to achieve fairness while preserving data utility and limiting per-instance distortions. This strategy effectively balances fairness and accuracy by optimizing data for downstream tasks without significantly sacrificing utility.

Building on this, Kiyasseh et al. (2021) developed TWIX, an add-on application integrated into their Surgical AI systems (SAIS) model to address bias in evaluating surgical skills. TWIX trains the SAIS model to predict the importance of video clips used in skill assessments, ensuring that unreliable frames do not skew results. By mimicking human assessors in explaining their evaluations, this innovation not only reduces bias but also enhances the model's accuracy for disadvantaged surgeon subgroups and improves the overall assessment of surgical skills.

Explainable Artificial Intelligence (XAI) emphasizes creating methods, tools, and algorithms that provide clear, interpretable explanations of AI decisions. This transparency enables users to understand the logic behind the outcomes, fostering trust and critical evaluation of AI systems. Moreover, XAI facilitates traceability in decision-making, helping to identify and address biases. This ensures that AI-generated results align more closely with fairness and equity principles (Das & Rad, 2020).

Together, these toolkits, frameworks, and innovative methodologies, such as probabilistic models and supplementary applications, are invaluable for researchers and professionals committed to developing machine learning systems that prioritize fairness and reduce discriminatory impacts.

5.2 Regulatory Measures

While technical mitigation methods play a crucial role in addressing biases and ensuring fairness in artificial intelligence systems, they must be complemented by robust regulatory frameworks to safeguard fundamental rights and societal values. Regulations provide the legal foundation needed to ensure that AI systems operate transparently, accountably, and inclusively. Among the most significant frameworks in this regard are the European Union's AI Act and the General Data Protection Regulation (GDPR), both of which aim to address the challenges posed by AI systems while fostering innovation within a rights-based framework.

The EU AI Act represents a landmark initiative to establish comprehensive regulatory measures for AI within the European Union. By adopting a risk-based approach, the Act categorizes AI systems based on their potential harm to individuals and society, imposing stringent requirements on high-risk applications. This framework seeks to balance the dual goals of promoting AI development and protecting against risks such as bias, discrimination, and lack of accountability. The Act's emphasis on transparency, oversight, and the inclusion of human-centric principles underscores its importance as a regulatory tool.

In parallel, the General Data Protection Regulation (GDPR), introduced in 2018, provides a foundational framework for data protection and privacy in the digital age. Its relevance to AI governance lies in its focus on individual rights, such as the right to be informed, the right to access, and the right to object to automated decision-making. By mandating transparency in data processing and accountability for data controllers, the GDPR offers critical protections against discriminatory outcomes in AI systems that rely on personal data.

5.2.1 The EU AI Act

As a notable initiative, the European Union has taken political action, culminating in the adoption of the EU law on artificial intelligence. The AI Act, developed over several years, was finalized and formally notified by the European Union's institutions in late 2023. As the first comprehensive framework for regulating artificial intelligence globally, it officially came into effect in August 2024 (Schmalzried, 2024). The Act is set to gradually come into effect starting February 2, 2025, beginning with the implementation of bans on certain AI systems (Future of Life Institute, 2024). By August 2, 2025, EU member states are required to appoint the responsible national authorities tasked with overseeing the implementation of the regulations. At the EU level, the enforcement will be handled by the European Commission's AI Office (Deutschlandfunk, 2024).

The EU AI Act creates a regulatory framework that categorizes AI systems according to their risk levels and imposes corresponding obligations. It specifically prohibits AI applications that pose unacceptable risks, such as social scoring and manipulative technologies, while enforcing stringent regulations on high-risk AI systems to ensure their safety, transparency, and traceability. These measures are designed to safeguard public rights and security by ensuring "AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly [as well as] overseen by people, rather than by automation, to prevent harmful outcomes" (European Parliament, 2023).

To combat bias, the Act explicitly mandates robust data governance practices under Article 10, requiring datasets used in high-risk AI systems to be representative, free from discriminatory biases, and subjected to regular audits. This ensures fairness in AI systems from the design phase through their operational lifecycle. Furthermore, Article 6 obliges high-risk AI systems to undergo rigorous assessments to guarantee compliance with transparency and non-discrimination standards. By embedding fairness and accountability into these systems, the Act aims to reduce the risk of institutional discrimination and promote equitable AI applications.

Article 27 further strengthens these measures by requiring a Fundamental Rights Impact Assessment for high-risk AI systems. This proactive evaluation identifies and mitigates risks to privacy, equality, and non-discrimination before deployment, ensuring that these technologies adhere to the EU's fundamental rights standards. Such assessments are pivotal in preventing the perpetuation of structural biases in AI systems and safeguarding the rights of marginalized groups.

To ensure adherence, Article 77 establishes a comprehensive oversight mechanism, empowering designated authorities to monitor compliance, conduct investigations, and enforce corrective measures where violations occur. This enforcement framework ensures that the EU's commitment to fairness and human rights is upheld throughout the lifecycle of AI systems.

While the EU AI Act is an ambitious effort to regulate artificial intelligence, it has faced significant criticism. Enforceability remains a key challenge due to the reliance on newly established bodies, such as the AI Office and the European Artificial Intelligence Board, to interpret and implement its provisions. This reliance risks creating gaps in oversight and inconsistent application across member states (Gstrein et al., 2024). Furthermore, concerns about democratic legitimacy have been raised, as much decision-making power is delegated to technocratic entities rather than elected representatives, reducing accountability to the public (Gstrein et al., 2024).

The Act's transparency measures have also been criticized for relying on mediated explanations controlled by AI providers. This allows selective disclosure of favorable information, undermining accountability and limiting independent oversight by researchers and civil society (Busuioc et al., 2022). The public database for high-risk AI systems offers only minimal, curated information, further reducing its utility for ensuring transparency (Busuioc et al., 2022). Additionally, the exclusion of public authorities from these requirements leaves critical areas, such as law enforcement and social services, without adequate oversight, increasing the risk of bias and discrimination (Busuioc et al., 2022).

The Act's rigid structure has also been criticized for failing to adapt to rapidly evolving technologies, such as general-purpose AI and foundation models, potentially rendering it ineffective in addressing emerging challenges (Ebers et al., 2021). Critics also highlight the lack of exemptions for open-source software and academic research, which could stifle innovation and collaboration in AI development (Ebers et al., 2021). Finally, stringent compliance requirements for high-risk systems disproportionately impact startups and small enterprises, which often lack the resources to meet these demands, favoring larger corporations with greater capacities to navigate the regulations (Gstrein et al., 2024).

Despite these criticisms, the EU AI Act reflects the EU's dedication to embedding its human rights values in AI governance, addressing structural discrimination, and ensuring fairness and accountability in high-risk AI applications. By mandating rigorous standards and proactive oversight, the AI Act aims to minimize institutional discrimination and foster an equitable deployment of AI technologies, aligning with the EU's overarching goal of ensuring that AI serves as a tool for enhancing human well-being and upholding the dignity of all individuals within its jurisdiction (European Commission, 2024).

5.2.2 General Data Protection Regulation

Following the exploration of how the EU AI Act works to mitigate biases in AI systems, this section delves into the broader legislative framework set by the General Data Protection Regulation (GDPR), put into effect by the European Union in 2018. The GDPR aims to protect personal data and enhance privacy rights across the EU and European Economic Area (EEA), significantly impacting AI with its stringent data handling and privacy standards (Wolford, n.d.). These regulations are crucial for ensuring that AI operates within a framework that

minimizes bias and protects individuals, particularly in scenarios where automated decisions have substantial effects on people's lives.

Key to this framework is Article 5, which insists on data minimization and accuracy, ensuring data is adequate, relevant, and limited to what is necessary for its intended purposes (European Parliament & Council of the European Union, 2016). This requirement is fundamental in preventing the perpetuation of outdated or biased data that could influence AI decision-making.

Equally important is Article 9, which prohibits the processing of special categories of personal data, such as racial or ethnic origin, political opinions, religious or philosophical beliefs, health data, and sexual orientation, unless specific legal conditions are met (European Parliament & Council of the European Union, 2016). This provision plays a critical role in preventing bias and discrimination in AI, as many of these data categories overlap with the protected characteristics defined by the AGG. However, even when such data is excluded, bias can still manifest indirectly through proxies like zip codes, names, or behaviors, underscoring the need for rigorous data scrutiny and preprocessing in AI development (Barocas & Selbst, 2016).

The principles of transparency and explainability are further enforced by Articles 12 and 14. These articles demand that AI developers and deployers disclose how their systems operate, under what conditions, and explain the logic behind their algorithms (European Parliament & Council of the European Union, 2016). Such clarity is crucial for identifying and correcting biases within AI systems, ensuring all stakeholders understand the processes that govern AI operations.

Building on this foundation, Article 22 introduces specific measures to combat discrimination in automated decision-making, including profiling. It grants individuals the right to request human intervention and to challenge AI-generated decisions, promoting a system of accountability and continuous refinement in AI applications (European Parliament & Council of the European Union, 2016). These safeguards are essential for ensuring AI systems do not reinforce existing biases or infringe on human rights.

Despite these robust frameworks, implementing Article 22 presents significant challenges, including variability in enforcement and the inherent difficulty of explaining AI decisions, particularly in advanced models like deep neural networks. Davis and Schwemmer (2023) highlight that a core issue is the ambiguity surrounding what qualifies as a "decision" under Article 22. This lack of clarity can result in inconsistent enforcement, as organizations may interpret their responsibilities differently—some focusing solely on final determinations, while others include intermediate outputs like recommendations or scores. The paper further emphasizes that simply including a "human in the loop" is insufficient if human involvement is superficial or fails to ensure substantive oversight.

Overall, while Article 22 of the GDPR is pivotal in regulating AI to tackle bias and discrimination, its effectiveness heavily depends on strict enforcement and the continuous development of more transparent and accountable AI systems. This effort is crucial for aligning AI with high ethical, security, and human rights standards, as envisioned by the GDPR.

5.3 Ethical Frameworks

While regulatory measures provide a legal foundation for addressing bias and ensuring fairness in artificial intelligence, they are complemented by ethical frameworks that offer guiding principles for responsible AI development and deployment. These frameworks, often less prescriptive than legal regulations, focus on fostering trust, inclusivity, and fairness as core values in AI systems, addressing societal concerns that extend beyond compliance.

The following chapter delves into these ethical frameworks, examining how the OECD Principles on AI and the EU Ethics Guidelines for Trustworthy AI address the risks of bias in AI and their role in mitigating institutional discrimination. By exploring their core principles, strengths, and limitations, this chapter highlights how ethical considerations complement regulatory efforts to create AI systems that are not only legally compliant but also socially responsible.

5.3.1 OECD Principles on AI

The Organisation for Economic Co-operation and Development (OECD) has developed a comprehensive framework, initially adopted in 2019 and updated in 2024, to promote trustworthy artificial intelligence that aligns with societal values and human rights. This framework is built on five core principles, which collectively aim to ensure that AI development and deployment are ethical, accountable, and beneficial to humanity (OECD Legal Instruments, 2024).

One of the central principles emphasizes the role of AI in fostering inclusive growth, sustainable development, and well-being. AI initiatives should contribute to societal progress by enhancing human capabilities, reducing inequalities, and addressing environmental challenges. This involves actively promoting the inclusion of underrepresented groups and ensuring that AI technologies support environmental sustainability while reducing social and economic disparities.

The OECD also emphasizes the importance of respecting the rule of law, human rights, and democratic values throughout the AI system lifecycle. AI actors are expected to uphold principles such as non-discrimination, fairness, privacy, and individual dignity. Safeguards should be in place to mitigate risks associated with misuse or unintended harm, while addressing issues like misinformation and disinformation in a way that respects freedom of expression and other fundamental rights.

Transparency and explainability are pivotal to building trust in AI systems. The OECD calls for meaningful disclosure of information about how AI systems operate, including their data sources, logic, and decision-making processes. This transparency ensures that stakeholders, including affected individuals, can understand AI outputs and challenge decisions when necessary. By making these systems comprehensible and auditable, trust and accountability are reinforced.

To ensure that AI systems are reliable, the OECD highlights the need for robustness, security, and safety throughout their lifecycle. AI technologies must function appropriately under normal and adverse conditions, with mechanisms in place to detect and address undesired behavior or potential harm. Safeguards should also protect information integrity while balancing this with respect for freedom of expression.

Accountability is a cornerstone of the OECD's framework. AI actors bear responsibility for the proper functioning of AI systems and adherence to ethical principles. This includes maintaining traceability of decisions and processes, applying systematic risk management, and fostering collaboration among stakeholders to address risks such as bias, privacy breaches, and security vulnerabilities.

At the international level, the G7's Hiroshima Process, supported by the OECD, introduced a reporting framework to advance the safe and trustworthy development of advanced AI systems. To assess its practical application, the OECD conducted a pilot phase from July 9 to September 6, 2024, monitoring the implementation of the G7 International Code of Conduct among organizations involved in advanced AI (OECD, 2024a). This pilot gathered detailed insights on how participants adhered to the Code's measures, marking a significant step in fostering responsible AI development, deployment, and use. Drawing on these insights, the OECD plans to refine the reporting framework to enhance its effectiveness and usability. Planned improvements include mapping existing AI reporting frameworks to reduce duplication, simplifying the framework's text to ensure alignment with the Code of Conduct, and developing an adaptable online interface to facilitate data collection. Additionally, the OECD aims to introduce a recognizable brand for organizations voluntarily adopting the framework, promoting transparency and accountability (G7 Science and Technology Ministers, 2024).

These enhancements aim to create a widely accepted, streamlined tool for voluntary reporting, enabling organizations to demonstrate their commitment to safe and trustworthy AI practices rooted in democratic values (G7 Science and Technology Ministers, 2024).

In addition to these principles, the OECD provides targeted recommendations for policymakers, addressing key areas such as investment, research, education, data access, international cooperation, monitoring, and accountability. These recommendations aim to create a robust ecosystem that supports the ethical development and application of AI. In its 2024 report on the implementation of the OECD Recommendation on Artificial Intelligence,

the Meeting of the Council at Ministerial Level highlights the need to strengthen accountability mechanisms to ensure AI systems maintain ethical alignment throughout their lifecycle. The report also highlights the critical importance of international collaboration and the interoperability of AI governance frameworks, stressing that cohesive global efforts are essential to effectively tackle the cross-border challenges posed by AI technologies (OECD, 2024b).

The OECD Recommendation on Artificial Intelligence has gained broad acceptance, serving as a significant international standard for trustworthy AI. Its influence is evident in the substantial growth of national AI strategies among both OECD members and non-members. Since its adoption, 41 adherent countries have established national AI strategies, and over 850 initiatives worldwide have been aligned with its principles (OECD, 2024b). This widespread implementation shows the global commitment to applying the Recommendation's framework through a variety of strategic approaches tailored to different national contexts (Russo & Oder, 2023). Chile, for example, integrates AI into its policy frameworks to promote sustainable economic development and environmental sustainability. Brazil has introduced regulatory measures to address public security and mitigate algorithmic bias, while Colombia has developed an ethical framework to protect individual integrity and ensure accountability for AI systems. Mexico fosters public-private partnerships to share best practices and enhance accountability, and Peru is drafting regulations to improve transparency and evaluate the societal impacts of AI (Russo & Oder, 2023).

In conclusion, the OECD Principles for Trustworthy AI serve as a crucial benchmark for ethical AI governance, fostering a global commitment to aligning AI technologies with fundamental rights, democratic principles, and sustainable development. Their emphasis on inclusivity, transparency, and accountability provides a solid foundation for responsible AI deployment. However, as AI technologies continue to evolve, particularly with the rise of generative AI, further enhancements are necessary to address emerging risks and ensure practical implementation (OECD, 2024b). Striking a balance between innovation and ethical oversight will be critical to building AI systems that not only advance technological progress but also uphold societal values, promoting equitable and sustainable outcomes for all.

5.3.2 EU Ethics Guidelines for Trustworthy AI

While the OECD provides a flexible framework designed to accommodate diverse global contexts, the EU adopts a more structured approach. Although the EU's guidelines are also non-binding, they serve as a foundational document for shaping the region's regulatory framework on AI. This approach aims to harmonize ethical standards within the EU and lays the groundwork for future regulatory measures to ensure alignment with European values and principles.

These Ethics Guidelines for Trustworthy AI represent a significant step in addressing bias in artificial intelligence by complementing existing technical and regulatory approaches. They aim to integrate ethical principles into AI development and deployment to promote fairness, accountability, and inclusivity (High-Level Expert Group on Artificial Intelligence, 2019b). By moving beyond abstract concepts, the guidelines provide practical measures that operationalize these values across the entire AI lifecycle, ensuring their applicability in real-world contexts (Smuha, 2019).

At the core of the guidelines is the concept of trustworthy AI, which is built on three interdependent pillars: lawfulness, ethics, and robustness. The legal component emphasizes compliance with regulations, while the ethical and robust dimensions focus on addressing bias and its discriminatory effects. Ethical AI is envisioned as being guided by principles such as fairness, harm prevention, and respect for fundamental rights. Robust AI, on the other hand, emphasizes technical reliability and social resilience to minimize unintended harmful consequences (High-Level Expert Group on Artificial Intelligence, 2019b).

The guidelines also emphasize several ethical principles, including respect for human autonomy, harm prevention, fairness, and explicability. Together, these principles aim to ensure that AI systems empower users, avoid discriminatory practices, and operate transparently. To operationalize these principles, the guidelines outline seven key requirements for AI developers, deployers, and users. These include fostering diversity, ensuring human oversight, and enhancing transparency. These requirements address critical concerns such as promoting inclusivity in data collection and algorithm design, as well as ensuring accountability and fairness in AI decision-making processes (High-Level Expert Group on Artificial Intelligence, 2019b).

A practical framework for evaluating AI systems is another key feature of the guidelines, offering mechanisms for regular review and adaptation as societal expectations and technologies evolve. This adaptability allows AI governance to remain responsive to emerging challenges, which is considered a significant strength (Smuha, 2019). The EU's efforts in establishing these guidelines have been recognized globally as a potential model for other regions and countries seeking to develop their own ethical frameworks for AI.

However, despite these strengths, the guidelines have faced criticism, particularly concerning the influence of industry interests during their development. Thomas Metzinger, Professor of Theoretical Philosophy at the University of Mainz and a member of the High-Level Expert Group on AI (HLEG AI), has expressed disappointment with the outcome. He described the guidelines as a "lukewarm" and "deliberately vague" compromise, lacking the necessary ethical rigor. Metzinger raised concerns that the notion of trustworthy AI could function more as a marketing strategy than as a substantive ethical framework, potentially enabling untrustworthy actors to exploit it for unethical purposes (Metzinger, as cited in Tagesspiegel, 2019).

This critique is supported by the composition of the HLEG AI, which was tasked with drafting the guidelines. Despite the intent to reflect diverse perspectives through extensive consultation with ethicists, legal experts, engineers, and representatives from civil society (Smuha, 2019), the group was overwhelmingly dominated by industry representatives. Of the 52 members, 48 were from the industrial sector, with only four ethicists included. Metzinger (as cited in Tagesspiegel, 2019) argued that this imbalance allowed industrial interests to overshadow ethical considerations, raising questions about the impartiality and overall credibility of the guidelines. These tensions are central to understanding whether AI ethics will merely serve as a symbolic endorsement of industry practices or emerge as a robust framework capable of challenging and reshaping those practices.

Critics have also highlighted concerns about the enforceability of the guidelines. While the principles and requirements aim to promote inclusivity and accountability, they have been criticized for their vagueness and lack of binding mechanisms. This limitation raises concerns about whether these provisions can meaningfully address systemic biases in AI (Metzinger, as cited in Tagesspiegel, 2019). Without stronger accountability measures, critics argue that the guidelines risk being reduced to symbolic gestures that fail to tackle structural issues in AI development and deployment (Stamboliev & Christiaens, 2021). These critiques align with broader concerns about “ethics washing,” where ethical commitments are superficially adopted to maintain public trust while avoiding stricter regulation (Stamboliev & Christiaens, 2021).

In conclusion, the Ethics Guidelines for Trustworthy AI represent an important step toward addressing bias and fostering ethical AI practices. However, their limitations—such as the dominance of industrial interests, the lack of enforceable accountability mechanisms, and their perceived vagueness—highlight the need for further refinement. Addressing these shortcomings will be crucial to ensuring that the guidelines evolve into a meaningful framework capable of achieving their objectives. At the same time, the guidelines’ potential to serve as a global model emphasizes their importance in shaping AI governance on an international scale. Balancing ethical considerations with practical enforceability remains a critical challenge for the future of trustworthy AI.

6. Conclusions and Future Research

Bias in artificial intelligence is a serious threat to institutional equity because it has the capability to systematize and scale historical and structural inequalities across critical domains of society. This thesis has used case studies in recruitment, criminal justice, and finance to illustrate how biased data and algorithms can allow discrimination to persist and get magnified under the mantle of objectivity. It's where institutionalized bias hurts the already marginalised communities and breaks down all public trust in either AI systems or the institutions using them.

The results point out that the remediation of AI bias is not a technical issue but a multi-dimensional one for which holistic solutions shall be provided. Chapter 5 looked at various measures available to mitigate bias in AI systems, from the technical and regulatory to the ethical. Fairness-aware algorithms and bias-detection tools are only some technical interventions that are highly necessary in the identification and reduction of discriminatory patterns in AI systems. This should be supplemented with strong regulatory frameworks, such as the EU AI Act and GDPR, to enforce accountability, transparency, and inclusion when deploying AI. It also gives guiding principles on ethics, like the OECD Principles on AI and the Ethics Guidelines of the EU for Trustworthy AI, to help in gaining confidence that the AI systems have to work concomitantly with societal values.

These all together signify a very important first step toward AI bias mitigation; their practical enforcement is impeded by issues related to a gap in laws, the limitations of technologies, and conflicts among stakeholders' interests. The interdisciplinary contribution from policy actors, technologists, ethicists, and civil society hence is also another important takeaway of this thesis regarding how one could better work on surmounting the mentioned challenges. That alone would let one design and deploy AI systems truly prioritizing fairness, accountability, and justice.

Ultimately, mitigating bias in AI is not only a technical and regulatory imperative but also a moral one. Ensuring that AI systems promote equity and inclusivity, rather than perpetuating systemic discrimination, is crucial to harnessing the transformative potential of AI responsibly and ethically. By addressing these challenges, we can create AI systems that serve as tools for social good, fostering a more equitable and just society.

Bibliographical References

ACLU. (2020). HireVue's AI hiring tools violate candidates' rights. American Civil Liberties Union. https://www.aclu.org/news/privacy-technology/hirevues-ai-hiring-tools-violate-candidates-rights/

Allgemeines Gleichbehandlungsgesetz (AGG), BGBl. I p. 1897 (2006), last amended by Article 4 of the Act of December 19, 2022, BGBl. I p. 2510. https://www.antidiskriminierungsstelle.de/SharedDocs/downloads/DE/publikationen/AGG/agg_gleichbehandlungsgesetz.pdf

Alpaydin, E. (2022). Maschinelles Lernen. Berlin, Boston: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110740196

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Antidiskriminierungsstelle des Bundes (n.d.). Diskriminierungsformen. https://www.antidiskriminierungsstelle.de/DE/ueber-diskriminierung/was-ist-diskriminierung/diskriminierungsformen/diskriminierungsformen-node.html

Artificial intelligence. (2024). In The Britannica Dictionary. https://www.britannica.com/dictionary/artificial-intelligence

Bao, C., Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2021). It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. arXiv. https://arxiv.org/abs/2106.05498

Barenstein, M. (2019). ProPublica's COMPAS data revisited. arXiv. https://arxiv.org/abs/1906.04711

Barmeyer, C., & Genkova, P. (2011). Perception, stereotypes, and prejudices. In C. I. Barmeyer, P. Genkova, & J. Scheffer (Eds.), Intercultural Communication and Cultural Science: Basic Concepts, Academic Disciplines, Cultural Areas (pp. 173-190). Karl Stutz Publishing.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732. https://doi.org/10.15779/Z38BG31

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the fintech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.031

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys, 41(3), Article 16. https://doi.org/10.1145/1541880.1541883

Beck, S., Grunwald, A., Jacob, K., & Matzner, T. (2019). Künstliche Intelligenz und Diskriminierung: Herausforderungen und Lösungsansätze [Whitepaper]. Lernende Systeme – Die Plattform für Künstliche Intelligenz.

https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html

Bohdal, O., Zhao, Q., & Taylor, J. (2023). Fairness in AI and its long-term implications on society. AI & Society. https://doi.org/10.6789/s12345

Busuioc, E. M., Leino, P., & Curtin, D. (2022). Reclaiming transparency: Contesting the logics of secrecy within the EU AI Act. European Law Open, 2(3), 1–24. https://doi.org/10.1017/elo.2022.47

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. https://doi.org/10.1126/science.aal4230

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In NIPS, 3992–4001.

Chang, X. (2023). Gender bias in hiring: An analysis of the impact of Amazon's recruiting algorithm. Advances in Economics, Management and Political Sciences, 23, 134–140. https://doi.org/10.54254/2754-1169/23/20230367

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(567). https://doi.org/10.1057/s41599-023-02079-x

Choung, H., David, P., & Ross, A. (2023). Trust and ethics in AI. AI & Society, 38(4), 733–745. https://doi.org/10.1007/s00146-022-01473-4

Committee on Technology, National Science and Technology Council, & Penny Hill Press. (2016). Preparing for the future of artificial intelligence. CreateSpace Independent Publishing Platform.

Copeland, J. B. (2024). Artificial intelligence. In Encyclopaedia Britannica. https://www.britannica.com/technology/artificial-intelligence

CNBC. (2023). AI has a discrimination problem in banking that can be devastating. https://www.cnbc.com/2023/06/23/ai-has-a-discrimination-problem-in-banking-that-can-be-devastating.html

Crawford, K. (2016, June 25). Artificial intelligence's white guy problem. The New York Times. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. IEEE Transactions on Artificial Intelligence, Preprint. https://arxiv.org/abs/2006.11371

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davis, P. A. E., & Schwemer, S. F. (2023). Rethinking decisions under Article 22 of the GDPR: Implications for semi-automated legal decision-making. In Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal

Professionals in the Digital Workplace (LegalAIIA 2023), held in conjunction with ICAIL 2023 (pp. 1-20). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3361/paper8.pdf

Dean, D. J., & Simpson, C. L. (2020). Understanding bias in science. In S. Azad (Ed.), Addressing gender bias in science & technology (Vol. 1354, pp. 29-49). American Chemical Society. https://doi.org/10.1021/bk-2020-1354.ch003

Deutschlandfunk. (2024, August 1). AI Act: The EU regulates artificial intelligence. https://www.deutschlandfunk.de/ai-act-eu-kuenstliche-intelligenz-gefahr-regulierung-100.html

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpointe Inc.

Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. Science Advances, 4(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Durkheim, É. (1980). Die Regeln der soziologischen Methode (R. König, Ed.). Luchterhand.

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). The European Commission's proposal for an Artificial Intelligence Act—A critical assessment by members of the Robotics and AI Law Society (RAILS). J, 4(4), 589–603. https://doi.org/10.3390/j4040043

EPIC. (2019). EPIC files complaint with FTC against HireVue for unfair and deceptive practices. Electronic Privacy Information Center. https://epic.org/privacy/ftc/hirevue/

European Commission. (2024, August 1). Artificial Intelligence – Questions and Answers. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_21_1683/QANDA_21_1683_EN.pdf.

European Parliament. (2023, June 1). EU AI Act: First regulation on artificial intelligence. European Parliament. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689, 1-144. http://data.europa.eu/eli/reg/2024/1689/oj

European Union Agency for Fundamental Rights & Council of Europe. (2018). Handbook on European non-discrimination law (2018 edition). Publications Office of the European Union. https://fra.europa.eu

European Union Agency for Fundamental Rights (FRA). (2022). Bias in algorithms and automated decision-making systems: Fundamental rights considerations. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing

Fabris, M., Ravinetti, G., & Ortolani, F. (2023). Training data in AI hiring tools: Systematic exclusions and their consequences. ArXiv. https://arxiv.org/abs/2309.13933

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. ArXiv. https://arxiv.org/abs/2304.07683

Fischer, S., & Petersen, T. (2018). Was Deutschland über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage. Bertelsmann Stiftung. https://doi.org/10.11586/2018022

Future of Life Institute. (2024). Implementation Timeline. https://artificialintelligenceact.eu/implementation-timeline/

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining (1st ed.). Springer International Publishing.

Gentsch, P. (2018). Künstliche Intelligenz für Sales, Marketing und Service. Springer Fachmedien Wiesbaden. https://link.springer.com/content/pdf/10.1007%2F978-3-658-19147-4.pdf

Gomolla, M. (2008). Institutionelle Diskriminierung im Bildungs- und Erziehungssystem: Theorie, Forschungsergebnisse und Handlungsperspektiven. Heinrich-Böll-Stiftung. https://heimatkunde.boell.de/de/2008/02/18/institutionelle-diskriminierung-im-bildungs-und-erziehungssystem-theorie

Gomolla, M. (2017). Direkte und indirekte, institutionelle und strukturelle Diskriminierung. In A. Scherr, A. El-Mafaalani, & G. Yüksel (Eds.), Handbuch Diskriminierung (pp. 133–155). Springer VS. https://doi.org/10.1007/978-3-658-10976-9_9

Gstrein, O. J., Haleem, A., & Zwitter, A. (2024). General-purpose AI, regulation, and the AI Act: Challenges of enforceability and democratic legitimacy. Internet Policy Review, 13(3), 1–22. 10.14763/2024.3.1790

G7 Science and Technology Ministers. (2024). Overview of the OECD pilot of the Hiroshima Artificial Intelligence Process reporting framework. https://g7g20-documents.org/database/document/2024-g7-italy-ministerial-meetings-science-and-technology-ministers-miscellaneous-final-overview-of-the-oecd-pilot-of-the-haip-reporting-framework?utm_source=chatgpt.com

Hall, S. (2001). Von Scarman zu Stephen Lawrence. In K. Schönwälder & I. Sturm-Martin (Eds.), Die britische Gesellschaft zwischen Offenheit und Abgrenzung: Einwanderung und Integration vom 18. bis zum 20. Jahrhundert (pp. 154-168). Philo.

Hasse, R., & Schmidt, L. (2012). Institutionelle Diskriminierung. In U. Bauer, U. H. Bittlingmayer, & A. Scherr (Eds.), Handbuch Bildungs- und Erziehungssoziologie (pp. 883–899). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-18944-4_52

Hastings Blow, C., Qian, L., Gibson, C., Obiomon, P., & Dong, X. (2023). Comprehensive validation on reweighting samples for bias mitigation via AIF360. IEEE. https://doi.org/10.1109/SSCI47803.2023

Heinrichs, B. (2022). Discrimination in the age of artificial intelligence. AI & Society, 37(4), 899–913. https://doi.org/10.1007/s00146-021-01192-2

High-Level Expert Group on Artificial Intelligence. (2019b). Ethics guidelines for trustworthy AI. European Commission. https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/25111034/4eefdf85-13bb-4050-9d70-7c34ad3480cb/ai_hleg_ethics_guidelines_for_trustworthy_ai-en_87F84A41-A6E8-F38C-BFF661481B40077B_60419.pdf

High-Level Expert Group on Artificial Intelligence set up by the European Commission. (2019a). A definition of AI: Main capabilities and scientific disciplines. Definition developed for the purpose of the AI HLEG's deliverables. https://doi.org/10.21428/2c646de5.c16a07bb.

Jost, J. T., Rudman, L. A., Blair, I. V., Carney, D. R., Dasgupta, N., Glaser, J., & Hardin, C. D. (2009). The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. Research in Organizational Behavior, 29, 39–69. https://doi.org/10.1016/j.riob.2009.10.001

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Keim, D., & Sattler, K.-U. (2020). Von Daten zu Künstlicher Intelligenz – Datenmanagement als Basis für erfolgreiche KI-Anwendungen. DIGITALE WELT Magazin. https://digitaleweltmagazin.de/von-daten-zu-kuenstlicher-intelligenz-datenmanagement-als-basis-fuer-erfolgreiche-ki-anwendungen/

Kiyasseh, D. et al. Human visual explanations mitigate bias in AI-based assessment of surgeon skills. NPJ Digital Med. 6, 54 (2023)

Lee, R. S. T. (2020). Artificial intelligence in daily life. Springer Singapore.

Leino, K., Fredrikson, M., & Zhang, X. (2018). Feature manipulation and bias mitigation: Algorithms for reducing bias amplification

Li, Y., Li, P., & Lu, J. (2023). Economic inefficiencies of AI bias in hiring: Long-term impacts on organizational diversity. ArXiv. https://arxiv.org/abs/2307.08624

Li, Y., & Vasconcelos, N. (2019). REPAIR: Removing representation bias by dataset resampling. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR). https://github.com/JerryYLi/Dataset-REPAIR

Lipton, Z. C. (2018, June 5). From AI to ML to AI: On swirling nomenclature & slurried thought. Approximately Correct. http://approximatelycorrect.com/2018/06/05/ai-ml-ai-swirling-nomenclature-slurried-thought

Mainzer, K. (2016). Künstliche Intelligenz – Wann übernehmen die Maschinen? Springer Verlag.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1956). A proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth College. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115. https://doi.org/10.1145/3457607

Metzinger, T. (2022, March 6). Ethics washing made in Europe. Der Tagesspiegel. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Mishra, I., Kashyap, V., Yadav, N., & Pahwa, R. (2024). Harmonizing intelligence: A holistic approach to bias mitigation in artificial intelligence (AI). International Research Journal on Advanced Engineering Hub, 2(7), 1978-1985. https://doi.org/10.47392/IRJAEH.2024.0270

Mujtaba, R., & Mahapatra, S. (2024). The myth of neutrality: Biases in AI hiring systems. ArXiv. https://arxiv.org/abs/2405.19699

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

OECD. (2024). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

OECD. (2024a, July 22). OECD launches pilot to monitor application of G7 Code of Conduct on advanced AI development. https://www.oecd.org/en/about/news/press-releases/2024/07/oecd-launches-pilot-to-monitor-application-of-g7-code-of-conduct-on-advanced-ai-development.html

OECD. (2024b). Report on the implementation of the OECD Recommendation on Artificial Intelligence (C/MIN(2024)17). https://one.oecd.org/document/C/MIN(2024)17/en/pdf

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, Article 13. https://doi.org/10.3389/fdata.2019.00013

Paassen, B., Bunge, A., Hainke, C., Sindelar, L., & Vogelsang, M. (2019). Dynamic fairness: Breaking vicious cycles in automatic decision making. Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019). https://arxiv.org/abs/1902.00375

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data and Cognitive Computing, 7(1), 15. https://doi.org/10.3390/bdcc7010015

Penedo, A. C., & Kramcsák, P. T. (2023). Can the European Financial Data Space remove bias in financial AI development? Opportunities and regulatory challenges. European AI & Society Journal, 2(4), 15–30.

Pessach, D., & Shmueli, E. (2021). Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. Expert Systems with Applications, 185, 115667. https://doi.org/10.1016/j.eswa.2021.115667

Ratzan, J., & Rahman, N. (2024). Measuring responsible artificial intelligence (RAI) in banking: A valid and reliable instrument. AI Ethics, 4, 1279–1297. https://doi.org/10.1007/s43681-023-00321-5

Räz, T. (2022). COMPAS: zu einer wegweisenden Debatte über algorithmische Risikobeurteilung. Forensische Psychiatrie, Psychologie, Kriminologie, 4/2022. https://doi.org/10.1007/s11757-022-00741-9

Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. AI and Ethics, 3(4), 735–744. https://doi.org/10.1007/s43681-022-00200-5

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in AI. In Proceedings of the Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA (pp. 539–544). ACM. https://doi.org/10.1145/3308560.3317590

Rudin, C., Wang, C., & Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.6ed64b30

Russell, S., & Norvig, P. (2023). Künstliche Intelligenz: Ein moderner Ansatz (4. Aufl.). Pearson Deutschland.

Russo, L. & Oder, N. (2023, October 31). How countries are implementing the OECD Principles for Trustworthy AI. OECD.AI. https://oecd.ai/en/wonk/national-policies-2

Sadok, H., Sakka, F., & El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. Cogent Economics & Finance, 10(1). https://doi.org/10.1080/23322039.2021.2023262

Schmalzried, G. (2024, August 5). Artificial intelligence AI Act in force: What's wrong with Europe's AI regulation? Bayerischer Rundfunk. https://www.br.de/nachrichten/netzwelt/kuenstliche-intelligenz-ai-act-in-kraft-was-stimmt-nicht-mit-europas-ki-verordnung,UKsMiAJ

Shams, R. A., Zowghi, D., & Bano, M. (2023). AI and the quest for diversity and inclusion: A systematic literature review. AI and Ethics. https://doi.org/10.1007/s43681-023-00362-w

Siddique, S., Haque, M. A., George, R., Gupta, K. D., Gupta, D., & Faruk, M. J. H. (2024). Survey on machine learning biases and mitigation techniques. Digital, 4(1), 1–68. https://doi.org/10.3390/digital4010001

Simonite, T. (2018). Amazon's gender-biased hiring algorithm has been scrapped—but the problem persists. Wired. https://www.wired.com/story/amazons-gender-biased-hiring-algorithm/

Simonite, T. (2021). HireVue drops facial analysis in hiring tools. Wired. https://www.wired.com/story/hirevue-drops-facial-analysis-hiring-tools/

Smuha, N. A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence: A continuous journey towards an appropriate governance framework for AI. Computer Law Review International, 4(3), 97-108. https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/25111034/05ed277e-aa6d-4b50-b326-e8a266dcf4c4/ssrn-3443537-1.pdf

Stamboliev, E., & Christiaens, T. (2024). How empty is Trustworthy AI? A discourse analysis of the Ethics Guidelines of Trustworthy AI. Critical Policy Studies. https://doi.org/10.1080/19460171.2024.2315431

Stevens, A., Deruyck, P., Van Veldhoven, Z., & Vanthienen, J. (2020). Explainability and fairness in machine learning: Improve fair end-to-end lending for Kiva. Institute of Electrical and Electronics Engineers Inc., 1241–1248. https://doi.org/10.1109/SSCI47803.2020.9308371

Suresh, H., & Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization. https://doi.org/10.1145/3465416.3483305

Vuttipittayamongkol, P., & Elyan, E. (2020). Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. Information Sciences, 509, 47–70. https://doi.org/10.1016/j.ins.2019.08.062

Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://arxiv.org/abs/1911.11834

Wilson, A., & Caliskan, A. (2024). Biases in AI resume screening: Discrimination in names, race, and gender. ArXiv. https://arxiv.org/abs/2407.20371

Wolford, B. (n.d.). What is GDPR, the EU's new data protection law? https://gdpr.eu/what-is-gdpr/

World Bank. (2021). Data for better lives. World Development Report 2021. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1600-2

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness constraints: A flexible approach for fair classification. Journal of Machine Learning Research, 20, 1–42. https://doi.org/10.1007/978-3-642-33486-3_3

Zweig, K. A. (2018). Wo Maschinen irren können: Verantwortlichkeiten und Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung. Bertelsmann Stiftung. https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/wo-maschinen-irren-koennen/

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa