2024

Bias as Evidence

"HOW, IF AT ALL, CAN ALGORITHMIC BIAS IN RISK PREDICTION TOOLS WORK AS EVIDENCE FOR POLICY?"

MARTA ZIOSI, OXFORD INTERNET INSTITUTE, UNIVERSITY OF OXFORD



Thesis submitted in partial fulfillment of the requirements for the degree of DPhil in Information, Communication and the Social Sciences in the Oxford Internet Institute at the University of Oxford

Marta Ziosi

St Antony's College

Hilary Term 2024

Supervisors: Prof. Mariarosaria Taddeo and Prof. Luciano Floridi

Nr. Of words: 60,024

Abstract

Risk prediction tools were introduced with the promise to reduce human bias in decision-making in areas like credit, health, and criminal justice. However, they reproduce existing demographic disparities based on factors like gender, race, and class (Barocas et al., 2022), raising concerns about 'algorithmic bias' amongst scholars and society. In this thesis, I argue that what is often referred to as algorithmic bias can also be conceived of as a manifestation of existing societal disparities.

My main research question is: '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*' Given the presence of unfair differences in the way risk prediction tools perform towards different demographic groups, I investigate whether I can qualify and quantify these differences in a way that is informative for policies on bias and discrimination. This does not entail testing for bias, but rather inquiring whether I can *conceive* of, *measure,* and *use* these unfair differences in performance to *inform* policies that address existing societal disparities. Consequently, this research unfolds across four distinct levels: a *conceptual*, a *methodological*, an *applied*, and a *policy* level.

These levels will be hosted within the framework of an integrated thesis, where chapters deal with algorithmic bias as evidence (Chapter 3), a set of proposed tools to measure this bias (Chapter 4), its application in the context of predictive policing (Chapter 5), and its possible integration into specific policy frameworks to inform discrimination claims (Chapters 6 and 7). The thesis will culminate in a series of policy recommendations tailored to researchers, practitioners and policymakers working at the intersection of algorithmic fairness and policymaking. A shift from framing bias as a problem to evidence can re-orientate some efforts of these engaged communities from mitigating its effects, towards looking for the conditions that give rise to it.

Table of Contents

Abstract **2**

List of Abbreviations **8**

List of Tables **10**

List of Figures **11**

Acknowledgments **12**

Chapter 1. Introduction ***16***

1.1 Definitions & Scope **18**

1.2 Background and Rationale **22**

1.3 Research Questions and Thesis Overview **24**

1.4 Methodology **27**

1.4.1 Research Stance 27

1.4.2 Research Methods 29

1.5 Ethical Considerations & Positionality **32**

1.5.1 Positionality 33

1.6 Contributions and Target Audience **34**

1.7 Thesis Outline **37**

Chapter 2. Literature Review ***45***

2.1 Algorithmic Bias **46**

2.1.1 Bias as a Problem 46

2.1.2 Fairness Metrics as a Solution? 48

2.1.3 Bias as Evidence 51

2.1.4 Explainability Tools as a Solution? 53

2.2 Risk Prediction Tools **56**

2.2.1 The Case of Criminal Risk Assessment 57

2.3 (AI) Policy for Bias and Fairness **60**

2.3.1 The Global AI Policy Landscape 60

2.3.2 AI Policy for Bias and Fairness 62

Chapter 3. Algorithmic Bias: To preserve or to transform? *66*

3.1 Introduction **67**

3.2 The Problem with 'Fair' Predictions **70**

3.3 Algorithmic Bias: a 'Thick' Concept **73**

3.3.1 Algorithmic Bias: The Case of Credit, Health, and Crime Risk Assessment 74

3.4 Fairness Metrics: To Preserve or to Transform? **77**

3.4.1 Demographic Parity and Calibration 79

3.5 From Fair Predictions to Interventions **80**

3.5.1 Preserving Bias as a Source of Evidence 80

3.5.2 Transforming Bias as a Problem: Toward Equity-driven Fairness? 85

3.6 Policy Implications and Recommendations **87**

3.7 Conclusion **91**

Chapter 4. A Genealogical Approach to Algorithmic Bias *93*

4.1 Introduction **94**

4.2 Explainability and Fairness **96**

4.3 A Genealogical Approach to Bias **98**

4.3.1 XAI Approaches as Questions 100

4.3.2 Questioning XAI Approaches 105

4.4 Recommendations for AI policy and governance **110**

4.5 Conclusion **114**

Chapter 5. Evidence of What, and for Whom? The Socially Contested Role of Algorithmic Bias in a Predictive Policing Tool *116*

5.1 Introduction **117**

5.2 Background and Related Work **119**

5.2.1 Algorithmic bias: from an object of social critique to a device for social change 119

5.2.2 The Chicago Crime Prediction Tool 122

5.3 Methods **124**

5.4 Results **127**

5.5 Discussion **141**

5.5.1 AI Hype Rhetoric 143

5.5.2 *'Bias as defined by who, right? So ultimately, who's the judge?'* [C001] 145

5.6 Conclusion **146**

5.7 Ethical Considerations **147**

Chapter 6. Evidence of bias: Bridging Information Gaps in The EU AI Liability Directive ***151***

6.1 The EU AI Liability Directive: background, scope, and purpose **152**

6.2 Harmonisation and legal (un)certainty **156**

6.3 AI systems' complexity, autonomy, and opacity **158**

6.4 The burden of evidence **160**

6.5 Recommendations **163**

6.6 Conclusion **165**

Chapter 7. Exploring Algorithmic Bias as Evidence in the California Racial Justice Act: Opportunities, Challenges, and Policy Implications ***168***

7.1 Introduction **169**

7.2 Opportunities to use algorithmic bias as evidence **171**

7.2.1 Evidence of Systematic Disparities 171

7.2.2 Counterfactual evidence 174

7.2.3 Less 'biased' evidence 175

7.3 Challenges of using bias as statistical evidence **177**

7.3.1 Access, opacity, and complexity of AI 178

7.3.2 AI expertise, convenience & scope 179

7.3.3 The burden of responsibility 181

7.4 Implications and policy recommendations **182**

7.5 Concluding remarks **186**

8. Conclusion ***189***

8.1 Synthesis of contributions and findings **190**

8.2 Implications and policy recommendations **195**

8.2.1 Overall Implications 195

8.2.2 Synthesis of policy recommendations 202

8.3 Limitations and avenues for future research 211

8.3.1 Methodological limitations and reflections 211

8.3.2 Limitations in scope and directions for future research 213

8.4 Concluding remarks 215

Bibliography 218

APPENDIX A - DISCLAIMER FROM LUCIANO FLORIDI 256

APPENDIX B - DISCLAIMER FROM DAVID WATSON 257

APPENDIX C – DISCLAIMER FROM DASHA PRUSS 258

APPENDIX D - DISCLAIMER FROM MARIAROSARIA TADDEO 259

APPENDIX E - DISCLAIMER FROM JAKOB MOKANDER 260

APPENDIX F - DISCLAIMER FROM CLAUDIO NOVELLI 261

APPENDIX G - DISCLAIMER FROM FEDERICO CASOLARI 262

APPENDIX H – QUESTION SHEET INTERVIEWS 263

APPENDIX I – CODE TABLE 264

List of Abbreviations

AI – Artificial Intelligence

AILD – Artificial Intelligence Liability Directive

DAGs - Directed Acyclic Graphs

DFS - Dieter Schwarz Foundation

EC – European Commission

FAccT – Fairness, Accountability and Transparency

FN – False negative

FNR – False Negative Rate

FP – False Positive

FPR – False Positive Rate

FTC - Federal Trade Commission

GPAI – General Purpose AI

LLM – Large Language Model

ML – Machine Learning

NeurIPS - Conference and Workshop on Neural Information Processing Systems

NIST - National Institute of Standards and Technology

OECD – The Organisation for Economic Co-operation and Development

OII – Oxford Internet Institute

PLD – Product Liability Directive

RQ – Research Question

TN – True Negative

TNR – True Negative Rate

TP – True Positive

TPR – True Positive rate

XAI – Explainability

List of Tables

TABLE 1. ROLES OF PARTICIPANT POPULATION 125

TABLE 2. PARTICIPANT EXPOSURE TO THE CRIMINAL LEGAL SYSTEM (CLS) 125

TABLE 3. SUMMARY OF POLICY RECOMMENDATIONS FOR TARGETED AUDIENCES 211

List of Figures

FIGURE 1. OVERVIEW OF THESIS CHAPTERS AND METHODS 26

Acknowledgments

This thesis would have not been possible without the support, both mental and material, of many people.

First, I want to thank Prof. Beatrice Borghi. This thesis started as an abstract idea, driven by the will to make it concrete and impactful. However, due to financial constraints, the acceptance email from the University of Oxford was received with a mix of excitement and panic: how am I going to afford it? This was followed by months of emails, calls, online searches, and advice from close friends (Thank you, Ceci!). I was determined to pursue this, but I needed someone to believe in the project with me. That someone replied to me on the 24/04/2020, and it was Beatrice. Despite never having met me (up until today!), Beatrice believed in my project and she 'moved mountains' (pardon me the Italian expression) to make this happen. Until she did. On the 15/10/2020 Beatrice informed me that the Rotary Club Bologna Est approved the funding for my project, and everything changed. Since that moment, I knew Beatrice's name would be on top of this list, as it will be in the future. I cannot thank Beatrice enough for teaching me to believe in others, with an open mind and an open heart. It is thanks to her that I am determined to do the same to that someone in the future who will need it.

Second, I want to thank my supervisor Prof. Luciano Floridi. Luciano was the other person who, just from an email, believed in this project. Through his words of encouragement and advice he helped me turn a rough idea into a thesis proposal. Among the valuable pieces of advice he gave me, a sentence stuck with me. He said, *'Think about this proposal like a keyhole. It has to be narrow enough to be feasible and yet, broad enough to allow one to see something bigger through it.'* This sentence was crucial to help me give importance to the small details in the proposal and, at the same time, put into perspective the general idea that drove it. This was just some of the valuable advice that I

will treasure with me now and always in my career (and life). Thank you for being the supportive and inspiring supervisor that you have been. I also want to thank Prof. Mariarosaria Taddeo. Even though she became my supervisor late along the way, she has always provided me with support and opportunities that fueled my PhD experience. I truly hope (and want) that this Ph.D. is just the first step towards a long way together.

Third, I want to thank the funding institutions without which this thesis would have not been possible. I want to thank the Rotary Foundation, especially the Rotary Club Bologna Est, the Maidenhead Club (especially Judith Diament for supporting me through Covid!), and the Headington Club (Richard, thank you for always bringing a smile to my face, and putting all this passion into what you do – including your amazing cooking!) around Oxford. I also want to thank my department, the Oxford Internet Institute (OII), and the Dieter Schwarz Foundation (DSF) for supporting me through two years of living expenses and crucial research budget during my PhD. The OII-DSF award arrived at the right time, and it turned a struggle into a possibility. Without its generous research funding, I would have never been able to attend the conferences that most shaped my thought and the chapters that the reader will find in this PhD.

Fourth, I want to thank the people who were crucial in the making of this PhD. I want to thank Cecilia Corsini for the precious support and advice. Ceci, thank you for always saying the right thing at the right time. Please, never hold back! I also thank you for greatly improving my taste in wine. I want to thank the people from the Digital Ethics Lab, which is today hosted in Yale. Especially, I want to thank Jess for her invaluable moral and academic support. You are a wonderful genius. I also want to thank Jakob for his precious advice, his constant support and the insightful walks we had together. I hope there will be many more in the future! I also want to thank David, Josh, Prathm, Andreas and all the others from the Lab. I am grateful to all my friends from the Oxford Internet Institute,

especially those that I have started, and I am still with: Jo (and Paul!), Liam, Clemi and many others. I also want to thank Ana Valdivia for showing me what a 'true' passion for academia (and people) looks like. I will bring it with me always.

Finally, I want to thank the people who are close to my heart. I want to thank my Indian family: Smita, Priaanshi, Praveen and Pranav. Mosi, thank you for showing me what a PhD and determination look like. I want to thank my favourite football team: the Foxes. You have kept me sane and insanely happy for all these years. You have also given me friends (and love) that will be with me forever. Lena, I am so grateful we met on that train and for all the support you have given me. I want to thank Cailean. You are my best friend, and I am so happy to have 'grown up' with you through all these adventures. I know you are physically far (Colombia), but we never really mentally are. I also want to thank Pauline. You have truly witnessed the ups and downs of this PhD. Thank you for living live to the fullest with me. I will always treasure it. Also, thank you to Elizabeth Hesse for allowing me to believe that this project was possible, and to Ulrich for his unmistakable humour. I want to thank Chiara. You have supported me through the last push with sports, presence and love but really, you have always been there. Finally, I want to thank my family: Chiara, Michele and Emilia. This thesis is really for you. I would not be here without you but most importantly, I would not be myself.

Chapter 1:

Introduction

Chapter 1. Introduction

Risk prediction algorithms are data-driven tools used to measure the risk of a future outcome in areas including but not limited to credit, e.g., the risk of default; criminal justice, e.g., the risk of someone re-committing a crime; and healthcare, e.g., the risk of heart failure. While these tools were introduced with the promise to reduce human bias in decision-making, they have been found to reproduce or exacerbate pre-existing societal biases: demographic disparities along the lines of gender, race, and class amongst others (Barocas et al., 2022). This is mirrored in the presence of unfair differences in the way risk prediction tools perform towards these demographic groups (Mitchell et al., 2021). This phenomenon has given rise to concerns about 'algorithmic bias' amongst scholars and society at large. In this thesis, I argue that what is often referred to as algorithmic bias can also be conceived of as a manifestation of existing societal disparities. Starting from this intuition, the main aim of my research is to understand, as an academic researcher, in which ways algorithmic bias can carry evidence of these disparities to inform policies that tackle the societal conditions that give rise to them.

My main research question is: '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*' Given the presence of unfair differences in the way risk prediction tools perform towards different demographic groups, I enquire whether I can qualify and quantify these differences in performance in a way that is informative for policies around bias and discrimination. This does not entail testing for bias, but rather inquiring whether I can *conceive*, *measure,* and *use* these unfair differences in performance to *inform* policies that address existing societal disparities. Consequently, this research unfolds across four distinct levels of inquiry: a *conceptual*, a *methodological*, an *applied*, and a *policy* level.

On the conceptual level, the research is guided by sub-question 1, *'How can bias as evidence be conceptualised?'* On the methodological level, sub-question 2 asks, *'How can bias as evidence be measured?'* On the applied level, sub-question 3 is, *'How can bias as evidence be applied in practice?'* Finally, on the policy level, sub-question 4 asks, *'How can bias as evidence be used to inform policy frameworks to tackle discrimination?'* These sub-research questions remain overly broad. For this reason, they will be asked within the framework of an integrated thesis, where each chapter is an inquiry into a specific, yet relevant topic for the research question. In order, the chapters deal with algorithmic bias as evidence (Chapter 3), a set of proposed tools to measure it (Chapter 4), its application in the context of predictive policing (Chapter 5), and its possible integration into specific policy frameworks to inform discrimination claims (Chapter 6 and 7), respectively. Going back to the main research question, the thesis will culminate in a *policy* output: a series of policy recommendations tailored to the audiences of the specific chapters, accompanied by some overarching considerations on how algorithmic bias can inform policies around bias and discrimination.

Conceptually, a shift from framing bias as a problem to framing it as evidence of problems (e.g., societal disparities) can re-orientate some efforts of the research community from mitigating its effects, towards looking for the conditions that give rise to it (e.g., historical practices of oppression, such as redlining). *Methodologically*, this shift can prompt a redirection of bias detection tools toward uncovering its root causes and thus, provide insights into its origins and mechanisms. In *practice*, trying to capture what bias carries evidence of could help rethink the predictive performance of risk prediction tools as a mirror of past conditions (Kleinberg et al., 2017; Mayson, 2018), rather than a future predicament. Finally, if what bias carries evidence of were to inform *policy*, this could

broaden the scope of policies around AI from how to best automate solutions toward how to tackle existing societal problems (Abebe et al., 2020).

1.1 Definitions & Scope

This thesis presents the idea that algorithmic bias, mirrored in inter-group disparities in predictive performance, provides ways to gather evidence of the societal conditions that influence it, which could be valuable for policy. Breaking away from the sole connotation of bias as a 'problem', this thesis attempts to approach it also from the perspective of *evidence*, defining algorithmic bias as when

> '*predictive performance - however specified - unjustifiably differs across groups along social axes such as race, gender, and class'* (Mitchell et al., 2021, p.1).

This bias may also be referred to as a 'skewed performance' along one of these social or demographic axes. This thesis does not conceive of algorithmic bias as inherently problematic. As the definition suggests, however, it can become so when it aligns with and perpetuates underlying societal inequalities, such as those related to gender, race and class. This not only results in discriminatory outcomes but also significantly exacerbates these societal disparities in a systematic and large-scale manner. This is when algorithmic bias is referred to as a 'problem'.

This framing of bias as a 'problem', as the literature review will argue, is largely adopted in the literature. This thesis aims to explore algorithmic bias from the perspective of 'evidence.' It examines the various ways in which algorithmic bias, precisely when reproducing existing disparities, can also be used to gather evidence of them. As the thesis chapters will present, algorithmic bias can be conceived of as a 'signal' to build hypotheses about the presence of societal disparities (Chapter 3 and 4), it can serve as a tool for diagnosing societal issues and justifying the need for interventions (Chapter 5),

and it can be understood in a legal sense as providing statistical data to establish facts or support arguments in legal proceedings (Chapter 6 and 7).

Additionally, it mainly focuses on bias in risk prediction tools. A risk prediction algorithm refers to a set of mathematical rules or processes designed to analyse input data and to predict the likelihood of a particular outcome (Express Analytics, 2022), such as the probability of reoffending, defaulting on a loan or developing a disease. In this thesis, 'risk prediction tools' refer to the application of such algorithms in specific contexts, and as such, the two terms are used interchangeably.[1] More broadly, risk prediction tools here refer to what the academic literature may alternatively present as data-driven, actuarial, predictive, risk assessment or algorithmic tools to predict risk.[2] Thus, risk prediction tools may also be referred to through these terms in the thesis. Importantly, my focus is on tools to measure individual or event-level risk, rather than on methods to assess the risk of algorithms, e.g., tools for auditing[3]. I do not refer to these tools strictly in terms of the algorithms they use, but also in terms of the training data, the data on which an algorithm is trained, and the selection of input and output variables (König & Krafft, 2021). These all play a crucial role in bias (Mayson, 2018) and thus, they ought to be considered when analysing bias as 'evidence'.

'Risk' refers to the likelihood that harm occurs in a system (Derczynski et al., 2023). The risk prediction tools in the areas cited above tend to individualise systemic risk. For example, even though the risk of crime derives from a series of socio-economic factors

[1] As the sentence above explains, risk prediction 'algorithms' and 'tools' are not precisely the same thing. They should be understood as 'interchangeable' in a colloquial rather than in a strictly logical sense.

[2] While the debate on whether these algorithms are 'AI' is beyond the scope of this thesis, the thesis will refer to these as 'AI systems' only insofar as they fall under this category in specific regulatory policies considered.

[3] For more on this, see (Mokander & Floridi, 2021).

(e.g., poverty and lack of job opportunities, among others) (Wikström & Treiber, 2017), crime prediction tools ascribe this risk to individuals or groups belonging to lower socioeconomic classes, defining them as 'high-risk'. Risk, thus, is a suitable lens through which to explore how bias affecting certain groups or individuals can mirror broader disadvantageous societal conditions.

Another important element is the context in which this bias arises. This thesis limits itself to analysing and presenting examples taken from three specific areas: that of credit, criminal justice and health risk prediction in the US. Strategically, considering multiple cases allows researchers to systematically compare variations between them (Campbell, 1975). However, by representing single cases, the resulting considerations should not be used as grounds for generalisation (Tellis, 1997). These areas are, indeed, not the only three where algorithmic bias occurs[4]. However, given a documented history of institutional inequalities in these sectors in the US (Solly, 2020), they are fitting as illustrative examples where bias is specifically problematic in its consequences, e.g., being denied a loan or being arrested, and tends to carry historical evidence of societal disparities, e.g., past redlining divisions or disparate arrest practices by the police.

Given this very history, these cases also help highlight the conditions by which this bias is perpetuated. Past redlining divisions, differential access to healthcare, and disparate arrest practices toward people of colour represent some examples of what I mean by these 'conditions'. Historical segregation in US neighbourhoods, for instance, profoundly affected the residents' access to credit, health insurance, and education (Agyeman, 2021; Perrino, 2020). In turn, this created the conditions of poverty, unemployment, and past

[4] See Barocas et al. (2022) for a more comprehensive view of cases on bias and discrimination in machine learning.

default history by which residents in these communities are considered 'not worthy' of credit when zip codes are used to calculate the risk of default. Given the especially damaging and widespread character of bias in crime risk prediction (Angwin et al., 2016; König & Krafft, 2021; Mayson, 2018), this thesis puts added emphasis on it throughout several chapters.[5]

At the level of policy, this thesis refers to it as in line with OECD's definition of 'regulatory policy' as

> *'the use of regulations, laws, and other instruments to deliver better economic and social outcomes'* (OECD, 2021).

Regulatory policy, as understood here, includes both hard and soft law initiatives that aim to create rules or guidance for designing, developing, and/or deploying AI. Specifically, I will focus on AI policies concerning bias, fairness and discrimination in the EU and in the US, two countries among the 'early movers' in AI regulatory policy (Roberts et al., 2023). In the EU, I will focus on the case of the EU AI Liability directive (AILD), a proposed legislation to determine who is liable for harms caused by AI systems, including user discrimination (AI Liability Directive, 2022). Given its overarching approach to AI harms and its complementarity with the EU AI Act (AI Act, 2021)[6], the proposal is representative of the EU's approach to addressing AI regulation horizontally (Roberts et al., 2023). The EU AI Act, with its focus on risk management, data bias, transparency and human oversight, is designed to create a regulatory environment to prevent and mitigate algorithmic bias. Differently, the focus of the AILD is on redress and thus, on establishing

[5] Especially, Chapter 5 and Chapter 7.

[6] Here and in the rest of this thesis I shall consider the first Draft (COM/2021/206 final) and the final draft of compromise amendments to the same, (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)). While I cite the first draft, a link to the latest one can be found here: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

mechanisms to provide evidence of harm, including algorithmic discrimination. This thesis focuses on the latter because it aligns more closely with its aims and scope. In the US, I will focus on the case of the California Racial Justice Act (CRJA), a landmark legislation that prohibits the state from seeking or obtaining a criminal conviction or imposing a sentence based on race, ethnicity, or national origin (California Racial Justice Act, 2020). In this case, I will focus on state-specific legislation because it reflects the US' approach to addressing AI regulation vertically through sector-specific laws (Roberts et al., 2023).

1.2 Background and Rationale

In the realm of risk prediction, where algorithms are used to calculate future risk based on historical data, algorithmic bias is simultaneously a problem and evidence by reproducing existing societal disparities. By using zip codes as an input feature, for example, credit risk assessment can result in unfair credit allocations while mirroring previous redlining divisions across protected groups (Blattner & Nelson, 2021; Citron & Pasquale, 2014). A health risk assessment tool can underestimate health risks for people of color while mirroring differential access to health insurance across the population (Obermeyer et al., 2019). In the context of recidivism, bias can lead to overestimating the likelihood of recommitting a crime while mirroring disparate arrest practices toward people of color (Angwin & Larson, 2016). The research community has responded to disparities in predictive performance by remaining agnostic toward the specific reasons for these disparities (Barabas et al., 2018). They have focused on developing metrics that can adjust predictive performance based on what is considered 'fair': fairness metrics. Some common fairness metrics are calibration, demographic parity and equalised odds[7]. Though different, these metrics generally reply to the question, 'How can we predict

[7] An exhaustive list can be found in 'Fairness and Machine Learning' (Barocas et al., 2022).

more fairly?' These are designed to detect and mitigate biases that may exist in the skewed performance of risk prediction tools, particularly those that affect historically marginalised groups along the lines of gender and race, among others (Barocas et al., 2022). These approaches are useful to address bias in its consequences, e.g., in its disparate outcomes toward certain marginalised demographics (Barabas et al., 2018). However, to truly address bias, it is also necessary to look at the conditions that give rise to it. As previously mentioned, past redlining divisions, differential access to healthcare, and disparate arrest practices toward people of colour represent some of these conditions.

One proposed solution for assessing these conditions is the use of explainability (XAI) approaches. Explainability approaches are techniques used to help understand how machine learning models arrive at their decisions or predictions (Packt, 2018). Some common explainability techniques include feature importance, partial dependence plots, and SHAP (SHapley Additive exPlanations). However, algorithmic bias is a complex phenomenon that cannot be fully captured by XAI tools alone. Specifically, XAI approaches tend to represent the role of protected characteristics in discriminatory outcomes unrealistically, e.g., as independent, intrinsic, and causal attributes[8]. In this regard, policy interventions[9] may serve as a broader tool for addressing the real-world problems of bias and fairness.

On the one hand, most policy documents concerning AI outline ethical guidelines to promote algorithmic fairness, which is often understood as the prevention, mitigation, and monitoring of unwanted bias and more rarely as the 'absence' of discrimination (Jobin et

[8] Chapter 4 elaborates further on this.

[9] Purposive attempts by governments or organizations to achieve specific goals through regulations, incentives, public programs or market mechanisms (Wiseman, 1979)

al., 2019). On the other hand, some of this policy work advocates using algorithms to ensure fairness and minimise human bias in decisions (Future of Life Institute, 2017; Kleinberg et al., 2017; Silberg & Manyika, 2019). Notwithstanding the importance of creating hard and soft procedures to avoid the risks and maximise the benefits of AI in society (Floridi et al., 2018), most of this work assumes the goal or aim of these systems to be the ethical or fair automation of human tasks (Greene et al., 2019). Much less attention, for example, has been focused on the potential of algorithms to be used for 'diagnostic' purposes – that is to say, to detect existing societal disparities rather than to automate decisions (Xiang, 2020) - and on how public policy can take advantage of that.

1.3 Research Questions and Thesis Overview

The above paragraphs suggest several important points. Firstly, the literature on algorithmic fairness correctly acknowledges algorithmic bias as a problem. However, bias can also indicate deeper issues that negatively affect predictive performance, such as societal disparities. Secondly, explainable AI (XAI) tools can be useful in identifying and addressing algorithmic bias in terms of the conditions that bring it about. However, these tools typically focus on a narrow aspect of bias, such as feature contributions or cause-and-effect relationships and may not address the full range of bias issues. Finally, bias poses a significant obstacle to the fair use of AI in society. However, once bias is identified, the knowledge gained can be used to inform policies that address the underlying societal conditions that give rise to it.

Far from providing solutions, the above considerations ground the questions that this thesis concerns itself with. The guiding research question is, '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*' It is important to qualify the assumptions of the research question and their implications. In that respect, the presence of bias is not something that this research tries to demonstrate. By asking 'how' it

can ‘work as evidence’, I am not questioning the presence of bias in a model, but rather its usability towards a certain goal. That is to say, its presence is the starting point. Given the presence of inter-group disparities in predictive performance, I enquire whether I can qualify and quantify them in a way that is informative for policy. This does not entail testing for bias, but rather inquiring whether I can *conceive of*, *measure*, and *use* these inter-group disparities to *inform* policymaking. Respectively, the research question hints towards a *conceptual*, a *methodological*, a*n applied,* and a *policy* component. However general, each of these components will be actualised in a series of specific studies that represent the chapters of this integrated thesis.

On a conceptual level, the research is guided by sub-question 1, *‘How can bias as evidence be conceptualised?’* As an answer, Chapter 3 undertakes a critical reconceptualisation of algorithmic bias not only as a problem but also as evidence of existing societal disparities, and its implications for the capacity of fairness metrics to preserve as well as transform the societal disparities that algorithmic bias makes manifest.

On the methodological level, sub-question 2 asks, *‘How can bias as evidence be measured?’* Chapter 4 offers an analytical framework to approach algorithmic bias ‘genealogically’: by measuring and addressing bias in the conditions that enable it, rather than in its consequences. It analyses explainable artificial intelligence (XAI) approaches with respect to their ability to gather evidence – note, not proof – of societal disparities.

On the applied level, the sub-question 3 is, *‘How can bias as evidence be applied in practice?’* Chapter 5 is concerned with the applied use of algorithmic bias as evidence of existing disparities in the context of predictive policing, with a special eye toward making evidence of police enforcement bias work in service of marginalised, affected communities.

Finally, on the policy level, sub-question 4 asks, *'How can bias as evidence be used to inform policy frameworks to tackle discrimination?'* This sub-question, circling back to the main research question, should be understood as an outreach of the thesis: an exploratory attempt to sketch the direction and shape of its implications for different policy frameworks in context. As a consequence, it is answered through two policy 'deep dives'[10]: Chapter 6 explores the possibility of leveraging evidence on AI systems to inform assertions of algorithmic discrimination in the context of the EU Liability Directive, and conversely, Chapter 7 explores the possibility of employing algorithmic bias as evidence to substantiate claims of discrimination under the California Racial Justice Act.

Going back to the main research question, the thesis will culminate in a *policy* output: a series of policy recommendations tailored to the audiences of the specific chapters, accompanied by some overarching considerations on how algorithmic bias can inform policies around bias, fairness and discrimination.

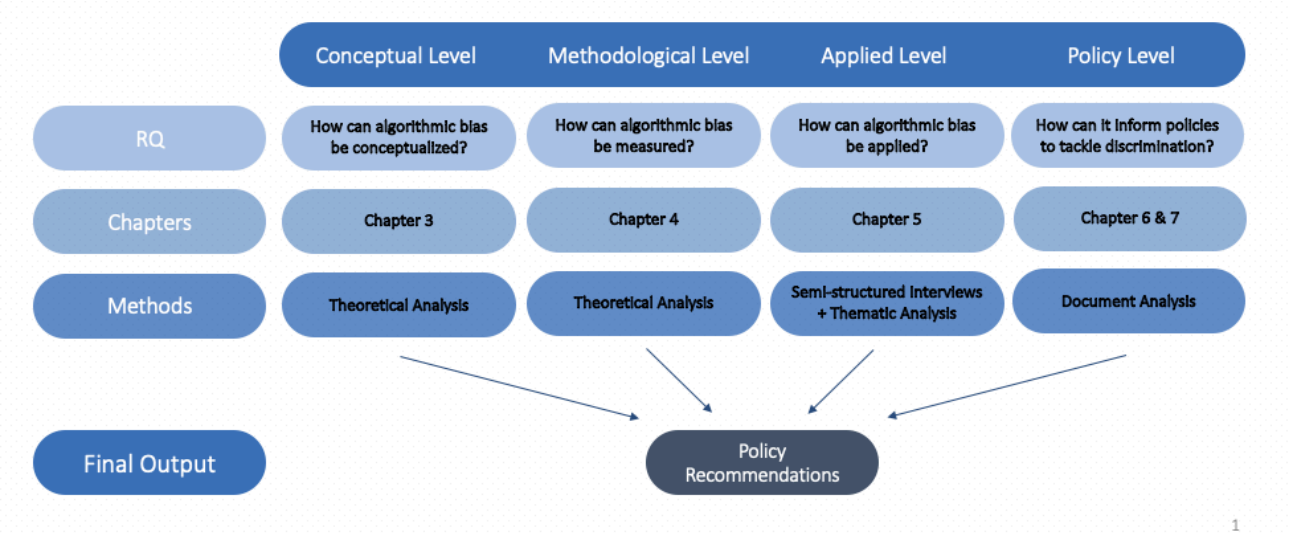


Figure 1. Overview of Thesis Chapters and Methods

[10] 'Deep dive' here refers to an in-depth or vertical review or exploration of a specific topic.

Figure 1 presents an overview of the four main levels of the thesis. It outlines the sub-research questions that guide them, the related chapters that aim to answer them, and their respective methods of analysis. It also presents the policy recommendations as the final thesis output. The paragraphs below will delve deeper into the methodological approach taken by each of these studies. Additionally, the Thesis Outline section at the end of this chapter will also provide a more in-depth overview of their respective content.

1.4 Methodology

This thesis approaches the research questions on four levels: conceptual, methodological, applied, and policy level. The five substantive chapters (Ch3-7), grouped under these levels, employ specific methods. Below, I elaborate on this thesis' broader research stance and the methodological approach taken by each chapter. For each method, I also outline some limitations.

1.4.1 Research Stance

In this thesis, I take a critical pragmatic stance. Like traditional pragmatism (Dewey, 1920; James, 1907; Peirce, 1903), this stance starts from real-world problems rather than a specific worldview or assumed 'truth' about the world (Salkind, 2010). Additionally, it evaluates solutions (and methodologies) depending on 'what works' rather than on what a worldview (e.g., positivism, constructivism) would prescribe (Bryman, 2013). It does not a priori privilege some types of methods over others, and it allows for the use of mixed methods. At the same time, when formulating pragmatic solutions, it takes a critical stance by interrogating how each problem is constructed (Kaushik & Walsh, 2019). The goal of critical pragmatism is to enhance critical reflection on the systematic power imbalances driving the status quo and use this understanding to ground and develop research projects (Kaushik & Walsh, 2019).

On the one hand, this thesis takes a critical stance toward the dominating conception of bias primarily understood as a 'problem'. Specifically, it claims that its predominant solutions are not informed by the systematic societal disparities that this bias partly carries. At the same time, it aims to use this understanding of bias as 'evidence' to inform practical policy solutions that address the conditions that enable these disparities, thus blending a critical and a pragmatic component. In line with this research approach, the methods employed in each chapter were chosen according to 'what worked' and thus, what fit a conceptual, methodological, applied, and policy level of inquiry.

As mentioned above, a pragmatic research stance does not prescribe nor privilege some types of methods over others a priori (Kaushik & Walsh, 2019; Tashakkori & Teddlie, 2003). Some scholars, such as Hampson & McKinley (2023), view this as a limitation of pragmatism. They argue that it can lead one to choose methods by convenience, rather than through a deliberate and structured selection. This critique highlights a potential risk of methodological inconsistency and lack of methodological rigor in pragmatic research. To address this limitation, the thematic division of chapters within this thesis, delineated into distinct 'levels', aims to provide a structured framework for its research. Thus, it hopes to facilitate a diversified, yet coherent selection of research methods tailored to each respective thesis 'level'. The added critical component to the research stance, additionally, fosters a selection of research methods according to the socially-situated nature of problems and thus, justifies a choice of methods that are most appropriate to answer a specific research question (Kaushik & Walsh, 2019). As we will see in the next paragraphs, this entails the application of theoretical frameworks for the conceptual chapter and the methodological chapter, qualitative interviewing for the applied chapter (Chapter 5) and document analysis for the policy chapters (Chapters 6 and 7). I will elaborate on these below.

1.4.2 Research Methods

First, each research question, and thus, each chapter, is motivated by the problems identified in the literature review (Chapter 2). Rather than engaging in a gap-spotting exercise, the literature review engages in an exercise of problematisation: it challenges how algorithmic bias is already discussed in the literature at the crossing of Computer Science, Science and Technology Studies (STS), Law and Policy. Gap-spotting and problematisation are two ways in which researchers develop research questions from existing literature (Sandberg & Alvesson, 2011). Gap-spotting, which can refer to neglect, confusion, or lack of application of a research topic, tends to reinforce already influential theories (Alvesson & Sandberg, 2011). Differently, problematisation challenges the assumptions that underlie these theories, making it an appropriate method for scrutinising the prevailing assumptions regarding bias as a 'problem' within the literature, particularly within a critical pragmatic framework. This exercise of problematisation is carried out through a narrative critical review (Stratton, 2019). This type of review critically assesses assumptions in the literature (Stratton, 2019) and thus, it helps motivate the research conducted in the specific chapters. While this method suffers from the danger that the author's opinions bias the review of the materials (Green et al., 2006), the considerations that I draw should not be understood as conclusions. Instead, they should be grasped as conjectures to further investigate or 'test' in the chapters that they motivate.

The conceptual chapter and the methodological chapter are both of a theoretical character. A theoretical study entails using analytical tools – 'conceptual frameworks' – to reply to a research question by structuring, guiding or substantiating an argument (Bhat, 2018). In Chapter 3, I use the theory of thick concepts by Williams (1985) to substantiate a simultaneously prescriptive and descriptive interpretation of algorithmic bias. Specifically, I use this theory to argue that algorithmic bias, as a thick concept, can be both a problem in

its consequences and evidence of existing societal disparities. In Chapter 4, I draw on Pearl's ladder of causation (2000, 2009) and on Kohler-Hausmann's (2019) constructivist theory of discrimination to adopt a constructive and critical approach to capturing and addressing algorithmic bias; namely, one that explains algorithmic bias in terms of the social and technical conditions that enable it and yet, that remains critical about its ability to represent salient aspects of discrimination through technical tools. An overreliance on theory can run the risk of preventing the importance of empirical data from shaping the research (Collins & Stockton, 2018). Rather than being an argument against the use of theoretical analysis itself, this is a warning on the importance of ensuring that a thesis welcomes both components. The following chapter, in fact, is of an empirical character.

Chapter 5 is an applied study. An applied study aims to find solutions to immediate problems that face individuals but also organisations and society by looking at how these problems are instantiated in context (Bickman & Rog, 1998). In this study, I explore how evidence of police enforcement bias captured by a predictive policing tool in Chicago is used by different actors to advance criminal justice interventions. I conducted semi-structured interviews (Edwards & Holland, 2013) with Chicago area community organisations, academic researchers, and public sector actors identified through snowball sampling (Parker et al., 2020). Semi-structured interviews allowed me to focus and dig deeply into an initial set of topics and assumptions about potential ways in which evidence of algorithmic bias can inform policy, while also letting participants' expertise guide the interview and allowing new themes and potential solutions to emerge (Roulston & Choi, 2018). Interview transcripts, however, often lack a consistent structure (Nowell et al., 2017). This runs the risk that a researcher's interpretations and findings are not properly derived from the data (Nowell et al., 2017). Thus, I analysed the interviews through a thematic analysis, identifying recurring themes, and switching between inductive coding

and data collection to clarify and strengthen themes and uncover opposing evidence (Miles et al., 2014). Given its empirical character and extensive methodology, Chapter 5 will further elaborate on the specifics of the methods employed.

The policy level branches out into Chapters 6 and 7. A policy study entails the evaluation or analysis of policies in their content, their process of development and/or their impact on society (Ascher, 1986). In this policy study, I conduct a document analysis of the EU Liability Directive in Chapter 6 and of the California Racial Justice Act in Chapter 7. Document analysis analyses the content of policies and technical documents written by people who are active participants in the process they describe (Karppinen & Moe, 2011). In Chapter 6, I analyse the EU AI Liability Directive's capacity to address the 'information gap'; namely, the ability to grant claimants the right to access information about high-risk AI systems to prove algorithmic discrimination. In Chapter 7, I analyse the opportunities and challenges of using algorithmic bias as evidence of systematic racial disparities in the California Racial Justice Act. In both cases, the document analysis mainly relies on primary sources (Golub, 2022); the legal texts involved. An analysis of primary sources offers first-hand evidence to support one's arguments and thus, provides stronger support for my claims. However, in order to minimise the risk of overreliance on and misinterpretation of one single text, I have triangulated the results with an analysis of adjacent pieces of legislation, policy guidance and commentaries on the respective legal texts (Frey, 2018). In line with the aim of policy studies (Kilonzo & Ojebode, 2023) and a pragmatic stance to research (Salkind, 2010), the output of both chapters is a series of policy recommendations for practitioners on how to use bias as evidence to inform policymaking.

1.5 Ethical Considerations & Positionality

It is important to consider both the ethical assumptions as well as the societal implications of my research. Regarding the former, this research assumes that bias carries evidence of societal disparities. It is important to warn against the risk of interpreting these disparities as 'intrinsic differences' between societal groups. For example, in the case of crime risk prediction, interpreting 'societal disparities' as intrinsic differences among groups might suggest that some people are more prone to committing crimes and thus, justify policy interventions that reinforce rather than address them. This assumption would be incorrect and unwarranted. It has been proven how, for example, crime risk prediction algorithms use past arrests as a proxy for crime (Fogliato et al., 2021; B. Harcourt, 2015a). As a consequence, bias in these algorithms is often a mirror of disparate arrest practices by the police.

However, stressing the theoretical incorrectness of an assumption does not alone prevent stakeholders from achieving their goals in practice. As Benjamin (2019) states, data does not ultimately speak for itself, nor does it alone change one's stance or policies. Importantly, as mentioned above, every research project comes with its own transformative effects on society (Denzin & Lincoln, 2024) which ought to be mitigated. Accordingly, Chapter 5 directly addresses the socially contested role of algorithmic bias; namely, the possibility for different stakeholders to employ it to justify policy interventions that align with their own goals and assumptions about the world. It provides a way to assess policy interventions depending on whether they reinforce or challenge existing power structures. While this contribution does not fully solve the concerns raised above, it provides a tool to call out and address the potential misuse of evidence of bias in society when it happens. More information about this can be found in Chapter 5.

Another important aspect of a research project is its compliance with established ethical rules and guidelines. My research project has been approved by the OII's Departmental Research Ethics Committee (DREC). It was given the research ethical approval reference number SSH_OII_CIA_21_104. Finally, as the chapters of the thesis have been submitted for publication in academic journals and conferences outside of the University, this research is also in line with the widely recognised ACM Code of Ethics and Professional Conduct[11].

1.5.1 Positionality

As a white woman and an academic researcher from Europe, I acknowledge that my personal and professional background shaped how I selected research questions and analysed data, as well as the manner in which I interacted with study participants from backgrounds similar and dissimilar to mine. I have a privileged academic affiliation, race, and lack of personal contact with the criminal legal system, which undoubtedly shaped the information that grassroots community organisations and other interviewees were comfortable sharing with me. Chapter 5 further elaborates on this and on the theme of exploitation of marginalised communities by academic researchers.

Especially in conducting interviews, my positionality as a woman played a double role. On the one hand, research shows that female researchers may witness inverted situations of power imbalances in interviews (Abels & Behrens, 2009; Puwar, 1997). In my research, this occasionally took the form of repeated interruptions, lengthy monologues and not being taken seriously (Boucher, 2017). At the same time, research recognises that this can occasionally result in heightened trustworthiness being placed on the female researcher (Deem, 1994; Ortbals & Rincker, 2009). This facilitated building trust with research

[11] https://www.acm.org/code-of-ethics

participants. Additionally, my privileged position as a white academic researcher from a well-recognised institution helped build rapport with other academic researchers in my research, who had similarly received training at elite academic institutions.

1.6 Contributions and Target Audience

The overarching research question is, '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*' Each level of the thesis brings its own contribution to the reply. Below, I elaborate on these contributions, together with introducing the audiences toward which these contributions are targeted.

Conceptually, a shift from framing bias as a problem to framing it as evidence of problems (e.g., societal disparities) can re-orientate some efforts of the research community from mitigating its effects, towards addressing the conditions that give rise to it (e.g., historical practices of oppression, such as redlining). Toward that end, Chapter 3 brings a 'thick' conception (Williams, 1985) of bias: one that ascribes both a prescriptive and a descriptive aspect to algorithmic bias. Specifically, I argue that algorithmic bias can be both a problem and an indicator of social inequalities. Through this conceptualisation, I claim that one should preserve the *evidence* of societal disparities carried in algorithmic bias, and, through this evidence, inform interventions to address the problems that these disparities represent. This chapter is aimed at a primarily academic audience, specifically at researchers working at the intersection of algorithmic bias and fairness metrics. Accordingly, part of this research has been presented at the Doctoral symposium of the Fairness, Accountability and Transparency (FAccT)[12] in ML 2023 conference. At the moment of writing, Chapter 3 is under review in Digital Society[13] (Indexed Springer

[12] https://facctconference.org/2023/

[13] https://link.springer.com/journal/44206

journal). This peer-reviewed journal publishes articles on topics related to governance, transformations, and developments of digital technology and their implications for society.

Methodologically, this shift from problem to evidence can prompt a redirection of bias detection tools toward uncovering its root causes and thus, provide insights into its origins and mechanisms. Accordingly, Chapter 4 provides a 'genealogical' approach to algorithmic bias; a constructive, epistemic critique to explain algorithmic bias in terms of the conditions that enable it. I focus on XAI feature attributions (Shapley values) and counterfactual approaches as potential tools to gauge these conditions and offer two main contributions. One is constructive: I develop a theoretical framework to classify these approaches according to their relevance for bias as evidence of societal disparities. The other contribution is critical: I evaluate these approaches in terms of their assumptions about the role of protected characteristics in discriminatory outcomes, especially in their ability to represent some salient aspects of discrimination. This chapter is also primarily aimed at an academic audience, specifically at researchers working at the intersection of algorithmic bias and explainability. It has been presented as an extended abstract at the NeurIPS[14] 2022 workshop 'Algorithmic Fairness Through the Lens of Causality and Privacy'.[15] It has now been accepted for publication in Minds & Machines[16] (indexed Springer journal). This peer-reviewed journal fosters a tradition of criticism within the AI and philosophical communities on problems and issues of common concern at the crossing of philosophy and computer science.

[14] https://nips.cc/

[15] https://nips.cc/virtual/2022/workshop/49967

[16] https://link.springer.com/journal/11023

At the *applied* level, trying to capture what bias stands as evidence of could help rethink the predictive performance of risk prediction tools as a mirror of past conditions (Kleinberg et al., 2017; Mayson, 2019), rather than a future predicament. Accordingly, Chapter 5 offers an analysis that can help recognise whether criminal justice interventions motivated by evidence of algorithmic bias in a predictive policing tool reinforce or challenge existing power structures. This also contributes to seeing data-driven policies as mirroring long-standing tensions and debates in contexts such as policing, such as debates around reforms and abolition. On the one hand, by focusing on applied aspects of algorithmic bias mitigation, this chapter also targets an academic audience such as the FAccT community. In fact, it has currently been accepted to FAccT 2024. On the other hand, given its engagement with the local communities and the intention to minimise the extractive character of academic research toward these, this chapter has been selected to be developed into a short documentary by Oxco[17], a video production company based in Oxford. The target audience for this specific output is policy practitioners and community members working around community safety in Chicago and in the US more broadly.

Finally, if what bias stands as evidence of was to *inform policy*, this could broaden the scope of policies around AI from how to best automate solutions, toward how to tackle existing societal disparities (Abebe et al., 2020). To that end, Chapter 6 offers ways to reduce the burden of evidence to prove algorithmic discrimination for end-users of AI systems under the EU AI Liability Directive. Conversely, Chapter 7 sheds light on the opportunities and challenges of using evidence of algorithmic bias to prove the presence of systemic racial disparities through the California Racial Justice Act. Chapters 6 and 7 can be most useful to an audience of policymakers and legal practitioners. Accordingly,

[17] https://oxco.video/

Chapter 6 has been published in the European Journal of Law and Technology[18]. This journal publishes on topics which relate generally to European issues in information and technology law. Chapter 7 has been developed with legal practitioners and data science experts working with the CRJA in San Francisco. Finally, to ensure that the results of the thesis reflect the interests of these different audiences, section 8.2.2 of the conclusion addresses the policy recommendations ensuing from each chapter to the corresponding audiences mentioned here.

1.7 Thesis Outline

The field of algorithmic bias has been growing at a fast pace. This is especially true at the moment of submission (April 2024), with the deployment of foundation models bringing about new cases of bias and discrimination across sectors in society (Martineau, 2022). An integrated thesis has allowed me to engage with the field in a dynamic and timely manner, allowing me to publish chapters in the form of articles or papers throughout my DPhil. For this purpose, I conceived my thesis in a modular fashion: as a system of components that can be separated and recombined without the meaning being lost. Below, I provide a brief description of each chapter for the reader to navigate my thesis with clarity and driven by their interest.

Some of these chapters have been completed in collaboration with other researchers, including my supervisors Prof. Luciano Floridi and Prof. Mariarosaria Taddeo. Importantly, I am the sole first author of each chapter. This means that I have formulated the research questions, gathered, and analysed the data, and wrote the content of the chapters. The co-authors contributed with their domain expertise, and they provided

[18] https://ejlt.org/index.php/ejlt

feedback to sharpen the content of each chapter. Appendix A-G provides statements from each co-authors that clarifies and confirms my role.

Chapter 2: Literature Review

Rather than engaging in a gap-spotting exercise, the literature review engages in an exercise of problematisation: it challenges the ways in which algorithmic bias is already discussed in the literature at the crossing of Computer Science, Science and Technology Studies (STS), Law and Policy. It does so by unearthing some rarely challenged and under-looked assumptions; the assumption of 1. Algorithmic bias as a (technical) problem, of 2. Fairness metrics and XAI tools as solutions for it, and of 3. AI Policy's focus on automating solutions, rather than addressing existing societal problems. It draws links between these assumptions and what motivates each thesis chapter.

Chapter 3: Algorithmic Bias: To Preserve or to Transform?

In this conceptual chapter, I argue that algorithmic bias can be conceptualised both as a problem and an indicator of societal disparities. Given that fairness metrics can be classified into bias preserving and bias transforming metrics (Wachter et al., 2021a), I argue that one should preserve the *evidence* of societal disparities carried in algorithmic bias, and, through this evidence, inform interventions to address the problems that these disparities represent. To do so, I propose to situate fairness metrics, one of the proposed solutions to tackle algorithmic bias, within the context of substantive fairness (Green, 2021b); namely, by shifting their goal from achieving 'fair(er)' predictions toward supporting a broader set of interventions to address existing disparities. I operationalise this proposal in two steps. First, I introduce a 'thick' conception (Williams, 1985) of bias: one that recognises both a descriptive and a prescriptive aspect of algorithmic bias. I provide examples from credit, health, and crime risk assessment to illustrate how

algorithmic bias, as a thick concept, can be both a problem and evidence of existing societal disparities. Second, I adopt Wachter et al. (2021)'s characterisation of fairness metrics into 'bias preserving' metrics, which maintain demographic disparities, and 'bias transforming' metrics, which mitigate these disparities. I apply this division to two specific fairness metrics, namely demographic parity and calibration. I use these frameworks to explore how fairness metrics could contribute to equity-driven fairness interventions in these three areas.

Chapter 3 has been submitted as an article to Springer's Digital Society Journal and is currently under review. The chapter has undergone minimal editing to ensure coherence with the thesis as a whole.

Chapter 4: A Genealogical Approach to Algorithmic Bias

The Fairness, Accountability, and Transparency (FAccT) literature tends to focus on bias as a problem that requires *ex post* solutions (e.g., fairness metrics), rather than addressing the underlying social and technical conditions that (re)produce it. In this chapter, I propose a complementary strategy that uses genealogy as a constructive, epistemic critique to approach algorithmic bias in terms of the conditions that enable it. I focus on XAI feature attributions (Shapley values) and counterfactual approaches as potential tools to gauge these conditions and offer two main contributions. One is constructive: I develop a theoretical framework to classify these approaches according to their relevance for bias as evidence of societal disparities. I draw on Pearl's ladder of causation (2000, 2009) to order these XAI approaches concerning their ability to answer fairness-relevant questions and identify fairness-relevant solutions. The other contribution is critical: I evaluate these approaches in terms of their assumptions about the role of protected characteristics in discriminatory outcomes. I achieve this by building on Kohler-Hausmann's (2019)

constructivist theory of discrimination. I derive three recommendations for XAI practitioners to develop and AI policymakers to regulate tools that address algorithmic bias in its conditions and hence mitigate its future occurrence.

This chapter has been published as a peer-reviewed article in the Springer Journal Minds & Machines (see (Ziosi et al., 2024)). The chapter has undergone minimal editing in order to ensure coherence with the thesis as a whole. The article was co-authored with Dr. David Watson and Prof. Luciano Floridi. Please see Appendix A and B for authorship statements.

Chapter 5: Evidence of What, for Whom? The Socially Contested Role of Algorithmic Bias in a Predictive Policing Tool

This chapter presents a critical, applied study of the social role of algorithmic bias in the context of the Chicago crime prediction algorithm, a predictive policing tool that forecasts when and where in the city crime is most likely to occur. Through interviews with 18 Chicago-area community organisations, academic researchers, and public sector actors, I show that stakeholders from different groups articulate diverse problem diagnoses of the tool's algorithmic bias, strategically using it as evidence to advance criminal justice interventions that align with stakeholders' positionality and political ends. Drawing inspiration from Catherine D'Ignazio's taxonomy of 'refusing and using' data (D'Ignazio, 2024), I find that stakeholders use evidence of algorithmic bias to *reform* the policies around police patrol allocation; *reject* algorithm-based policing interventions; *reframe* crime as a structural rather than interpersonal problem; *reveal* data on authority figures in an effort to subvert their power; *repair* and heal families and communities; and, in the case of more powerful actors, to *reaffirm* their own authority or existing power structures. I identify the implicit assumptions and scope of these varied uses of algorithmic bias as

evidence, showing that they require different (and sometimes conflicting) values about policing and crime. This divergence reflects long-standing tensions in the criminal justice reform landscape between the values of liberation and healing often centered by communities impacted by incarceration and the values of surveillance and deterrence often instantiated in data-driven reform measures. I conclude by reflecting on the possibility of using algorithmic bias as evidence to challenge the status quo to protect the interests of marginalised, affected communities.

A version of this chapter has been accepted to the Fairness, Accountability and Transparency (FAccT) in ML 2024 conference. The chapter has undergone editing in order to ensure coherence with the thesis as a whole. The article was co-authored with Dr. Dasha Pruss. Please see Appendix C for authorship statements.

Chapter 6: Evidence of bias. Bridging Information Gaps in The EU AI Liability Directive (AILD)

The proposed European AI Liability Directive (AILD) is an important step towards closing the 'liability gap', i.e., the difficulty in assigning responsibility for harms caused by AI systems. However, if victims are to bring liability claims, they must first have ways of knowing that they have been subject to algorithmic discrimination or other harms caused by AI systems. This 'information gap' must be addressed before a discriminated user can build a claim of algorithmic discrimination and for the AILD to meet its regulatory objectives. Through qualitative analysis of the AILD proposal document (H. Morgan, 2022), this chapter examines its capacity to close the information gap and ultimately, meet its regulatory objectives. I argue that the current version of the AILD reduces legal fragmentation but not legal uncertainty; privileges transparency and disclosure of evidence of high-risk systems over notification of harm and discrimination; and shifts the burden on

the claimant from proving fault to accessing and understanding the evidence provided by the defendant. I conclude by providing four recommendations on how to improve the AILD to address the 'information gap' and indirectly the 'liability gap', necessary steps toward using evidence of algorithmic bias to inform the regulatory framework.

This chapter has been published in the European Journal of Law and Technology (see (Ziosi et al., 2023)). The chapter has been expanded in order to ensure coherence with the thesis as a whole. The original article was co-authored with Dr. Jakob Mokander, Dr. Claudio Novelli, Prof. Edoardo Celeste, Prof. Luciano Floridi and Prof. Mariarosaria Taddeo. Please see Appendix A-G for authorship statements.

Chapter 7: Exploring Algorithmic Bias as Evidence in the California Racial Justice Act: Opportunities, Challenges, and Policy Implications

In California, the use of risk assessment instruments (RAIs) and predictive policing tools has raised concerns regarding their reinforcement of existing racial inequities in the CJS, especially through algorithmic bias. Despite regulatory efforts like Proposition 25, which aimed to limit the use of RAIs, their deployment persists across California's counties, alongside predictive policing algorithms in major cities. The California Racial Justice Act (CRJA) (2020) (AB 2542) emerges as a ground breaking initiative to actively address racial injustices within the state's CJS. By allowing statistical disparities as evidence of intentional discrimination, the CRJA challenges legal precedents and aims to rectify systemic racial biases ingrained in the justice system. This chapter conducts a qualitative analysis of the CRJA document (H. Morgan, 2022), exploring its potential to incorporate algorithmic bias as evidence to combat racial disparities in charges, convictions, and sentencing. I argue that algorithmic bias, as an inherent feature of these algorithms, represents an alternative form of evidence – as a 'trace' – that can complement more

traditional sources used to build discrimination claims in courts. The findings offer insights to policymakers and stakeholders in the CJS, aiming to inform decisions and actions regarding the integration of algorithmic bias in the CRJA and in regulatory initiatives more widely.

This chapter has been submitted as an article to the Big Data & Society Journal and it is currently being considered for review. The chapter has undergone minimal editing in order to ensure coherence with the thesis as a whole.

Chapter 8. Conclusion

Going back to the main research question, the conclusion of the thesis culminates in a *policy* output: a series of policy recommendations tailored to the audiences of the specific chapters, accompanied by some overarching considerations on how algorithmic bias can inform policies around bias, fairness and discrimination. It also presents a synthesis of contributions and findings of the thesis and, importantly, it outlines a series of limitations and avenues for future research.

The next sections will present the above chapters in full, in the order expressed in this thesis outline.

Chapter 2:

Literature Review

Chapter 2. Literature Review

In this literature review, I present how the academic community has approached the three principal elements of the main research question in three sections; 2.1 algorithmic bias, 2.2 risk prediction tools, and 2.3 policies on AI with a focus on bias and fairness. In section 2.1, I show that the literature on algorithmic bias leaves unspecified the link between the types of bias and how they differentially relate to inter-group disparities in predictive performance. Additionally, I suggest that the literature on algorithmic fairness often presumes bias as the central issue that fairness metrics ought to fix. This assumption may inadvertently discourage efforts to address what makes bias problematic in the first place: the conditions that give rise to it. Although explainability approaches are proposed as a solution, there is limited clarity regarding their relationship with bias and fairness. Specifically, it remains unclear which XAI approach to employ to address which aspect of bias. In section 2.2, I argue that the literature unquestioningly assumes that the sole purpose of risk tools is to predict outcomes and, where this is not the case, this technology is occasionally used for intervention, a task for which they are not explicitly designed. Concerning bias and fairness in AI policy, in section 2.3. I suggest that there is a strong focus on how to best automate 'fair' solutions and less on how to tackle existing problems around fairness and bias in society. While the latter aspects are typically tackled by economists and sociologists, their insights do not seem to hold a central place in the AI policy discourse.

Overall, the literature review introduces the main themes that the core chapters of this thesis focus on. As a narrative critical review (Stratton, 2019), this literature chapter stands, on the one hand, as a narrative introduction to the research topics and, on the other hand, it provides critical considerations that motivate further research conducted in the specific chapters (Stratton, 2019). This method of review entails describing and

synthesising the work in the field in a way that establishes the context of research. At the same time, however, it also allows one to express opinions and draw considerations on the materials reviewed (Stratton, 2019).

2.1 Algorithmic Bias

I defined bias in section 1.1. Here, I elaborate on how the academic literature around fairness, accountability and transparency approaches it as a problem (Section 2.1.1) as well as evidence (Section 2.1.2) of existing societal disparities. For each of these sections, I also review the solutions that are typically associated with each of these standpoints: fairness metrics for the first and explainability methods for the second.

2.1.1 Bias as a Problem

From a technical standpoint, bias is a property of a system rather than a problem, per se (Mullaney et al., 2021). Bias can be used positively to tailor a certain model around the specific needs of a vulnerable group or, in the guise of positive discrimination, it can play a mitigating role against a model that favours a majority population over a minority one (Chander, 2017). As a result, bias is not problematic in itself. However, as mentioned in the introduction, it can become so when it aligns with and perpetuates underlying societal inequalities, such as those related to gender and race. This is when I refer to it as a 'problem'. There is a growing amount of literature that has been focusing on characterising this 'type' or aspect of bias. It tends to do so by describing it in terms of its types or sources as well as its manifestations in predictive performance (Ashoka, 2013; Hurley & Adebayo, 2017; Traub, 2013). However, it does not provide an exhaustive account of how these elements are conceptually related, specifically in terms of how they might be differentially reflected in predictive performance.

In the literature, algorithmic bias is often attributed to *distortions* – a "skew" - in

the data or in the predictive process (Mayson, 2018) resulting in measurable effects on predictive performance and leading to problems such as disparate impact or disparate treatment towards a certain group (Green, 2020). In this respect, the data on which a certain algorithm learns or on which a certain method is applied can be incomplete or wrong, non-representative, and skewed (Crawford et al., 2019). This can depend on characteristics such as the quality of the data, e.g., whether it is incorrect or non-representative (Barocas & Selbst, 2016; Crawford, 2013); on its provenance, e.g., whether the data on which a system is trained shares any similarity with the data it is tested on or with the context it is applied in (Danks & London, 2017); and on the underlying distribution, e.g., whether it fails to represent different groups in accurate proportions (Baeza-Yates, 2013; Barocas & Selbst, 2016).

It may also depend on which outcome measures or target variables are used and on how they are measured (Barocas & Selbst, 2016; Daumé, 2017) or on which input variables (features) are selected (Olteanu et al., 2019). Both the choice of target variables and of features might include elements that work as proxies for protected categories such as race or gender. For example, as I will discuss in the upcoming chapters, a healthcare algorithm that relies on the cost of patients' past medical care to predict their future medical needs can underestimate health risk for Black patients (Obermeyer et al., 2019). Generally, Black patients have less access to medical care and they tend to generate less costs than White patients with similar health needs (Benjamin, 2019a).

Additionally, *distortions* may also depend on the method used for the predictive process (Mayson, 2018). Risk prediction models based on supervised learning (e.g., classification trees or logistic regression models), for example, entail assumptions and method-specific mechanisms that might make them more prone to certain types of distortions in predictive performance (Fuster et al., 2020; Mehrabi et al., 2019). Firstly,

these models prioritise highly predictive variables (Barabas et al., 2018). This means that they tend to treat factors belonging to a dominating population as 'universal', thus ignoring factors belonging to sub-groups in the data (Greiner, 2008). Secondly, these models favor individual factors, such as psycho-social factors, over broader structural drivers of predictive outcomes, unduly eliminating the effects of the latter on prediction results (Morley & Floridi, 2020). Finally, regression-based models struggle to differentiate intermediate outcomes from covariates without a causal inference model (Greiner, 2008). This might lead to a misinterpretation a model's results, reinforcing the potential for bias. Intensive policing, for example, causes more re-arrests and anti-law enforcement resentment ('antisocial cognition'). Antisocial cognition, however, is an intermediate outcome. It is caused by intensive policing and it should not be interpreted as a causal driver of crime (Prins & Reich, 2017).

The literature on bias has listed an increasing number of different types of bias, as in the above. However, it often fails to specify how each type might be differentially reflected in predictive performance. Admittedly, this is no easy task. Distortions in the data and in the predictive process might not be the sole reason for inter-group disparities in predictive performance. The latter could also suggest inherent *differences* in the event to predict (Mayson, 2018), which complicates the task at hand. Inter-group disparities in predictive performance might be a product of different groups behaving differently, such as having different incidence or base rates (Hao & Stray, 2019). Testing the difference between noise (distortions) and signal (real differences) represents a challenging task.

2.1.2 Fairness Metrics as a Solution?

As Chapter 3 will present, part of the reply from the research community has been to remain agnostic toward the specific reason for that disparity, focusing on devising metrics that will adjust predictive performance according to a selected standard of fairness (S.

Mitchell et al., 2021). While these efforts are important to minimise the negative consequences of bias in the specific contexts of deployment, they should be also informed by what might give rise to this bias in the first place. These efforts can be traced in the literature that focuses more specifically on 'algorithmic fairness' (Barocas et al., 2022; Berk et al., 2017; Calders & Verwer, 2010; Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2011; Feldman et al., 2015; Hajian & Domingo-Ferrer, 2013; Kamiran & Calders, 2012; Kleinberg, 2018).

There are a host of bias mitigation techniques varying from pre-processing, in-processing and post-processing methods (Wan et al., 2023). Pre-processing methods tackle bias at the level of the training data. They aim to address signals and correlations that contribute to bias and discrimination (Biswas & Rajan, 2021). Examples are re-weighing and 'massaging' the dataset (Nazer et al., 2023). In-processing techniques integrate fairness constraints directly into the optimization process of the model during training (Wan et al., 2023). Examples include Fairness-aware Data Augmentation (Köse & Shen, 2021) and Adversarial Debiasing (Kumar, 2018). Post-processing techniques involve adjusting the model's output to mitigate bias after the model has been trained (Lohia et al., 2018). Examples are Calibrated Equalized Odds and Reject Option Classification (Fernandez, 2023).

Notwithstanding this variety of techniques, the literature on algorithmic fairness has been increasingly engaged in the formalisation of fairness criteria from political philosophy e.g., the Rawlsian notion of equality of opportunity (Lee et al., 2020) , or from law, e.g., disparate treatment and disparate outcomes (Green, 2020; Wachter et al., 2021a) into formal metrics against which to evaluate predictive performance. This field has led to the creation of a host of so-called 'fairness metrics' against which predictive performance is to be adjusted to conform to the above ideas of fairness. Examples are 'statistical parity',

'equalised odds', 'predictive parity', and 'calibration', to cite a few. Statistical parity states that the probability of a positive prediction should be equal across protected groups (Castelnovo et al., 2022). Equalized odds requires that the true positive and false positive rate are equal across all protected groups (Garg et al., 2020). Predictive parity and calibration will be elaborated upon in Chapter 3.

Additionally, some fairness metrics formalize a causal understanding of fairness. This is the case of 'counterfactual fairness' (Kusner et al., 2017). In practice, counterfactual fairness methods analyse whether an outcome would change if the protected attribute was different while keeping all other relevant factors constant. This is based on the intuition that 'similar' – according to a defined criterion – individuals should be treated similarly. An exhaustive list can be found in 'Fairness and Machine Learning' (Barocas et al., 2022).

However different, the metrics respond to the general question guiding most of this field of research: '*How can we predict more fairly?*' Concerningly, these metrics have been found to be in contrast both with each other and with the demands of predictive accuracy (Berk et al., 2017). Kleinberg et al. (2016) proved that no three fairness metrics can be simultaneously satisfied but in the case of perfect classification or equal base rates; plainly, unless there are no errors in prediction or groups behave the same. For instance, Chouldechova (2017) showed how satisfying predictive parity was at odds with obtaining equal false positive (FP) and false negative (FN) rates across groups when base rates differ across those groups. Hardt et al. (2016) confirmed this impossibility even for any two metrics.

On a more conceptual level, Green (2020) and Narayanan (2018) remarked how equality of outcomes and equality of treatment simply cannot be simultaneously satisfied

in an unequal world. Finally, a survey by Berk et al. (2017) presented not only how these metrics and their variations are in conflict with one another, but also with the request of predictive accuracy maximisation. This has also been confirmed in the work of Dwork et al. (2011), Corbett-Davies et al. (2017) and Kamiran & Calders (2012), who have resorted to operationalising any different conception of fairness as a constraint to predictive accuracy.

Overall, this apparent impossibility of finding a solution has been attributed to formal causes (Berk et al., 2017) or to contrasting demands for fairness in society (Narayanan, 2018). Nevertheless, this impasse could also be telling of the attempt of the research community to solve for bias by adjusting predictive performance to fairness metrics, without tackling the societal conditions for its occurrence in the first place. This importantly reveals two separate goals that should not be conflated; building an algorithm that is not discriminatory and shedding light on what gives rise to bias.

2.1.3 Bias as Evidence

The concept of bias as 'evidence' is an aspect of bias that, even though more rarely, appears in the literature. On the one hand, bias in algorithms can reflect human bias, in terms of normative assumptions and technical decisions embedded in algorithms (Abebe et al., 2020; Malik, 2020). This intuition has been at the core of research efforts in disciplines such as science and technology studies (Agre, 1997; Jasanoff, 2006; Winner, 1980), law (Citron, 2008; Lessig, 1999; Reidenberg, 1997) and values-in-design (Friedman & Hendry, 2019; Lucas D. Introna, 2000). The assumptions of developers, the composition of their teams, and their decision about which proxies to use to measure an abstract construct are factors that can be reflected in biased performance.

On the other hand, bias in algorithms can represent a mirror of existing demographic disparities in society (Abebe et al., 2020; Benjamin, 2019a; Eubanks, 2018; Mayson, 2018; Zajko, 2021). For example, Zajko (2021) presents bias as a 'computalisation' of social inequality or as the intersection of different structures of inequality that are reproduced by algorithms. Referring specifically to bias in risk assessment algorithms in the criminal justice system in the US, Barabas et al. (2018) and Mayson (2018) refer to it as a reflection of disparate arrest practices towards people of color. Referring to bias in health risk assessment, it has been found that using health expenses as a proxy to measure health needs across populations can reflect socio-economic disparities, e.g., ability to pay and differential access to healthcare (Obermeyer et al., 2019), as well as unequal health insurance structures (Benjamin, 2019a). In the case of credit risk assessment, Blattner & Nelson (2021) found that differential noise in credit scores across demographic groups, when due to features of the underlying credit data, can carry evidence of past disparities in credit access. They attributed this differential noise to differences in the composition of credit report data such as sparsity in the sense of having data from few loans or few years of history, past default history, and account diversity in the sense of holding multiple types of loans (Blattner & Nelson, 2021). Chapter 3 will further delve into these examples.

While useful, these conceptualisations of bias as evidence tend to remain conceptually underspecified and vague. Except for the specific examples presented above, they rarely focus on a specific and concrete aspect of bias, such as proxies, and on a specific context. They also often fail to be conceptually pragmatic, as they usually stop at the point of critique. As a rare exception, Eubanks (2018) proposes to conduct a structural analysis of bias to identify which more systematic changes would be effective, including where automated decision-making solutions should be removed from the process (Eubanks,

2018). When being pragmatic, some might still place their primary focus on identifying, removing or managing statistical biases as the main solution, leaving the inquiry of what gives rise to bias open (Fazelpour & Danks, 2021). It is in this context that explainability approaches tend to be considered as fairness-relevant tools to shed light on the relationship between protected features and skewed performance.

2.1.4 Explainability Tools as a Solution?

When the focus is shifted from perceiving bias solely as problematic due to its outcomes, toward considering the insights bias may provide as evidence, it becomes feasible to recognise explainability approaches as a suitable remedy. This is because, differently from fairness metrics, their primary aim is to shed light on, rather than to remain agnostic toward, the factors that contributed to a biased outcome. There is an increasing number of ML libraries that advertise tools that are relevant both for explainability and fairness. Examples are IBM360[19] and WhatIf tools[20]. There is also a growing set of explainability methods[21] as well as fairness methods[22] which can be used alone or in concert to mitigate algorithmic bias.

While I presented fairness metrics above, examples of the former are feature-attribution approaches such as SHAP (Shapley Additive exPlanations) (Lundberg & Lee, 2017). Such methods can estimate how input features contribute to a biased performance (Begley et al., 2020). This can help one understand how a model has arrived at a particular discriminatory outcome (Packt, 2018). Another example is methods for counterfactual

[19] AI Explainability 360 (https://aix360.mybluemix.net/) and AI Fairness 360 (https://aif360.mybluemix.net/)

[20] https://pair-code.github.io/what-if-tool/

[21] For an encompassing review of existing explainability methods, see Carvalho et al., (2019) and Islam et al. (2022)

[22] For an overview of the pre-processing, in-processing and post-processing techniques for fairness see Rovatsos et al. (2019)

explanations. These explain what could have happened to an outcome had an input feature to a model been changed in a particular way (Barocas et al., 2020; S. Verma et al., 2020). By showing how predictions change when protected attributes are altered, these explanations can reveal discriminatory patterns. Counterfactual explanations can also point users towards actionable recourse for an undesired negative outcome by suggesting alterations to their features that could improve their outcome (e.g., suggesting a loan applicant to increase their income). These approaches vary both in the type of counterfactuals and of explanations (VanNostrand et al., 2024). Regarding the former, counterfactual explanations can be point-based as well as region-based. While the former identify an instance – a 'point' –, the latter focus on identifying a region or a range in the feature space in which the prediction changes (VanNostrand et al., 2024). Regarding the latter, these approaches vary in how the counterfactual information is conveyed, ranging from straightforward numerical representations (Ustun et al., 2019) to textual descriptions (Slack et al., 2023), and even visual illustrations (Gomez et al., 2020).

These tools increase in number and get updated at a fast pace. However, they are also often introduced without specifying a clear link between the potential of explainability for fairness. Specifically, it remains unclear which XAI approach to employ to address which aspect of bias. While Chapter 4 of this thesis will elaborate on specific XAI methods, I here focus on reviewing the relationship between algorithmic bias and explainability more in general. I suggest that the relationship between fairness and explainability is currently vague. Additionally, it is mostly focused on enhancing possibilities for fairness rather than on surfacing evidence of unfairness.

In terms of their relationship, explainability can enhance fairness (broadly understood) to different extents and on different levels. First, explainability can enhance transparency (Abdollahi & Nasraoui, 2018). This is relevant to fairness through

accountability, granted by the ability to see how a model has arrived at a discriminatory outcome. Second, explainability can increase or enable trust in a model (Dodge et al., 2019). It can help, for example, to determine if qualities relevant to algorithmic fairness are met (Doshi-Velez & Kim, 2017) by, for example, ensuring that an outcome was not arrived at or correlated with protected features. This is also associated with the idea of 'responsible AI', AI that takes into account moral and ethical considerations as well as societal values (Adadi & Berrada, 2018).

At the same time, explainability can also be detrimental to fairness. Explainability, for instance, can influence the *perception* of fairness. On that note, some highlight the risks for fairness that more explainability, and more reliance on it, can bring. Examples are the risks of 'fairwashing' and of the rationalization, and potential justification, of some types of discrimination (Aivodji et al., 2019). Concerning 'fairwashing', the utilisation of explainability tools for fairness may be presented as a responsible commitment to the public, yet function merely as a superficial checkbox or a marketing tactic to encourage more users to adopt an algorithm. Regarding the potential rationalisation of certain forms of discrimination, the capacity of XAI tools to elucidate which features serve as stronger predictors of an outcome might facilitate the justification of employing discriminatory proxies under the pretext of enhanced predictive accuracy. The relationship between explainability and fairness can also be a matter of trade-offs. Barocas (2022) highlighted the tensions between, on the one hand, requests for simpler models to ensure transparency and thus facilitate algorithmic fairness and, on the other hand, simpler models being less able to satisfy some fairness demands, e.g., allowing for specific fairness 'fixes' in a model.

On the one hand, the above portrays an unclear, and sometimes counterproductive relationship between XAI approaches and bias. On the other hand, where positive, this relation is framed as bringing about enhanced capabilities or more opportunities for

algorithmic fairness. This is expressed in values such as transparency, trustworthiness, and responsible AI, understood as fairness-relevant properties that follow from what explainability can offer. However, the focus of this thesis is different; it questions their ability to surface evidence of unfairness, e.g., of existing disparities in society.

2.2 Risk Prediction Tools

As presented in section 1.1, risk prediction tools are mathematical models fit on historical data, used to measure the risk of a certain future outcome in areas including but not limited to credit, e.g., the risk of default, criminal justice, e.g., the risk of someone re-committing a crime, and healthcare, e.g., the risk of heart failure (Express Analytics, 2022). While to different degrees, most of the risk prediction tools used in these areas feature a host of methods that can vary as widely as from statistical to supervised and unsupervised learning models (König & Krafft, 2021; Robinson Yu, 2014). They not only vary in model complexity but also in the type of data they use. The most basic models may use only a few demographic variables, such as age and employment status, to predict risk (Express Analytics, 2022). More complex models may use a variety of data sources, including medical records and genetic data for healthcare, and lifestyle and social media information for credit (Robinson Yu, 2014). They are joint in their purpose: to employ mathematical models to *predict* the probability of an outcome occurring.

Concurrently, as I will present below, history shows how the main task of credit, health and criminal justice institutions in the US has not always necessarily been to predict. Additionally, the present shows how these algorithms are not strictly used for prediction, but also, in some cases, to engage in interventions[23]. Given the inherent bias in

[23] As previously mentioned, this refers to purposive attempts by governments or organizations to achieve specific goals through regulations, incentives, public programs or market mechanisms (Wiseman, 1979)

these tools, the unwarranted extension of their use toward interventions risks perpetuating rather than challenging or changing the existing status quo. While each chapter will delve deeper into their respective topics, I here employ the case of actuarial (or 'criminal') risk assessment in the US Criminal Justice System as an overarching example to illustrate the considerations above. This is especially relevant given that Chapter 5 and Chapter 7 will deal more in-depth with the cases of predictive policing and crime risk assessment tools. Additionally, the history and methods employed in these actuarial risk assessment tools also influenced models that are nowadays used to calculate other types of risks covered in this thesis, such as credit and health (Lewin, 2007).

2.2.1 The Case of Criminal Risk Assessment

Actuarial risk assessment is a form of risk prediction based on assessing the likelihood of an individual (re)committing a crime based on a set of features as varied as criminal record, marital status, and employment among others. The history of actuarial risk assessment in the US Criminal Justice System (CJS) reveals how, starting in the 1920s, risk assessment was not initially based on statistical methods and the main aim of the assessment was to intervene on, rather than predict, an individual's future risk. These assessments were based on semi-structured clinical evaluations to point offenders toward rehabilitative treatments (Andrews et al., 2006). Yet, the beginning of the 1930s already witnessed the introduction of the first actuarial tool, the 'Illinois parole system', based on the Burgess method (B. E. Harcourt, 2008). The actuarial assessment of risk became increasingly commonplace in the 1970s with the widespread adoption of regression modelling for the prediction of risk in the CJS. This method guaranteed increased accuracy and it was perceived as more objective (Hannah-Moffat, 2013).

While this shift was motivated by the desire for accuracy and objectivity, it also came with some implications. On the one hand, regression entailed framing risk as a static

component, which suggests what should not be done, e.g., release, rather than what can be improved, e.g., rehabilitation (B. E. Harcourt, 2010). This marked a move away from interventions. On the other hand, conceptually speaking, it re-framed the focus of the CJS from rehabilitative interventions to the administration of punishment (Maurutto & Hannah-Moffat, 2006). Finally, this shift was accompanied by an increase in incarceration rates (Harcourt, 2010).

By the 1980s, these worrying outcomes motivated the expansion of the application of regression from previous 'risk' factors to more dynamic 'needs' and 'responsivity' factors, such as substance abuse and employment history for the former and psychological disorders and levels of self-esteem for the latter (Andrews et al., 2006). The aim of this inclusion was to reduce the risk of recidivism and thus, incarceration rates (Cullen & Gendreau, 2001). The main difference resided in the fact that the incorporation of more dynamic factors allowed the CJS to move beyond punishment, as it informed potential treatment interventions.

Confirming this trend towards interventions, the literature shows that today ML techniques are not strictly confined to the task of prediction, but their premises of competence in the arena of actuarial risk assessment are increasingly being extended to intervention on 'risk', 'needs' and 'responsivity' factors (Andrews et al., 2006). Linear regression techniques used for predicting crime can identify which variables are linked to an increase in crime risk. Nevertheless, they do not provide clear guidance on how much one needs to adjust each variable to effectively reduce future crime rates. Generally, these methods are not alone suitable for interventional approaches to risk reduction (Barabas et al., 2018).

This trend does not appear to be isolated to actuarial risk assessment, nor to the field of criminal justice alone. Recent efforts in predictive policing have proposed a repurposing of the application of predictive tools to redirect policy resources toward socio-economically disadvantaged areas (Rotaru et al., 2022). Some interventions are tailored to the type of crime that is expected in a certain area, e.g., violent crime vs burglary, and can take the form of placing a 'capable guardian' where a burglary is expected. Beyond crime prediction, health risk assessment algorithms such as OPTUM use insurance data to make predictions about population-level health outcomes. These predictions are then used to create interventions for patients with complex health needs (Obermeyer et al., 2019). These interventions direct patients towards programs where they receive additional resources and care from healthcare providers.

These considerations do not seek to undermine the overall utility and reliability of risk prediction tools. There is no conclusive evidence, for example, that human judges are better or less biased in assessing risk (Lin et al., 2020). Instead, by situating the discourse surrounding these tools within its broader historical and contemporary context, it becomes possible to achieve a more comprehensive understanding of their origins and to better inform decisions around their application in specific contexts. This encompasses instances where these tools are employed beyond their intended scope, as we have seen above, extending their predictive capabilities to inform or intervene in tasks of institutional or social significance. Additionally, when these tools exhibit algorithmic bias, their predictions may perpetuate historical patterns into future outcomes. As a result, there is a risk that the interventions informed by these predictions may reinforce, rather than challenge or alter, the existing status quo. As we will see in Chapter 3, even the use of some fairness metrics may not be sufficient to warrant such risks in some cases. Yet, differently from human judges, these tools' systematic reproduction of societal disparities at

scale can make of their skewed performance a focal point for investigating bias as a form of evidence.

2.3 (AI) Policy for Bias and Fairness

AI policy is a branch of public policy focused on curbing the adverse effects of AI (Floridi & Cowls, 2019; Jobin et al., 2019), as well as on harnessing its potential (Floridi et al., 2020). As mentioned in section 1.1, this is achieved through a set of hard and soft law initiatives that aim to create rules or guidance for designing, developing, and/or deploying AI. In the following paragraphs, I first provide a general overview of the current global AI policy landscape. Then I focus on how these policies approach the problem of algorithmic bias and fairness, specifically in the US. Overall, I argue that most of the rules and guidelines issued in AI policy assume the goal or aim of AI systems to be the ethical or fair automation of human tasks (Greene et al., 2019). There has been relatively little focus, in this specific field, on addressing the current societal issues related to bias and discrimination.

2.3.1 The Global AI Policy Landscape

Since 2016, policymakers around the world have increasingly turned their attention to the development and implementation of strategies and initiatives for governing AI systems (Roberts et al., 2023). The AI policy landscape has grown more defined over the past years, with several governments proposing regulatory policies for AI within their jurisdictions.

In the US, for example, the Blueprint for an AI Bill of Rights (2022) outlines a set of high-level principles, and it explains how they can be enforced through existing federal- and state-level legislation within particular sectors. More recently, on October 30th, 2023, President Biden issued an Executive Order (EO) on Safe, Secure, and Trustworthy Artificial Intelligence (2023). This EO sets forth a federal government-wide strategy,

encouraging Federal agencies to undertake various initiatives related to AI safety and security, privacy, equity, civil rights, consumer, patient, and student protections, workers' rights, AI innovation and competition, international leadership and engagement, and responsible and effective utilisation of AI by the government. Additionally, the order establishes the White House AI Council to supervise these endeavors.

In March 2023, the UK published its AI Regulation White Paper, A Pro-Innovation Approach to AI Regulation (2023). The White Paper proposes a sector-led approach that relies on regulators to address the impacts of AI in their specific context. In March 2024, the EU AI Act was approved by the European Parliament (Mackrael & Schechner, 2023). The AI Act introduces requirements and proportionate obligations for AI systems which are categorised according to four risk levels, from unacceptable risk (banned) to no risk. It also introduces a specific tiered approach for foundation models. The Act is interconnected with a series of other AI regulatory initiatives in the EU, including a civic liability framework (AI Liability Rules), accompanied by a revision of sectoral safety legislation (i.e., Machinery Regulation, General Product Safety Directive), as well as an enhancement of the regulations governing digital services (Digital Markets Act, Digital Services Act). China has introduced several different legislative initiatives to target specific types of AI technologies. This includes regulations for recommender systems (2022), 'deep synthesis' technologies (2022), and generative algorithms (2023). Recently, countries such as Brasil (Sakiotis et al., 2023), Canada (Christiaen, 2022) and South Korea (Roh & Nam, 2023) have also engaged in similar efforts to regulate AI[24].

At the same time, there have been numerous efforts to establish international agreement on AI governance, such as through Organisation for Economic Co-operation

[24] I am aware that, from a 'global' standpoint, this is by far not an exhaustive list.

and Development's (OECD) AI Principles and OECD Expert Working Groups, the Council of Europe's Committee on AI, the UNESCO Ethics of AI Principles, the Global Partnership on AI, and the EU-US Trade and Technology Council, amongst others. Despite the many different initiatives, the common goal is to promote discussion and create shared frameworks for AI governance (Roberts et al., 2023).

2.3.2 AI Policy for Bias and Fairness

As issues around risk prediction tools have been increasingly surfacing in AI Policy documents (Crawford et al., 2019; European Commission, 2021; Future of Life Institute, 2017; UNESCO, 2020) and as the final goal of this thesis is to inform policy, it is important to review how the field currently approaches issues around bias in algorithmic tools.

On the one hand, many of these policy documents outline ethical guidelines to promote algorithmic fairness, understood more often as prevention, mitigation, and monitoring of unwanted bias and more rarely as avoidance of discrimination (Jobin et al., 2019), in compliance with existing discrimination law. On the other hand, some of this policy work advocates using algorithms to ensure fairness and minimise human bias in decisions (Future of Life Institute, 2017; Kleinberg et al., 2017; Silberg & Manyika, 2019). Notwithstanding the importance of creating hard and soft procedures to avoid the risks and maximise the benefits of AI in society (Floridi et al., 2018), most of this work assumes the goal or aim of these systems to be the ethical or fair automation of human tasks (Greene et al., 2019). Concurrently, little emphasis has hitherto been placed by AI policy initiatives on solving existing societal problems around bias and discrimination.

In the US, this can be observed in how current efforts to tackle algorithmic bias and discrimination in AI policy are mostly focused on accountability and transparency. The

Algorithmic Discrimination Protections section listed in the Blueprint for an AI Bill of Rights (2022) echoes several cases of algorithmic bias I cited under credit (Student Borrower Protection Center, 2020), health (Obermeyer et al., 2019) , and crime risk assessment (Johnson, 2022a, 2022b; National Institute of Justice, 2021). The section recommends several measures, such as conducting demographic disparity assessments to document any demographic disparities within the AI system, using disparity mitigation to evaluate the adverse impact of multiple models and choosing the one with the least adverse impact, and ongoing monitoring and mitigation, such as conducting ongoing disparity assessments against equity standards to tackle any biases that may arise.

Similarly, the US National Institute for Science and Technology (NIST)'s 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence' (Schwartz et al., 2022) introduces preliminary guidance for addressing bias in datasets, in testing and evaluation, and in human factors in AI systems. It encourages system and procedural transparency through datasheets, model cards, and documentation of potential sources of human bias. It also suggests promoting accountability by conducting impact assessments, such as such as identifying risks and documenting any mitigation activities that can be taken to prevent potential harm.

The re-proposed Algorithmic Accountability Act (2023) may require organisations to conduct impact assessments, to create transparency around when and how AI systems are used, and to reasonably address any identified biases or security issues in a timely manner. Accountability and transparency in the development and deployment of AI systems are important values to ensure. However, they run the risk of assuming that the main goal of AI Policy concerning algorithmic bias is the fair automation of existing solutions; namely, they risk endorsing most AI applications as long as their process is documented and accounted for by the above procedures (Kaminski, 2022).

Much less attention, for example, has been focused on the potential of algorithmic systems to be used for 'diagnostic' purposes; that is, to detect existing problems rather than to automate decisions (Xiang, 2020), and on how public policy can take advantage of that. In the literature, there are rare invitations to use computing tools as a diagnostic to help us *'to understand and measure social problems with precision and clarity'* (Abebe et al., 2020, p.252) as well as to make **'***long-standing social problems newly salient in the public eye'* (Abebe et al., 2020, p.252). These remain a general invitation, not directly concerned with addressing current societal issues related to bias and discrimination. Additionally, as presented above, some authors started to question the default use of algorithmic systems for prediction rather than for intervention on existing problems (Barabas et al., 2018) and to speculate on their potential use as a mirror of social problems (Mayson, 2018). However, none of these efforts has resulted in the formulation of a concrete policy framework, from guidelines to other alternatives, for implementation that matches the efforts in AI policy towards systems used for prediction. Chapters 6 and 7 of this thesis 'deep dive' into the possibilities of this alternative policy potential.

Chapter 3:

How can bias as evidence be conceptualised?

Chapter 3. Algorithmic Bias: To preserve or to transform?

Abstract

In this chapter, I argue that algorithmic bias can be conceptualised both as a problem and an indicator of societal disparities. Given that fairness metrics can be classified into bias preserving and bias transforming metrics (Wachter et al., 2021a), I argue that one should preserve the *evidence* of societal disparities carried in algorithmic bias, and, through this evidence, inform interventions to address the problems that these disparities represent. To do so, I propose to situate fairness metrics, one of the proposed solutions to tackle algorithmic bias, within the context of substantive fairness (Green, 2021b); namely, by shifting their goal from achieving 'fair(er)' predictions toward supporting a broader set of interventions to address existing disparities. I operationalise this proposal in two steps. First, I introduce a 'thick' conception (Williams, 1985) of bias: one that recognises both a descriptive and a prescriptive aspect of algorithmic bias. I provide examples from credit, health, and crime risk assessment to illustrate how algorithmic bias, as a thick concept, can be both a problem and evidence of existing societal disparities. Second, I adopt Wachter et al. (2021)'s characterisation of fairness metrics into 'bias preserving' metrics, which maintain demographic disparities, and 'bias transforming' metrics, which mitigate these disparities. I apply this division to two specific fairness metrics, namely demographic parity and calibration. I use these frameworks to explore how fairness metrics could contribute to equity-driven fairness interventions in these three areas.

Note: This chapter has been submitted as an article to Springer's Digital Society Journal and is currently under review. The chapter has undergone minimal editing in order to ensure coherence with the thesis as a whole.

3.1 Introduction

In the realm of risk prediction, where algorithms are used to calculate future risk based on historical data, algorithmic bias simultaneously perpetrates and reflects existing societal disparities. A credit risk assessment tool can result in unfair credit allocations while mirroring previous redlining divisions across protected groups (Blattner & Nelson, 2021; Citron & Pasquale, 2014). A health risk assessment tool can underestimate health risks for people of color while mirroring differential access to health insurance across the population (Obermeyer et al., 2019). In the context of recidivism, bias can lead to overestimating the likelihood of recommitting a crime while mirroring disparate arrest practices toward people of color (Angwin & Larson, 2016). 'Algorithmic bias', in this context, is generally taken to refer to cases in which 'predictive performance (however defined) unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class' (Mitchell et al., 2021, p.1). This bias is also referred to as a 'skewed performance' along one of these social or demographic axes.

Fairness metrics are presented as a solution in this context. These refer to a set of measures that enable one to detect bias in a model, to consequently adjust it. As mentioned in the literature review (Chapter 2), these measures follow different definitions of fairness, reflecting ideas from political philosophy (ranging from equality of opportunity to equality of outcomes) and law (ranging from disparate treatment to disparate outcomes) (Lee et al., 2020). Despite these differences, fairness metrics aim to achieve 'fair(er)' predictions. Namely, they put constraints on the distribution of predictions across different populations. These metrics, however, are in contrast with each other, as well as with predictive accuracy. Some scholars attribute this to the so-called impossibility theorem (Chouldechova, 2017; Corbett-Davies et al., 2017; Kleinberg et al., 2016). The theorem states that, unless base-rates are the same or there are no errors in predictions, it is formally

impossible to satisfy multiple metrics simultaneously. As a result, they argue for trade-offs between different fairness conceptions. Others argue that multiple fairness metrics, such as equality of outcome and equality of treatment, are necessarily at odds in a world grounded on societal disparities (Narayanan, 2018; Green, 2020).

In this chapter, I argue that the problem with fairness metrics might also lie in the attempt of the research community to address bias by addressing predictive performance, without taking into consideration the societal disparities that bias makes manifest. I aim to address this concern by situating fairness metrics within the context of substantive fairness (Green, 2021b); namely, by shifting their goal from achieving 'fair(er)' predictions toward supporting a broader set of interventions to address existing injustices. As mentioned in Chapter 2, interventions refer to purposive attempts to change the social conditions under which people live (Wiseman, 1979). Of course, fair predictions can be understood as interventions (e.g., adjusting a biased outcome). However, framing fairness metrics more broadly as interventions can open the door for a host of other interventional roles that metrics can play. I do so by introducing a 'thick' conception (Williams, 1985) of bias: one that ascribes both a descriptive and a prescriptive aspect to algorithmic bias. I provide examples from credit, health, and crime risk assessment to illustrate how algorithmic bias, as a thick concept, can be both a problem and carry evidence of existing societal disparities. Second, I adopt Wachter et al. (2021)'s characterisation of fairness metrics into 'bias preserving' metrics, which maintain societal disparities, and 'bias transforming' metrics, which mitigate these disparities. I apply this division to two specific fairness metrics, namely demographic parity and calibration. These two metrics are some of the most commonly used to assess bias in risk assessment (Barocas et al., 2022). In their work, Wachter et al. (2021) have predominantly focused on the concept of bias preserving

metrics. Here, I further develop their interplay with bias transforming metrics in support of equity-driven fairness interventions.

Overall, I argue that one should preserve the evidence of societal disparities carried in algorithmic bias, and, through this evidence, inform interventions to address these disparities. However, this chapter provides a more nuanced answer. Bias preserving metrics, such as calibration, can be used as a diagnostic tool to measure societal disparities and gain a deeper understanding of them. Additionally, in combination with competing metrics, calibration can help reframe trade-offs between fairness metrics as different opportunities for fairness. On the one hand, the disagreement between two metrics can help formalise the different opportunities they respectively offer, and to help reason on different choices regarding bias mitigation. On the other hand, it can highlight the multiple, irreconcilable costs involved for marginalised populations and thus, create an opportunity to reconsider some of the risks and benefits brought by the use of risk tools. These informative differences, captured by bias preserving metrics, could be used to inform the application of bias transforming metrics toward more equitable, rather than strictly equal, fairness interventions. Overall, these efforts can contribute to addressing the root causes of unfair predictions, rather than solely tackling their consequences.

In the following paragraphs, Section 3.2 will introduce the literature on fairness metrics. Specifically, it will challenge its primary focus on achieving 'fair' predictions. Section 3.3 will introduce bias as a 'thick concept' and Section 3.4 will introduce the distinction between bias preserving and bias transforming metrics. Section 3.5 will consider the implications of framing bias as a thick concept relative to the role that fairness metrics can play in general, as well as the role they can play in relation to each other; specifically, it will explore the ways in which fairness metrics could contribute to equity-driven fairness interventions in the areas of credit, health, and crime. Section 3.6 will offer

some policy recommendations for AI practitioners and policymakers working on algorithmic fairness, adding some important limitations to the argument. Section 3.7 will add some concluding remarks.

3.2 The Problem with 'Fair' Predictions

The efforts to address algorithmic bias can be traced in the literature that focuses more specifically on 'algorithmic fairness' (Berk et al., 2017; Calders & Verwer, 2010; Chouldechova, 2017; Corbett-Davies et al., 2017; Dwork et al., 2011; Feldman et al., 2015; Hajian & Domingo-Ferrer, 2013; Kamiran & Calders, 2012; Kleinberg, 2018).

As mentioned in the literature review (Chapter 2), the literature on algorithmic fairness is mostly engaged in the formalisation of fairness criteria from political philosophy, e.g., the Rawlsian notion of equality of opportunity (Lee et al., 2020) or from law, e.g., disparate treatment and disparate outcomes (Green, 2020; Wachter et al., 2021a) into formal metrics against which to evaluate predictive performance. This field has led to the creation of a host of so-called 'fairness metrics' against which predictive performance is to be assessed and adjusted to conform to the above ideas of fairness. Examples are 'statistical parity', 'group fairness', 'demographic parity', 'equalised odds', 'predictive parity', 'calibration' and, 'counterfactual fairness'. An exhaustive list can be found in 'Fairness and Machine Learning' (Barocas et al., 2022).

However different, the metrics respond to the general question guiding most of this field of research: '*How can we predict more fairly?*' As mentioned in the introduction above, these metrics are in contrast both with each other and with the demands of predictive accuracy (Berk et al., 2017). For instance, Kleinberg et al. (2016) proved that no three fairness metrics can be simultaneously satisfied but in the case of perfect classification or equal base rates; plainly, unless there are no errors in prediction or groups

behave the same. Hardt et al. (2016) confirmed this impossibility even for any two metrics. Chouldechova (2017) demonstrated how metrics on disparate impact are mathematically incompatible with those on disparate treatment.

On a more conceptual level, Green (2020) and Narayanan (2018) remarked how equality of outcomes and equality of treatment simply cannot be simultaneously satisfied in an unequal world. Finally, a survey by Berk et al. (2017) presented not only how these metrics and their variations conflict with one another, but also with the request of predictive accuracy maximization. This has also been confirmed in the work of Dwork et al. (2011), Corbett-Davies et al. (2017) and Kamiran & Calders (2012), who have resorted to operationalising any different conception of fairness as a constraint to predictive accuracy.

Overall, this apparent impossibility of finding a solution has been attributed to formal causes (Berk et al., 2017) or contrasting demands for fairness in society (Narayanan, 2018). Nevertheless, this impasse could also be telling of the contradiction represented by the attempt of the research community to *solve* bias without *solving for* bias first. Fairness metrics attempt to address bias by suggesting 'constraining' a biased performance according to different criteria. However, their use is not necessarily informed by what that specific bias might carry evidence of (e.g., historical disparities). Letting this evidence inform the solutions offered by fairness metrics will not resolve their incompatibility per se. However, it can help to make their incompatibility redundant by tailoring the application of specific fairness metrics to the different fairness concerns that bias can represent.

In this chapter, I take the stance of substantive fairness to explore ways in which fairness metrics can help surface existing disparities and thus inform efforts to address them. Green (2023) has drawn an important difference between formal algorithmic

fairness, where mathematical representations of fairness are applied to algorithmic decision-making, and substantive algorithmic fairness, where algorithms are incorporated into broader agendas for equity reform. Of course, fairness metrics alone cannot provide reliable evidence or proof of injustices. On that note, one might even argue that relying on base rates is sufficient or even preferable. Base rates provide a more direct and accurate picture of ground truth (Mayson, 2018). At the same time, evidence gathered from fairness metrics can work as a 'signal' that can be corroborated – and might not have otherwise so starkly emerged – through base rates. Additionally, by being a function of ground truth and predictions, fairness metrics can help capture discrepancies between expected and observed outcomes (e.g., FN and FP rates). This is key to diagnosing not only existing disparities but also the mechanisms through which these get magnified and the disparate impacts they might bring. For example, as will be discussed in the case of a health risk assessment tool later in this chapter, it is through the use of calibration that health disparities among populations with the same healthcare costs were signalled, and the role of algorithmic predictions in exacerbating these disparities was identified.

Research has shown how metrics in general can help get a fuller and more nuanced picture of a problem, especially when combined with qualitative accounts (Thomas & Uminsky, 2022). Fairness metrics can thus provide a signal or a piece of evidence to develop hypotheses on the presence of societal disparities and, in turn, support broader equity efforts. Abebe et al. (2020) show how computational tools in general (e.g., algorithms) can play a variety of roles in this sense. Given the exercise of formalisation that they require, for example, they can help articulate and reason on how social problems are formulated in algorithms and highlight the limitations presented by computational techniques. At the same time, given their use or reliance on historical data, they can help diagnose emerging injustices and let historical ones resurface.

Here, I will draw on some of these considerations to map out the roles that fairness metrics more specifically, rather than computational tools in general, can play to address the existing disparities that they capture. Specifically, I do so by showing that algorithmic bias is a problem as much as it is a source of evidence, and on the fact that some fairness metrics preserve while others transform historical biases captured in the data. I elaborate on this below.

3.3 Algorithmic Bias: a 'Thick' Concept

The double aspect of bias presented in the introduction, both as a problem and as a source of evidence, can be conceptualised by applying the theory of thick concepts. Thick concepts were introduced by Bernard Williams in his book 'Ethics and the Limits of Philosophy' in 1985 (Williams, 1985). They became the focus of meta-ethics[25] in the second half of the twentieth century, but their application has since spread to other philosophical fields such as epistemology, moral psychology, and philosophy of law (Väyrynen, 2021). Thick concepts are specific evaluative concepts – concepts that express a value or a judgment - that are also substantially descriptive. They are usually presented in contrast to so-called 'thin' concepts, which are merely evaluative. Paradigmatic examples of evaluative terms are 'good', 'bad', 'right' and 'wrong', while descriptive terms are terms like 'water', 'small' and 'solid'.

According to Williams (1985), a thick concept includes both aspects. As such, 'courage' could count as a thick concept as it denotes an ethical concept (evaluative) that is guided by the facts or acts of courage (descriptive). In this respect, Williams (1985) thought that thick concepts are both action-guiding, as they indicate a reason for action, and world-guided, as their interpretation depends on the state of the world or a specific context. For

[25] In philosophy and ethics, metaethics is the study of the nature, scope, and meaning of moral judgment.

example, the virtuous character attached to the concept of 'courage' can act as a motivation to engage in a courageous act. Concurrently, what constitutes a courageous act will change in its content depending on conditions such as historical time and place. An act of chivalry might be deemed courageous in the Middle Ages. However, it is likely to be considered overprotective in the present day. The relationship between the evaluative and the descriptive component of a thick concept should not be understood conjunctively ('and'), nor coextensively (that they entail each other). Burton (1992) and Tappolet (2004) suggest rather that a thick concept entails an evaluative component in virtue of some specific actualisation of its descriptive component. For instance, 'courageous' means virtuous in relation to some instance of behaving courageously.

3.3.1 Algorithmic Bias: The Case of Credit, Health, and Crime Risk Assessment

Several of the considerations above, I argue, can be intuitively applied to the concept of algorithmic bias considered here. Bias can be alternatively conceived of as a 'problem' (evaluative) and as 'evidence' of (descriptive) problems. In the realm of risk prediction algorithms, algorithmic bias can simultaneously be a problem and a source of evidence by reproducing existing societal disparities. Here, I illustrate this through three examples taken from the areas of credit, health, and crime in the US. Strategically, considering multiple cases allows researchers to systematically compare variations between them (Campbell, 1975). However, by representing single cases, the resulting considerations should not be used as grounds for generalisation (Tellis, 1997). These areas are, indeed, not the only three where algorithmic bias occurs[26]. However, they are here considered as an illustrative device. They represent specific areas where bias is simultaneously a problem in

[26] See Barocas et al. (2022) for a more comprehensive view of cases on bias and discrimination in machine learning.

its consequences, e.g., being denied a loan, and evidence of societal disparities, e.g., past redlining divisions.

Credit risk assessment refers to the process of determining the likelihood of default on a loan, using credit scores as a key metric. Credit scores can increase access to credit for previously credit-constrained populations and reduce human bias in credit decisions (Bartlett et al., 2017; The Federal Reserve Board, 2007). However, concerns have been raised that credit scoring systems might perpetuate existing disparities in credit allocations (Citron & Pasquale, 2014; Hurley & Adebayo, 2017). A study found that Vantagescore 3.0, one of the main credit scores in the US, is a statistically noisier indicator of default risk for minority and low-income applicants, resulting in disparities in credit outcomes or lower loan acceptance rates for low-income applicants (Blattner & Nelson, 2021). This noise was attributed to data sparsity (in the sense of having data from few loans or few years of history), past default history, and limited account diversity (in the sense of only holding multiple types of loans) in the scores of these applicants. These results, as the authors of the study confirmed, point to past disparities in credit access and persisting differential access to credit-enabled activities like homeownership for minority populations (Blattner & Nelson, 2021).

Health risk assessment algorithms are used to predict complex health conditions and develop interventions to address them. Despite their intended benefits, concerns have been raised around algorithmic bias in health risk assessment (Foryciarz et al., 2022; Obermeyer et al., 2019; Park et al., 2021), highlighting racial bias in widely used health risk assessment algorithms in the US such as OPTUM (Obermeyer et al., 2019). In this specific case, Black patients were found to be considerably sicker than White patients for the same predicted risk score. This bias resulted in unfair predictions, underestimating the health risk of people of colour compared to Whites. It has been suggested that these results

also evidenced unequal health insurance structures (Benjamin, 2019b). People of colour face higher barriers to healthcare access, even when holding health insurance. Access varies by factors such as geography and differential availability of transportation, competing demands from jobs or childcare, and knowledge of reasons to seek care (Adler & Newman, 2002; Adler et al., 2016; Fiscella et al., 2000). All these factors can bias predictive performance.

Risk assessment algorithms are also used in the criminal justice system to calculate the risk of recidivism (Monahan & Skeem, 2015). Concerns have been raised around bias in predictive performance, especially regarding racial disparities in pretrial risk assessment algorithms (Hannah-Moffat, 2015; Kleinberg et al., 2018; Simonite, 2019; Stevenson, 2017). As the COMPAS scandal showed (Angwin & Larson, 2016), these algorithms can result in disparities in predictions toward people of colour, resulting in their higher likelihood of being labelled as "high risk" in bail decisions compared to Whites. As these algorithms use data on past arrests to calculate such risk, these differences are likely to be influenced by disparate practices toward people of colour by the police, as Black Americans are more likely to be stopped, searched, and arrested by the police than White Americans (Epp et al., 2014; Gelman et al., 2007; Goel et al., 2018; B. Harcourt, 2015b).

The bias in these examples is 'action-guiding' in that, as a problem, it indicates a reason for (corrective) action (Williams, 1985). This is evident in the concerns and disparate consequences for minority populations that the examples above raised across the board. It is 'world-guided' in that, as evidence, it reflects the state of the world that the data captures. A biased performance that systematically prefers a dominant demographic is 'action-guiding' in terms of signaling a problem. Similar to a technical bug, it calls for a fix. However, it does not alone tell one how to solve it. It is 'world-guided' in terms of that bias being relevant as a source of evidence of demographic disparities in society. Such

information can inform potential strategies to tackle the problem, for example, anti-discriminatory policies or reparations for past injustices.

As suggested above, these two components of bias are related not conjunctively nor coextensively, but contextually so. Not every performance that is systematically skewed toward one demographic group needs to entail both the evaluative and the descriptive components. For instance, the bias could stem from the model rather than the data, indicating a flaw in the predictions rather than a reflection of demographic disparities. Moreover, algorithmic bias might not always be problematic; it can also reveal justifiable policy decisions. For instance, predictive performance could be skewed in favour of marginalised populations to rectify past injustices, such as by intentionally overestimating the risk of strokes among women[27]. However, bias becomes problematic when it aligns with and perpetuates underlying societal inequalities, such as those related to gender and race. In these instances, bias is more than just a technical flaw; it is a reflection of broader societal disparities. This is where the concept of algorithmic bias as a 'thick concept' comes into play – it not only describes a deviation or a technical flaw but also carries a normative dimension, prescribing the need for corrective action. When algorithmic bias mirrors existing demographic disparities, it becomes essential to recognise the 'normative' component, which highlights the ethical reasons for addressing bias as a problem. Concurrently, the 'descriptive' component offers valuable insights into the evidence that supports the existence of these disparities, thus reinforcing the need for intervention.

3.4 Fairness Metrics: To Preserve or to Transform?

As previously mentioned, fairness metrics are often proposed as a solution to bias in the context of risk assessment. Specifically, Wachter et al. (2021) categorise fairness metrics

[27] This is because the condition is more frequent among members of that demographic.

as either ‘bias preserving’ or ‘bias transforming’. This classification characterises fairness metrics concerning their treatment of demographic disparities embedded in the data; specifically, whether they aim to preserve or to transform them. In the next paragraphs, I take the example of demographic parity (bias transforming) and calibration (bias preserving) to illustrate how this classification works.

Machine learning aims to predict some abstract concepts (e.g., ‘creditworthiness’, ‘health’ or ‘crime’). As these concepts are hard to measure in practice, they need to be approximated through labels that act as proxies for them (Malik, 2020). A classic example is the use of arrest rates as a proxy for crime. Another example is the use of health insurance costs or expenses to measure health needs. These proxies function as substitute measures. They are drawn from available empirical data and thus, they are likely to contain traces of historical disparities between demographic groups. Depending on their definition, fairness metrics can either preserve or transform these disparities.

Bias preserving metrics require that the predictions of an algorithm match the ‘reality’ or the ground truth[28] expressed in the target labels data. These metrics assume that predictions are fair when they accurately match the data on which they are trained. Metrics such as calibration or methods to equalise error rates fall within this category. Bias transforming metrics are free from this requirement. They assume that the ground truth expressed in the training data is not necessarily a primary target to adjust predictions to. These metrics match decision rates between demographic groups so that they are ‘equal’ or ‘similar’, regardless of the ground truth. Metrics such as demographic parity and conditional demographic parity belong to this category. This tension is readily apparent in

[28] Ground truth refers to real-world data used to train and test AI model outputs.

their formal definitions, which are tied to the target label for bias preserving metrics but not for bias transforming metrics.

3.4.1 Demographic Parity and Calibration

Demographic parity can be expressed as:

$$P(\hat{y} = 1|A = a) = P(\hat{y} = 1|A = a')\ \forall a, a',$$

where:

- $\hat{y}$ represents the predicted outcome (e.g., a positive classification)
- A is the protected demographic attribute (e.g., race, gender)
- a and a' are any two values of the attribute A

This equation states that the probability of a positive prediction ($\hat{y} = 1$) should be the same for all demographic groups A regardless of the attribute. Taking the example of credit, this states that the predicted probability of successfully being granted a loan is the same regardless of which value of the protected group (e.g., 'woman' or 'man') one belongs to. Differently, calibration can be expressed as $P(\hat{y} = 1|y = 1, A = a) = P(\hat{y} = 1|y = 1, A = a')$ $\forall a, a'$. As suggested above, this bias preserving metric would require that the predicted probability of being granted a loan ($P(\hat{y} = 1)$) given that, as the training data suggests, *one actually got granted a loan* ($y = 1$) is the same regardless of the value of the protected attribute (woman or man)[29].

[29] Other bias preserving metrics such as equalised odds or equalising error rates are similar to the latter but are concerned with ensuring an equal True Positive Rates (TPR) and False Positive Rate (FPR) for the former and equalising False Positive Rate (FPR) and False Negative Rate (FNR) for the latter. For full list see Wachter et al. (2021).

Demographic parity is concerned with equalising acceptance rates or the probability of being accepted across protected groups. Calibration is concerned with ensuring that predicted risk scores carry the same evidentiary value or 'mean the same thing' across groups. In a recidivism example, demographic parity would require that Black and White people receive a score such as 7 at the same rate. Calibration would require that a *predicted* score of 7 over 10 expresses an *actual* 70% probability of being re-arrested for Black and White people alike.

3.5 From Fair Predictions to Interventions

Given the considerations in the previous paragraphs, I propose to shift the goal of fairness metrics from equalising results or rates toward supporting interventions to address fundamental problems of justice and equity. In this section, I consider the implications of framing bias as a thick concept for the role that fairness metrics can play in general, as well as the role they can play with each other; specifically, the role that bias preserving metrics can play vis-a-vis bias transforming metrics in tackling algorithmic bias. Below, I claim that we should preserve the evidence of societal disparities which is present in algorithmic bias, and, through this evidence, inform interventions to address these disparities. In concert, bias preserving metrics can inform the work of bias transforming ones toward equity-driven fairness outcomes. The next paragraphs will elaborate on these claims and help justify them.

3.5.1 Preserving Bias as a Source of Evidence

3.5.1.1 From reproducing problems to preserving evidence

Bias preserving metrics could be employed to measure, understand, and further investigate existing demographic disparities. Similar to what Abebe et al. (2020) proposed for general computational tools, they can help formalise and reason about the choices of specific

variables, datasets, and models. Together with additional data to corroborate the results (Thomas & Uminsky, 2022), they can also help measure disparities across groups and compare their respective situations.

In analysing bias in the OPTUM health risk assessment tool, for example, Obermeyer et al. (2019) relied on calibration as a metric. This required checking whether the predicted health risk scores matched the actual health outcomes across races. By that metric, the algorithm did not seem to convey a significant disparity in predictions across groups. Unconvinced, the researchers checked how health risk was calculated and they noticed that health insurance costs were being used as a proxy for one's health condition. While it is not fully incorrect to assume that greater health expenditures reflect greater health needs [30], this assumption might introduce a skew due to the presence of socioeconomic disparities. Even though easily retrievable from insurance claims data, health costs are the result of an aggregation process that features several distortions due to structural inequalities (e.g., inability to pay), inefficiencies (e.g., lack of access), and incentives (Obermeyer et al., 2019).

The researchers took data on patients' health conditions and retrained the algorithm to find that, at a given risk score, Black patients were found to be considerably sicker than White patients, as expressed by signs of previously uncontrolled health conditions. The researchers proceeded to compare calibration results for two new predictive models trained on the same input data but with two different target labels: health insurance costs and health conditions. Calibration results were used to measure and compare the disparities in health conditions for similar health expenses expressed at a given risk level. The fact that Black and White patients tended to generate similar costs for the same health risk scores,

[30] Many actuarial algorithms in the US, for example, are based on similar assumptions.

but Black patients showed worse health conditions, suggested a concerning difference between these two groups. This could be evidence of socioeconomic disparities, such as differences in ability to pay or access to healthcare, as well as unequal health insurance structures that negatively affect Black patients' ability to receive treatment (Benjamin, 2019b).

Calibration would not have been effective in fixing unfair results as it would have presented them as 'fair'. However, it allowed the researchers to measure and analyse the social issue they were investigating: disparities in healthcare. Calibration was used not only to measure the problem but also to gain a deeper understanding of it. By controlling for specific health conditions, they found that Black patients tend to generate more costs for emergency visits and dialysis, rather than routine costs (Obermeyer et al., 2019). When used to fix unfair predictions, calibration can preserve and reproduce bias. However, when consciously used to retain underlying data patterns – to 'preserve' evidence -, it can serve as a mirror to formalise and diagnose, allowing researchers to measure and address existing disparities.

3.5.1.2 From fairness trade-offs to opportunities for fairness

Concerning the relationship between fairness metrics, authors tend to frame their incompatibility in terms of trade-offs incurred in choosing one over another (Kleinberg et al., 2016; Mashiat et al., 2022). However, referring to algorithms' metrics in general, Thomas & Uminsky (2022) have shown that using a variety of metrics can provide a more complete understanding of a detected phenomenon (Thomas & Uminsky, 2022). Once we frame fairness metrics as tools in support of different interventions, the tension between these metrics can also inform their cooperation rather than necessitate their mutual exclusion.

This tension can be made explicit by comparing the trade-offs involved in the choice between calibration and equalising false positive (FP) and false negative (FN) rates. Let us take the example of crime risk assessment (Hao & Stray, 2019). On the one hand, in a well-calibrated tool, risk scores should mean the same for two groups. For example, a score of 8 over 10 should reflect the actual rate at which both Black and White people are considered 'high-risk' and thus, carry the same 'real' meaning for both groups. However, it has been shown that judging both groups by the same standard or by an equally meaningful score entails that a significantly higher percentage of people of color is wrongfully classified as high-risk (FP rate) and that a significantly higher percentage of White people is wrongfully classified as low risk (FN rate). Echoing the results of the impossibility theorem, the literature concludes that it is impossible to keep the error rates comparable between groups and treat people with the same risk scores in the same way (Hao & Stray, 2019). They use this to echo that two definitions of fairness might be both defensible and yet together unachievable.

Once calibration is satisfied, however, a resulting difference in error rates could carry evidence of historical injustices; for example, of past unfair treatment. Namely, a significantly higher percentage of people of color being wrongfully classified as high-risk given that every group is judged equally by the same risk score could be an indicator of past disparate arrest practices toward Black people. This is substantiated by evidence that Black people are arrested at higher rates in the US and that this biases the data (Angwin & Larson, 2016). Concurrently, once error rates are balanced, a violation of calibration could be an indicator of possible disparate treatment. To ensure that Black and White people are wrongfully classified at comparable rates, White people would have to be considered 'high-risk' with a lower threshold or risk score ('more easily') than Black people (Hao & Stray, 2019). This would result in disparate treatment toward White people. On the one

hand, this could be argued to be the cost to account for past (and present) inequalities in treatment suffered by racialised communities. At the same time, the need to use different risk thresholds to address this inequality might corroborate evidence of previous disparate arrest practices toward Black people.

These considerations can help re-cast the relationship between fairness metrics from trade-offs to opportunities for fairness. Fairness metrics can help formalise evidence to inform or justify a course of action over another. Suppose there is historical evidence that some protected group is generally disadvantaged in relation to another in an area such as crime risk assessment, and that evidence from fairness metrics suggests that the cost of disparate treatment that Whites would suffer under one metric is smaller than the cost of disparate treatment that Blacks would suffer (and have been suffering) under the other. In that case, this might provide one piece of evidence in support of the former course of action.

3.5.1.3 Evidence for refusal and reform

At times, however, the tension between metrics and the evidence that one might gain through their contrast can just express the limits of what can be done with computational tools (Abebe et al., 2020). Their disagreement can provide evidence against the choice to employ risk assessment tools and in support of non-computational changes.

The disagreement between the fairness metrics above might stand as proof that, no matter what metric is used, any way to assign risk estimates to express recidivism will necessarily result in disparities between protected groups. This provides evidence that no other choice of risk assessment tool can solve this. Thus, the disagreement can create an

opportunity to reconsider some of the risks and benefits brought by the use of risk tools.[31] In turn, this can inform the discourse around their use *vis a vis* the use of non-computational interventions. The extent of the disparities revealed by the disagreement between metrics can provide evidence to justify or back efforts to, in this context, reform the criminal justice system and the police. This has already been suggested by authors and activists such as Benjamin (2019a). As mentioned above, once risk scores have been calibrated, significantly higher error rates for Black people might not only evidence disparate arrest practices toward them, but also denounce how widespread and systemic their nature is. Chapters 5 and 7 will further delve into these considerations.

3.5.2 Transforming Bias as a Problem: Toward Equity-driven Fairness?

In concert, the evidence carried by bias preserving metrics can inform the work of bias transforming ones toward equity-driven fairness outcomes. Let us take the example of credit risk assessment to illustrate this. A bias transforming metric such as demographic parity would require that the predicted probability of being granted a loan is the same, and thus equal, across protected groups regardless of ground truth. Given evidence of historical credit disparities captured through a bias preserving metric[32], bias transforming metrics could be used to ensure that the predicted probability of being granted a loan is not strictly equal across groups but *equitable*: that it is responsive or informed by those historical disparities.[33] This is a way in which bias preserving and bias transforming metrics could collaborate.

---

[31] Of course, this should just be a piece of information in taking into consideration a wider range of factors about their risks and benefits. For example, the disagreement between fairness metrics is not in itself an argument against risk tools where these tools are found to decrease incarceration rates for Black people.

[32] As previously mentioned, this evidence should be corroborated through other data and methods too.

[33] Bias preserving metrics could not do that, as they are required to match their predictions to the ground truth, while bias transforming metrics are free from such a requirement.

Chander (2017) has shown how bias can be positively used to tailor a model around the specific needs of a vulnerable group or, in the guise of positive discrimination, how it can play a mitigating role against a model that favors a majority population over a minority one. On the one hand, the bias captured through bias preserving metrics could work as evidence to motivate fairness interventions. On the other hand, bias transforming metrics could be used to ensure that these fairness interventions respond to equity concerns. As a result, this could lead to more loans being granted to low-income applicants and, in return, create more correct data to reduce the present noise in the scores. In their study of bias in the credit score Vantagescore 3.0, Blattner & Nelson (2021) themselves suggested increasing credits to low-income applicants. The interplay between bias preserving and bias transforming metrics, I argue, could work toward equity-driven fairness.

Bias preserving metrics can together collect pieces of evidence of wider demographic disparities. However, it is important to remember that these disparities hide multiple forms or aspects of disadvantages which it is important to consider and disambiguate. As previously indicated, Blattner & Nelson (2021) conducted an investigation into the bias against low-income applicants within the credit scores they analysed. As previously mentioned, their findings revealed that approximately half of the observed disparity between low- and high-income applicants in credit scores could be ascribed to variations in the composition of credit reports. These variations encompassed factors such as sparsity, defined as having limited data from either few loans or a restricted history spanning few years, as well as past default history and account diversity, characterised by the possession of multiple types of loans (Blattner & Nelson, 2021). The analysis by Blattner & Nelson (2021) assigned the residual gap to credit files categorised as ‘clean’. Clean credit files were identified as expressing disparities in default reporting, as well as between-group distinctions in the intrinsic predictability of default.

Capturing these differences can be informative in deciding whether to target fairness interventions toward a model or toward society. One could be done by augmenting datasets to account for under- and over-representations that disadvantage some minority groups (Pastaltzidis et al., 2022). Given the example above, this could be done to counter aspects that affect disadvantaged populations, such as data sparsity. The other can be done, as suggested above, by incentivising credit for previously credit-constrained populations in order to break these cycles of disadvantage in society.

3.6 Policy Implications and Recommendations

Given the considerations above, this section develops the implications for policy in the form of practical recommendations and caveats. Especially, it places the theoretical propositions above in front of some practical limitations. These policy recommendations are addressed to an interdisciplinary audience of developers, researchers, and policymakers working at the intersection of Fairness, Accountability, and Transparency (FAacT) in AI, especially on algorithmic fairness.

1. *Evidence captured through bias preserving metrics should be corroborated through qualitative and participatory methods.*

On the one hand, bias preserving metrics cannot alone be considered a reliable source of evidence for societal disparities. Any of their results should be validated through other methods. For example, model cards (M. Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2021) can be used to support or falsify the evidence of bias discovered through these metrics. Both these approaches propose documenting much richer and more comprehensive details about a given model or dataset, including more qualitative aspects, such as intended use cases, ethical considerations, underlying assumptions, caveats and more.

On the other hand, bias transforming metrics alone cannot 'solve' societal disparities. It is important to involve a wide range of stakeholders, especially those demographics that are impacted by the bias of these algorithms, in devising any interventions to tackle the problems that affect them (Thomas & Uminsky, 2022). It is also essential to consider non-computational interventions, such as social welfare policies. Future research could further explore how these qualitative accounts can support as well as favor from evidence of bias.

2. *Bias preserving metrics should be used to collect evidence of - rather than prescribe – the differential treatment of groups.*

Utilising fairness metrics for equity interventions may encounter legal challenges. U.S. law exhibits reluctance towards the explicit consideration of race or other protected characteristics in shaping policies, even when motivated by a desire to redress past discrimination. In 2023 for example, the U.S. Supreme Court ruled that race-conscious University admissions violate the equal protection clause in the U.S. Constitution (K. Meyer, 2023). Advocates supporting the decision contended that considering race to 'positively' bias the admissions process for some racial minority groups would constitute racial discrimination against groups such as Asian and White Americans (Rios, 2023). Therefore, the proposition to employ bias preserving metrics to advocate for a more equitable treatment of a protected group risks being construed as discrimination.

However, it is crucial to note that scholars such as Hellman (2019) observe that the mere practice of collecting information on racial categories does not inherently constitute discrimination based on race. According to this interpretation, utilising bias preserving metrics to gather evidence of racial discrimination is likely not to raise legal concerns. For example, the U.S. Census, mandated by the U.S. Constitution, collects information about the U.S. population, including factors such as race (US Census Bureau, n.d.). As Hellman

asserts, the collection of information differs from its utilisation, with the former not constituting disparate treatment and thus not triggering concerns about discrimination.

As previously mentioned, by respecting these considerations, bias preserving metrics such as calibration could then serve the multiple roles outlined above. As diagnostic tools, they could help measure societal disparities and gain a deeper understanding of them. Additionally, when employed alongside competing metrics, calibration could assist in reframing the trade-offs between fairness metrics as different opportunities for fairness. On one hand, the discrepancy between two metrics can formalise the distinct opportunities they respectively offer and aid in deliberating on various choices regarding bias mitigation. On the other hand, it can serve as evidence of the multiple, irreconcilable costs involved for marginalised populations, thereby serving as an argument against the use of risk tools.

3. *Bias transforming metrics could be used to monitor and reason on – rather than enact – differential outcomes along task-relevant features.*

While the mere collection of information on protected characteristics can be legal, the use of this information to enact equitable outcomes is likely to be precluded by US law. Rather than prohibiting any possible use of this information, however, the law is concerned with the manner in which it is used (Hellman, 2019). For example, the direct use of race to employ different risk thresholds for different racial categories – such as considering White people as 'high-risk' by a lower threshold as mentioned above (Hao & Stray, 2019) - is likely to raise legal concerns. Using race information to weigh differently the same predictors by which risk is calculated across racial groups would count as disparate treatment (Hellman, 2019).

However, the situation is unclear where the use of race information is used to find which features are relevant for predicting the outcome for different groups, where 'relevant' is defined as improving the predictive power of the algorithm for the task at hand (Hellman, 2019). Once we know that some features are more predictive of an outcome for some protected groups than others, race information could be used to determine how to differentially employ these features in making predictions for different racial groups. Corbett-Davies & Goel (2018), for example, advise applying different group tracks in the algorithm and using different relevant information to evaluate each track to improve predictive performance while increasing fairness. Differently from judging racial groups differently by the same predictors, this option would use differently relevant predictors in order to judge the groups 'equally'[34] and thus, might not raise the same legal concerns. It is also important that this information is not used to determine a judgement, but rather as an additional 'plus' factor to consider while making a decision.

While this solution would not allow for the direct enactment of an equitable skew in performance, it would not prohibit the differential use of predictors in equitably influencing it. Overall, this can indirectly contribute to reducing existing disparities. Under this reading, bias transforming metrics could help reason on, rather than enact, potential improvements in the equality of outcomes after such interventions. For example, they could help check whether the probability of being accepted for a loan across protected groups is more equal after these more indirect equity interventions. The legality of such a proposition, however, would be highly dependent on the context and character of the law and would need to be further tested through future research.

[34] For the different racial groups to obtain the same or a similar predictive performance.

3.7 Conclusion

Overall, this chapter aims to narrow the gap between technical and policy solutions around fairness and equity. It has achieved this by providing a blueprint to repurpose fairness metrics as tools in support of fairness interventions; specifically, as tools tailored toward specific aspects of societal disparities which are mirrored in algorithmic bias. In the last instance, it also advocated for the possibility of fairness metrics to be used in concert to shape equity-driven fairness interventions. In practice, however, the correct application of a type of fairness metric is not as simple, nor is their application to law and policy. Future research could investigate putting into practice the proposed uses of fairness metrics in different contexts such as those of credit, health and crime cited here. I hope that this chapter – and especially the recommendations above - can help researchers and practitioners with those intentions to strike the right balance between technical feasibility and legal concerns in practice.

Chapter 4:

How can bias as evidence be measured?

Chapter 4. A Genealogical Approach to Algorithmic Bias

Abstract

The Fairness, Accountability, and Transparency (FAccT) literature tends to focus on bias as a problem that requires *ex post* solutions (e.g., fairness metrics), rather than addressing the underlying social and technical conditions that (re)produce it. In this chapter, I propose a complementary strategy that uses genealogy as a constructive, epistemic critique to approach algorithmic bias in terms of the conditions that enable it. I focus on XAI feature attributions (Shapley values) and counterfactual approaches as potential tools to gauge these conditions and offer two main contributions. One is constructive: I develop a theoretical framework to classify these approaches according to their relevance for bias as evidence of societal disparities. I draw on Pearl's ladder of causation (2000, 2009) to order these XAI approaches concerning their ability to answer fairness-relevant questions and identify fairness-relevant solutions. The other contribution is critical: I evaluate these approaches in terms of their assumptions about the role of protected characteristics in discriminatory outcomes. I achieve this by building on Kohler-Hausmann's (2019) constructivist theory of discrimination. I derive three recommendations for XAI practitioners to develop and AI policymakers to regulate tools that address algorithmic bias in its conditions and hence mitigate its future occurrence.

Note: This chapter has been published as a peer-reviewed article in the Springer Journal Minds & Machines (see (Ziosi et al., 2024)). The chapter has undergone minimal editing in order to ensure coherence with the thesis as a whole. The article was co-authored with Dr. David Watson and Prof. Luciano Floridi. Please see Appendix A and B for authorship statements.

4.1 Introduction

Risk prediction tools can increase decision efficiency in contexts such as credit, health, and criminal justice. They may bring more neutrality, countering subjective and prejudice-driven human judgment, and improve accuracy, resulting in more efficient and resource-effective decision policies (Barabas et al., 2018). However, for years, algorithmic tools have been criticised for reflecting and potentially exacerbating pre-existing biases (Barocas & Selbst, 2016; Citron & Pasquale, 2014). As mentioned in section 1.1, algorithmic bias in this context is generally taken to refer to cases in which

> *'predictive performance (however defined) unjustifiably differs across disadvantaged groups along social axes such as race, gender, and class'* (Mitchell et al., 2021, p.1).

This bias is also referred to as a 'skewed performance' along one of these demographic axes.

In the Fairness, Accountability, and Transparency (FAccT) literature, bias tends to be characterised as a *problem* in its consequences, that is, an issue requiring an *ex post* solution. As Chapter 3 has presented, an example of this type of solution is fairness metrics, a set of measures that enable one to detect and adjust bias in a model. Algorithmic bias is rarely considered as *evidence* of the underlying social and technical conditions that (re)produce it – that is, as an issue requiring an *ex ante* solution. As mentioned in the literature review (Chapter 2), this tendency promotes the design of solutions *ex post*, by addressing the consequences, rather than (at least also) *ex ante*, by addressing the conditions. In this chapter, I seek to rebalance the overall strategy. I analyse a set of explainable artificial intelligence (XAI) approaches with respect to their ability to gather evidence – note, not proof – of societal disparities. I focus specifically on feature attribution approaches that rely on Shapley values and counterfactual approaches. These enable one to examine the relationship between protected characteristics such as race or gender and skewed performance.

Although the relation between explainability and fairness is key to approaching algorithmic bias as 'evidence,' it remains analytically vague. Additionally, some of the bias-relevant applications of feature attribution approaches tend to represent the role of protected characteristics in discriminatory outcomes unrealistically – e.g., as independent, intrinsic, and causal attributes. A complementary strategy is to approach bias *genealogically*. In this chapter, I use genealogy as a constructive, epistemic critique,[35] with a double role. Constructively, it allows us to explain algorithmic bias in terms of the conditions that give rise to it, *ex ante*. Critically, it helps explain algorithmic bias not in terms of a single origin ('cause'), but with respect to a broader set of social and technical conditions at play that (re)produce these disparities.

In this respect, I make two main contributions. I offer a theoretical framework to classify XAI approaches according to their relevance to gather evidence of societal disparities. I take inspiration from Pearl's ladder of causation (2000, 2009) to characterise XAI approaches into observational, interventional and counterfactual approaches – namely, concerning their ability to detect a) whether – and, if so, b) how – a protected characteristic contributed to skewed performance, and c) what can be done to change it. The goal is to consider these XAI methods not only as technical tools but as means to investigate and collect evidence about unfair differences in performance alongside protected characteristics. The second is to critique these methods concerning their ability to represent the role of protected characteristics in discriminatory outcomes. Drawing from Kohler-Hausmann's (2019) constructivist theory of discrimination, I question observational, interventional, and counterfactual XAI approaches concerning the

[35] In philosophy, following the tradition of Nietzsche and Foucault, a genealogy is a form of historical critique, designed to overturn norms by revealing their origins (Hill, 2016). I use the term more in its philological sense, to mean a constructive critique that looks at the conditions of possibility for a problem to address it successfully.

independence, *responsibility,* and *epistemic* assumptions they make towards protected characteristics, respectively. The aim is to question XAI approaches in their ability to help capture salient aspects of discrimination.

The remainder of this chapter is structured as follows. In Section 4.2, I review the relationship between explainability and fairness. In Section 4.3, I present the genealogical approach to bias. In Section 4.3.1, I characterise XAI approaches concerning their relevance for fairness. In Section 4.3.2, I question their capacity to address algorithmic discrimination. Finally, in section 4.4, I derive three main recommendations for XAI practitioners to develop and for policymakers to regulate tools that address algorithmic bias in its conditions and thus mitigate its future occurrence. Section 4.5 provides some concluding remarks and outlines directions for future research.

4.2 Explainability and Fairness

As the literature review (Chapter 2) has suggested, explainability can enhance fairness-relevant properties to different extents and on different levels. Explainability can enhance transparency (Abdollahi & Nasraoui, 2018), granted by the ability to see how a model has arrived at a discriminatory outcome or whether the model itself is putting undue weight on certain characteristics. Additionally, explainability can increase or enable trust in a model (Dodge et al., 2019). It can help, for example, to determine if qualities relevant to algorithmic fairness (such as fairness metrics) are met (Doshi-Velez & Kim, 2017). Explainability can also enhance accountability, as it can provide explanations for AI-informed (un)fair decisions (Leben, 2023; Zhou et al., 2022). These properties concern fairness within the context of 'responsible AI', AI that takes into account moral, and ethical considerations as well as social values (Adadi & Berrada, 2018).

At the same time, the relationship between explainability and fairness is not always positive. Explainability, for instance, can influence the *perception* of fairness. On that note,

some highlight the risks for fairness that more explainability, and more reliance on it, can bring. Examples are the risks of 'fairwashing' and of the rationalisation, and potential justification, of some types of discrimination (Aivodji et al., 2019). Ananny & Crawford (2018) have highlighted how the ability to 'see' a model does not equate to the ability to 'govern' it nor 'understand' it and, thus, to mitigate bias. Additionally, authors such as Barocas (2022) have recently highlighted the tensions between calls for simpler models to ensure transparency (and thereby facilitate algorithmic fairness), and the inconvenient fact that such models may be less able to satisfy some fairness demands (e.g., allowing for specific parameter tweaks). In a scenario involving a highly skewed dataset, for example, the transparency typically afforded by simpler models might aid in identifying sources of bias. However, these models might also offer fewer opportunities to adjust parameters in ways that could mitigate this bias. This shows how the relation between fairness and explainability need not be positive and may require trade-offs.

Notwithstanding the ability of XAI methods to enhance or mitigate the potentially negative consequences of algorithmic bias, my focus here is different. Specifically, I am interested in the potential of XAI's methods to gather evidence of the conditions that enable it. I see algorithmic bias as the *object,* and XAI approaches as a *means* to uncover and learn about the underlying conditions of social inequality.

Within the realm of explainability approaches, I focus specifically on feature attribution (e.g. SHAP (Lundberg & Lee, 2017)) and counterfactual approaches (e.g., Galhotra et al., 2021; Karimi et al., 2020, 2021). Feature attribution approaches can estimate how input features contribute to performance biases (Begley et al., 2020). Among

feature attribution approaches, this chapter focuses specifically on Shapley values.[36] These help identify variables that are drivers of unfair outcomes by allocating responsibility for observed disparities, defined through a specific measure of fairness, among the considered input variables of the model (Lundberg, 2018). This is arguably the most popular feature attribution method, as it unifies several related methods and comes with axiomatic guarantees (Lundberg & Lee, 2017). However, various methods exist for computing Shapley values that may provide different attributions for the same prediction (Sundararajan & Najmi, 2020). Following Heskes et al., (2020), I provide a division into marginal, conditional, and interventional or causal approaches. I also examine counterfactual approaches. These explain what could have happened to an outcome had an input feature to a model been changed in a particular way (Barocas et al., 2020; S. Verma et al., 2020). Together with the latter, as I will present, these approaches can be easily interpreted through a causal ladder framework.

4.3 A Genealogical Approach to Bias

The above suggests that XAI methods can be relevant to explaining an algorithm's performance concerning estimating the contribution of protected characteristics. When this performance reveals disparities about gender or race, the approaches I consider can support the goal of explaining this performance in terms of the input variables that conditioned it. However, I need an overarching strategy for how to apply the XAI methods I consider

[36] Shapley values were adapted from Shapley's (1951) foundational work in cooperative game theory, where the original goal was to quantify the contribution of individuals to a given coalition. In the XAI context, Shapley values represent the marginal contribution of a feature to the model's output when averaged over a specific distribution of all possible feature coalitions. They are the unique solution to the attribution problem satisfying certain desirable properties – e.g., linearity, symmetry, and efficiency. Concerning fairness, they can help identify variables that are drivers of unfair outcomes. They do so by allocating responsibility for observed disparities, defined through a specific measure of fairness, among the considered input variables of the model (Lundberg, 2018).

towards that goal. In this chapter, I propose to adopt a genealogical approach to algorithmic bias.

Genealogy refers to a form of historical critique, designed to overturn social norms by revealing their origins (Hill, 2016). Here, I use the term in its philological sense, to mean a constructive critique that looks at the conditions of possibility of a problem to address it successfully. In this case, it refers to a constructive critique designed to understand algorithmic bias by focusing on the plural, dynamic, and contingent conditions for its possibility, and the potential of XAI methods to surface evidence of them. Specifically, a genealogical approach to algorithmic bias invites one to explain bias in terms of the conditions for its occurrence and understand its explanation not in terms of a single cause, but of gathering evidence on the set of conditions that produce it.

As cited above, past redlining divisions, differential access to healthcare, and disparate arrest practices towards people of colour represent some examples that, through the use of proxies such as zip-codes, represent what I mean by these 'conditions'. Historical segregation in US neighbourhoods, for instance, profoundly affected the residents' access to credit, health insurance, and education (Agyeman, 2021; Perrino, 2020). In turn, this created the conditions of poverty, unemployment, and past default history by which residents in these communities are considered 'not worthy' of credit when zip codes are used to calculate the risk of default. I aim to express this genealogical approach by adopting both a constructive and a critical stance, taking inspiration from Pearl's ladder of causation (2000, 2009) and Kohler-Hausmann's (2019) constructivist theory of discrimination, whose contribution will become more evident in the following sections.

4.3.1 XAI Approaches as Questions

By taking inspiration from Pearl's ladder of causation (2000, 2009), I provide an ordering principle for XAI approaches to clearly distinguish between their utility for fairness along three levels – specifically, their differential ability to 'see', 'govern,' and 'understand' what influences skewed performance. Overall, this can help clarify the vague relationship between explainability and fairness. In the specific case of the feature attribution and counterfactual approaches that this chapter focuses on, this can help answer the following questions: (1) Is a protected characteristic unfairly associated with outcomes? (2) Would intervening to alter a protected characteristic directly affect outcomes? (3) Given observed values for protected characteristics and outcomes, would a hypothetical intervention to alter a subject's protected characteristic have changed the outcome? I propose so-called 'observational' approaches as relevant for procedural fairness, 'interventional' approaches for consequential recommendations, and 'counterfactual' approaches for algorithmic recourse.

4.3.1.1 What Bias: Observational Approaches for Procedural Fairness

I refer to marginal and conditional feature attribution methods as 'observational approaches', as they can help *observe* whether a protected characteristic is unfairly associated with an outcome. Marginal variable importance measures estimate the importance of features, assuming that these are independent of each other. An example is given by Datta et al. (2016)'s Quantitative Input Influence (QII) method, where Shapley values are used to calculate the average marginal influence of input features. Another example is provided by Štrumbelj & Kononenko's (2014) work. They use marginal Shapley values to develop a sensitivity analysis-based method to estimate individual feature contributions. Even though common, this assumption of independence might lead to incorrect or counterintuitive explanations when the features are, in fact, highly

correlated. Additionally, it allows these methods to represent only the direct effects of variables.

This motivates the use of conditional variable importance measures. These can represent indirect effects, and estimate importance by conditioning on a variable. Aas et al. (2021) propose conditioning strategies to compute more accurate Shapley values. Another example is Frye et al.'s (2020) so-called asymmetric Shapley values. They are called 'asymmetric' because, when computing Shapley values, they restrict the possible permutations of the features to those consistent with a partial causal ordering. They then apply conditioning by observation to check that their explanations respect the multivariate distribution of the data. Thus, they check that they do not produce misleading or nonsensical explanations because of feature dependence.

Overall, these observational approaches make it possible to check whether a protected characteristic, such as race or gender, is unfairly associated with an outcome. This is relevant for ensuring procedural fairness, i.e., the fairness of the decision-making process (Grgić-Hlača et al., 2018). In the example of credit risk assessment, it would allow a regulator or an auditor to see whether, for example, gender or race was (illegally) considered in arriving at an assessment for a loan. In this respect, 'unfairness' would amount to direct discrimination and be illegal (Prince & Schwarcz, 2019). Procedural fairness is related to commitments to values such as accountability and transparency (Rueda et al., 2022). Additionally, it allows the fulfillment of requests such as compliance in finance or due process in law.

These considerations suggest that these measures could be relevant for fairness in an algorithmic context for tasks such as auditing (Watson, 2022); specifically, ethics-based auditing (EBA) (Mökander et al., 2021a; Mokander & Floridi, 2021). EBA refers to

> *'a structured process whereby an entity's present or past behaviour is assessed for consistency with relevant principles or norms'* (Mökander et al., 2021, p.2).

According to Mökander et al. (2021), EBA can

> *'contribute to good governance by promoting procedural regularity and transparency'* (p.16).

The feature importance approaches mentioned here can contribute to assessing whether a protected characteristic played a role in the decision-making process.

In this respect, while both marginal and conditional approaches are concerned with answering the question above, they might be differentially relevant for procedural fairness. Authors like Watson (2022) suggest that the former can provide insights into model mechanics, while the latter is more informative about the underlying data-generating process. Accordingly, marginal measures could help shed light on discrimination at the level of the model. This could be relevant to EBA when reviewing source code audit (Mökander et al., 2021a). Conditional measures could help shed light on discrimination at the system level. In this respect, a model could play an instrumental role, and conditional measures could be useful to check for procedural fairness in more complex decisional settings such as institutions for credit, justice, etc.

4.3.1.2 How Bias: Interventional Approaches for Consequential Recommendations

Interventional approaches can help investigate whether intervening to alter a protected characteristic directly affects outcomes. Beyond showing whether a feature is correlated with a discriminatory outcome, these approaches allow us to tell *how* (i.e., the extent to which) an intentional change in one feature changes the outcome. So-called interventional or causal Shapley values can help answer this question, as they are designed to capture causal contributions (Heskes et al., 2020). They do so by quantifying the effects of each

input on a model's output in accordance with a user-supplied causal graph. Examples include do-Shapley values (Jung et al., 2022) and Shapley flow (Wang et al., 2021).

Unlike marginal observational approaches, most of these approaches consider the relations between input features (Heskes et al., 2020). They do not assume independence between them. They tend to do so by relying on a causal representation of the model or the 'world' through a causal diagram. This entails making explicit some assumptions about the features considered in a model. Such a representation can help understand the interaction of the input features within a model or a system, and to reason about potential interventions (Wang et al., 2021). When modelled as a causal driver of both re-arrests and anti-law enforcement resentment ('antisocial cognition'), for example, intensive policing could be reasoned about as a site to intervene on (e.g., to be reduced) to mitigate these outcomes.

Regarding fairness, this can potentially respect and enhance one's agency. The ability for someone to exercise their agency is close to values such as self-determination and autonomy (Christman, 2020). Interventional approaches could do so in the form of consequential recommendations. Given a negative outcome, consequential recommendations provide the minimum intervention required to obtain a better result (Karimi et al., 2020). For example, they might suggest how much a credit applicant needs to increase their credit score or income to raise their chances of receiving a loan. However, since most protected characteristics cannot be changed, these recommendations suggest interventions on so-called 'intervenable' factors, such as income in credit risk assessment, employment in crime risk assessment or nutrition in health risk assessment.

4.3.1.3 Why Bias: Counterfactual Approaches for Algorithmic Recourse

Given observed values for protected characteristics and outcomes, these approaches can suggest whether a hypothetical intervention to alter a subject's protected characteristic would have changed the outcome. Examples are provided by Galhotra et al. (2021) and Karimi et al. (2020, 2021). Most notably, Galhotra et al. (2021) propose 'probabilistic contrastive counterfactuals' which not only help quantify the direct and indirect effects of a feature on outcomes but also provide actionable recourse to individuals negatively affected by such an outcome.

These methods can allow users to check whether a protected characteristic (e.g., race or gender) was the cause of a specific discriminatory outcome, and what can be done to change that outcome. Beyond interventional approaches, counterfactual ones allow one to know not only how to act, but also to understand what brought about a discriminatory outcome. Additionally, interventional approaches base their recommendations on the *consequences* they can bring about. By contrast, counterfactual approaches can base recommendations on what *caused* a discriminatory outcome. Typically, one is concerned with improving outcomes, the other with reversing unfavourable outcomes (Karimi et al., 2021). In that respect, their relevance to fairness is still related to agency, but more closely to 'recourse'.[37]

In law, recourse refers to actions that individuals or corporations undertake to remedy unfair or unfavourable legal outcomes (Wallin, 1992). Algorithmic recourse refers to

[37] Fairness metrics such as demographic parity, equalised odds or disparate impact are generally concerned with the distribution of outcomes (e.g., whether different groups are treated similarly or 'equally' by the algorithm). Differently from the counterfactual methods presented here, they do not directly address why or how those outcomes occur.

> *'the systematic process of reversing unfavourable decisions by algorithms and bureaucracies across a range of counterfactual scenarios'* (Venkatasubramanian & Alfano, 2020, p.284).

For example, a rejected loan applicant who is a woman can argue for recourse if there exists a positive counterfactual instance to hers; an applicant who is similar or 'close' to her in every other feature but for gender, and who was granted a loan. Karimi et al. (2021) consider that algorithmic recourse is met when a candidate is provided both with an explanation of why the loan was rejected and offered recommendation(s) on how to obtain the loan in the future (Karimi et al., 2021). Algorithmic recourse, they claim, is achieved when one

> *'can understand and accordingly act to alleviate an unfavourable situation'* (p.2) (Karimi et al., 2021).

Given that these explanations are formed by looking at an opposite outcome in a unit that is the same or similar *but for* a protected feature, these explanations are often referred to as contrastive explanations (Galhotra et al., 2021; Karimi et al., 2021). These formulate explanations in terms of explaining why this outcome rather than another ('the opposite') happened.

4.3.2 Questioning XAI Approaches

What does it mean, however, for a protected feature such as gender or race to 'be associated with', 'alter' or 'cause' a discriminatory outcome? This section takes inspiration from Kohler-Hausmann's (2019) constructivist theory of discrimination to question how these XAI methods approach it concerning protected features.

4.3.2.1 From Procedural Explanations to Evidential Observations

From a statistical standpoint, the main objection to using observational approaches for procedural fairness arises from their assumption of feature independence. Marginal Shapley values ignore the fact that a change in one input feature may cause a change in

another. If protected characteristics present spurious correlations with the discriminatory outcome via another, possibly unobserved, feature this will produce misleading explanations (Heskes et al., 2020; Nabi & Shpitser, 2018). Conditional Shapley values recognise the presence of other features, and how this can influence the contribution of other features under consideration. However, they do not usually rely on a causal representation (e.g., a causal graph) of the relationship between characteristics. Similarly to some interventional approaches, they might not distinguish intermediate outcomes from covariates (Greiner, 2008) and produce unreliable explanations.

From a critical standpoint, one should ask what it means for a protected characteristic such as race or gender to be independent of other input features. I claim that the problem with assuming independence is that it assumes that protected features can be represented as discrete units, existing in isolation rather than in relation to a host of other variables or features.

The independence assumption comes with important conceptual repercussions. In reality, one cannot separate being a woman or being a person of colour from one's socioeconomic circumstances, at least in societies where gender or racial inequalities are present (Hu, 2023). Even when this dependence is acknowledged by conditional approaches, e.g., by 'controlling for' or 'conditioning on' gender, it should be understood in relation to the entire system of other features within which it is embedded. To attribute a role to gender in a credit decision without considering how variables such as income, marital status, and education relate to each other is to misunderstand or even deny the very role that gender plays in credit settings. One is not classified as less likely to pay back a loan just because they are a woman, but because of how that 'comes with', and thus influences, a host of other input features, e.g., lower income because of the gender pay gap.

These approaches are thus more appropriate to *observe* whether gender played a role, rather than to explain *which* role gender played vis-à-vis other features. While this aligns with the aims of procedural fairness, these considerations should serve as a caveat to avoid using these approaches beyond their means. It is essential to avoid using the contributions they estimate as full explanations, but rather as mere evidence of a potential link with gender, which needs to be further tested with, for example, causal inference methods or by engaging with the individuals or communities affected. I thus suggest referring to explanations produced through these approaches as 'evidential observations', explanations that may trigger further investigation into the conditions that brought about a discriminatory outcome.

4.3.2.2 From Consequential to Constructive Recommendations

One objection that can be brought to consequential recommendations obtained through interventional approaches is precisely that they do not rely on causes, but on whatever brings about the best consequences. As a result, the actions they offer can be either inconsistent or have little to do with discrimination. For example, recommendations tailored towards making rejected applicants more likely to be accepted for a loan might require them to change their race. As this is impossible (and wrong), other input features on which an individual can more plausibly intervene are often used instead. In this case, recommendations can provide a relatively sensible suggestion, such as changing one's job. However, if one's rejection was, in fact, a result of (racial) discrimination, this does little to resolve the original injustice.

One ought to ask, what is problematic about 'intervening on' a protected characteristic such as race or gender? This request, I argue, entails an assumption about responsibility. Namely, it ascribes to the person with the protected feature the responsibility to change their situation. However, protected features are traits *of* the unit.

Not only can one not practically intervene on or change their race, but also, normatively, one should not be held responsible for it. This is specifically relevant considering the previous claim; that interventional approaches can be useful for fairness with regard to enhancing one's agency. When developing consequential recommendations, the responsibility for changing an outcome must rest on the people who designed and deployed the algorithm (e.g., developers or providers), not the end users subjected to their results.

As such, any consequential recommendation that focuses on the effect of protected characteristics on a discriminatory outcome should not simply rely on another feature that can be 'intervened on'. It should instead provide a reason for the developers to reflect on how its model represents these protected characteristics. Given that interventional and causal Shapley values usually rely on a causal representation of the model or the 'world' through a causal graph, these approaches can promote fairness when developers use them to understand how protected features such as race or gender are represented in their model alongside a set of other features (Hu, 2019). This is especially relevant given that causal graphs encode direct and indirect associations. Thus, they allow one to reason about positive and negative contributions to skewed performance.

As Kohler-Hausmann (2019) suggests, protected features do not have causal effects so much as structural properties: they are *embedded* within structures, whether social systems or algorithmic models, influencing and constructing their meaning and role. As such, these approaches could be used to provide constructive recommendations to developers on how to change the way gender or race is represented in the model vis-à-vis other input features. For example, this could be done by realising that some variables seemingly unrelated to race, e.g., zip codes, are proxies for it in their model and how they could reconfigure this relation to change it. Rather than bringing about desired

consequences for and from users, these approaches can be better framed as constructive recommendations from developers toward users.

4.3.2.3 From Contrastive to Constitutive Explanations

Regarding counterfactual approaches, one could contend that, even though theoretically sound, they are often too demanding to realise in practice (S. Verma et al., 2020). Computing counterfactuals generally requires not just a causal graph but also knowledge of the structural equations that govern the relationships between nodes. As I noted previously, some XAI approaches that rely on this intuition overcome these limits by framing the search for the counterfactual as an optimisation problem. The counterfactual is found by characterising a notion of distance that allows us to identify the nearest hypothetical point, which is classified differently from the one considered (Wachter et al., 2017).

While this practical solution reduces the epistemic demands of counterfactuals, one should ask what it means for gender or race to 'cause' a discriminatory outcome. The solution provided above relies on the assumption that the two 'similar' or 'nearest' units are the same *but for* the protected features. For a protected feature to be causal, as I suggested before, would mean that it is the only factor by which these two similar units differ and that, given that these lead to two different outcomes, the protected feature must be the cause. However, as Kohler-Hausmann (2019) suggests, a counterfactual unit is not the same *but for* the protected feature. It is a different unit precisely *because of* the protected feature. The epistemic assumption that contrastive explanations make is that gender or race are 'causes' rather than that they constitute or characterise different units or 'worlds'.

In this respect, one should talk about constitutive rather than counterfactual explanations. To say that a given system has a causal effect because of how it is constituted is to suggest that if one changed parts of that system, it would have a different causal effect. However, by the logic above, it would also be a different system. In this sense, a constitutive explanation attempts to explain different outcomes by pointing not to the 'cause' but to the parts that constitute these different units. With relevance to fairness, explaining different outcomes concerning these parts and their organisation can help shed light on the conditions by which similar individuals are treated differently. For example, given two similar individuals with different credit outcomes, a constitutive explanation would entail naming the features by which these individuals are considered similar (e.g., income, zip codes) precisely as what makes their (different) treatment unfair. On the one hand, being treated differently based on the same relevant factors to predict default risk and yet, being of a different gender or race points towards potential discrimination. On the other hand, using historically relevant predictors of default such as income or zip codes as a basis for this comparison, can itself add a discriminatory layer. Going back to the aim of this chapter, it would help approach bias *ex ante* by focusing on the conditions that constitute it, rather than only *ex post* by focusing on fixing the consequences.

4.4 Recommendations for AI policy and governance

This chapter argues that the Fairness, Accountability, and Transparency (FAccT) literature tends to focus on bias as a problem that requires *ex post* solutions, e.g., fairness metrics, rather than also addressing the underlying social and technical conditions that (re)produce it. It proposes a complementary strategy that uses genealogy as a constructive, epistemic critique to approach algorithmic bias in terms of the conditions for its possibility. In this respect, the chapter has focused on XAI feature attribution approaches (Shapley values) and counterfactuals as potential tools to shed light on these conditions.

Given the considerations above, I conclude with three recommendations that can be useful for XAI practitioners (researchers, developers, and providers of XAI tools) when developing or deploying the XAI approaches and AI policymakers when regulating AI with fairness in mind.

1. *The relevance of explainability for fairness should be explicitly articulated and integrated in AI development and regulation.*

Section 4.2 reports an unclear, and sometimes counterproductive relationship between XAI approaches and bias. This is backed by an increasing number of libraries, such as IBM360[38] and WhatIf tools[39], that offer explainability and fairness tools without providing specific guidance about which tools can best tackle which aspect of bias. IBM360, for example, provides two entirely separate libraries for fairness and explainability: AI Fairness360 and AI Explainability360. This leaves open the possibility that these XAI approaches may be mistakenly or deceitfully employed in their application to fairness problems; for instance, that an XAI method relevant for procedural fairness is used beyond its capacity, to investigate *which* role a protected characteristic played rather than *whether* it played a role. As I claimed in this chapter, when developing and deploying XAI approaches, XAI practitioners should think about them concerning the fairness-relevant questions they can answer and the fairness-relevant solutions that they help identify; for example, as presented in this chapter, whether they are useful for procedural fairness or to enhance or protect one's agency. Additionally, they should provide clear instructions about how their approaches can be used in concert with others and list their limitations and strengths concerning the fairness-relevant purpose they can play. It is also crucial that AI

[38] AI Explainability 360 (https://aix360.mybluemix.net/) and AI Fairness 360 (https://aif360.mybluemix.net/)

[39] https://pair-code.github.io/what-if-tool/

regulatory initiatives introduce measures that recognise and promote the coherence and complementarity of the properties of XAI methods for fairness. Without such recognition, one might miss out on the opportunities they offer but also enhance the risks entailed by their misuse. An example can be found in how one of the latest amendments to the AI Act weakened the fairness-relevant potential of explainability by shifting the focus on ensuring oversight and traceability rather than empowering end-users rights to explanations (Nannini et al., 2023).

2. *The responsibility to act on discriminatory outcomes should not lie exclusively with discriminated users.*

XAI approaches should not only be designed for discriminated users seeking advice or recourse after being subjected to discrimination. As we will encounter in Chapter 6, research suggests that many users might not even know that they are interacting with an algorithm, let alone that they have been discriminated against. It is crucial that XAI practitioners develop methods that not only help recognise instances of discrimination but also provide constructive explanations on how to address their negative impact. These explanations could suggest interventions that AI providers can undertake to prevent discrimination or to intervene on it once it occurs (Karimi et al., 2021). This would not only help protect consumers from discrimination but also help AI providers prevent future liability claims under upcoming legislative proposals (Hacker, 2022). AI regulatory initiatives increasingly rely on ensuring compliance through risk management, audits and certification (Roberts et al., 2023). Additionally, it has been suggested that external validation of models by trusted third parties can ensure the reproducibility of results and surface biases (Haibe-Kains et al., 2020). Interventional XAI approaches could be used to provide feedback to AI providers in the form of constructive recommendations from third-party audits (Watson, 2022) to help them adhere to AI regulations.

3. *Explanations of discriminatory outcomes should name conditions, rather than just the main cause(s) of discrimination.*

Hu (2019) suggests that causal graphs can be useful to explicitly state the assumptions one makes about a problem and to examine the social biases at play. Similarly, causal graphs could help represent the features that constitute a discriminatory outcome, and how they relate to each other. This approach would provide explanations that identify the primary factor responsible for discrimination and enhance our comprehension of how it relates to the set of similar conditions that contribute to an unjustifiably different outcome and how these conditions relate to one another. As mentioned, this can help shed light on the conditions by which similar individuals are treated differently. While research proposes the potential use of some fairness metrics to create a prima facie case for discrimination (Wachter et al., 2021a), XAI approaches could go further and help provide a richer understanding of systemic discrimination. This is especially relevant in light of new regulatory initiatives such as the one presented in Chapter 7 (the California Racial Justice Act (2020)), which allows the use of statistical evidence in letting people charged with (or convicted of) a crime raise issues of racial bias and discrimination. As the law relies on a counterfactual intuition [40] to prove the existence of racial disparities, counterfactual approaches could provide a richer picture of what constitutes these disparities. As suggested in the section above, this could be done by naming the features by which these two ethnically different yet similar individuals receive a different outcome as what makes their differential treatment one firmly rooted in systemic racism.

[40] Racial disparity can be suggested when a longer or more severe sentence was imposed on the defendant than was imposed on other similarly situated individuals convicted of the same offense.

4.5 Conclusion

By shifting the focus to the conditions for rather than the consequences of discriminatory outcomes, this chapter hopes to emphasise the importance of understanding and preventing algorithmic discrimination. The genealogical approach proposed here, both in its constructive and critical components, can help tailor the application of XAI approaches not only to 'see' discrimination but also to hopefully 'govern' and 'understand' its workings (Ananny & Crawford, 2018). At the same time, it can help us recognise these approaches' limitations in representing and addressing algorithmic discrimination. Pragmatically, I have provided a set of policy recommendations by which both these constructive and critical components can be integrated into AI development and regulation.

Significantly, I also recognise that XAI approaches can only go so far in matters of discrimination. Similar to what has been suggested for fairness metrics in Chapter 3, XAI approaches should be envisioned as part of a more comprehensive strategy. Thus, in this chapter, I evaluate them in their potential to support, rather than supplant, approaches to address algorithmic discrimination. Thinking with a genealogical approach in mind, future research could explore how these XAI approaches could be used in concert with qualitative and contextual efforts to (re)construct how historical disparities emerge and are reproduced in AI systems. For example, their ability to surface hints of the conditions that enable discrimination could be corroborated through qualitative data and the participation of the communities involved to build a more comprehensive and accurate picture.

Chapter 5:

How can bias as evidence be applied?

Chapter 5. Evidence of What, and for Whom? The Socially Contested Role of Algorithmic Bias in a Predictive Policing Tool

Abstract

This chapter presents a critical, applied study of the social role of algorithmic bias in the context of the Chicago crime prediction algorithm, a predictive policing tool that forecasts when and where in the city crime is most likely to occur. Through interviews with 18 Chicago-area community organisations, academic researchers, and public sector actors, I show that stakeholders from different groups articulate diverse problem diagnoses of the tool's algorithmic bias, strategically using it as evidence to advance criminal justice interventions that align with stakeholders' positionality and political ends. Drawing inspiration from Catherine D'Ignazio's taxonomy of 'refusing and using' data (D'Ignazio, 2024), I find that stakeholders use evidence of algorithmic bias to *reform* the policies around police patrol allocation; *reject* algorithm-based policing interventions; *reframe* crime as a structural rather than interpersonal problem; *reveal* data on authority figures in an effort to subvert their power; *repair* and heal families and communities; and, in the case of more powerful actors, to *reaffirm* their own authority or existing power structures. I identify the implicit assumptions and scope of these varied uses of algorithmic bias as evidence, showing that they require different (and sometimes conflicting) values about policing and crime. This divergence reflects long-standing tensions in the criminal justice reform landscape between the values of liberation and healing often centered by communities impacted by incarceration and the values of surveillance and deterrence often instantiated in data-driven reform measures. I conclude by reflecting on the possibility of using algorithmic bias as evidence to challenge the status quo to protect the interests of marginalised, affected communities.

Note: A version of this chapter has been accepted to the Fairness, Accountability and Transparency (FAccT) in ML 2024 conference. The chapter has undergone editing in order to ensure coherence with the thesis as a whole. The article was co-authored with Dr. Dasha Pruss. Please see Appendix C for authorship statements. In the following text, only quotes which are longer than one line have been indented. This is simply for graphic reasons.

5.1 Introduction

Data-driven place-based predictive policing is an increasingly popular strategy for addressing a range of issues in policing, from resource limitations to unconscious biases in decision-making (Selbst, 2017). Predicting 'hotspots' of crime is a particularly contentious instance of the broader trend of algorithm-assisted decision-making and predicting crime risk in the criminal legal system, with proponents lauding it as an evidence-based policing reform (Kennedy et al., 2018; Pearsall, 2010) and critics arguing that it exacerbates the very issues it is meant to address (B. E. Harcourt, 2008; Shapiro, 2019). Algorithmic bias – as defined in the previous chapters – has been at the center of the discourse about predictive policing, including concerns about the algorithms' reliance on 'dirty data' and the possibility of feedback loops in officers' arresting behavior (Ensign et al., 2018; Lum & Isaac, 2016; Richardson et al., 2019a; Selbst, 2017).

Standard algorithmic fairness approaches have focused on identifying, removing, or managing statistical biases (Fazelpour & Danks, 2021), with strategies ranging from pre-, during and post-deployment strategies to preventing models from learning or outputting biases (Ensign et al., 2018; Fahse et al., 2021; Yeung et al., 2021; Zhang et al., 2018). However, as mentioned in the literature review, many critics have pointed out that verifying that an algorithm is correctly 'calibrated' may not be the right question to ask when underlying data reflect structural injustices (Balayn & Gürses, 2021; Hoffmann, 2019). Indeed, the bias exhibited by predictive policing algorithms is considered by some to be an inevitable artifact of higher police presence in historically marginalised communities (Lum & Isaac, 2016; Zilka et al., 2023), a reflection of the structural race- and class-biased patterns of the criminal legal system and an unequal society more broadly, rather than a mere technical 'glitch' (Benjamin, 2019c). In response, some researchers have called for more substantive notions of algorithmic fairness that take algorithms' objectives

and sociotechnical contexts into account (Barabas, 2019; Green, 2021b; Green & Viljoen, 2020). Others have focused on providing affected communities with alternative tools to respond to the deeper societal issues indicated by algorithmic bias, such as the use of civilian dashcam[41] data to detect disparities in police deployments (Franchi et al., 2023). In short, there has been widespread disagreement about how to practically respond to evidence of algorithmic bias.

This chapter presents a critical, applied study on evidence of algorithmic bias in the context of the Chicago crime prediction algorithm (Rotaru et al., 2022), a predictive policing tool that predicts the geographic regions in Chicago in which crime is most likely to occur. This algorithm is a fitting case as its developers demonstrated that it exhibits 'enforcement bias' – that is, the uneven or disproportionate patrolling, arrests, or other actions by police toward certain neighborhoods. Through interviews with Chicago-area academic researchers, civil society organisations, and public sector employees, I find that evidence of algorithmic bias has the potential to challenge as much as to affirm the status quo. I show that stakeholders from different groups articulate different problem diagnoses of the algorithm's bias, strategically using these diagnoses to advance criminal justice interventions that align with individuals' political ends. These strategies include using data to *reform* the police that generated the biased data in the first place, such as by allocating police patrols more equitably; *reaffirm* or strengthen existing models of policing; *reject* the use of predictive policing tools; and *reframe* the narrative away from reforms like data-driven policing toward abolitionist aims, such as *repairing* communities.

[41] A video camera mounted on the dashboard or windscreen of a vehicle and used to continuously record the view of the road, traffic, etc. through the windscreen.

I identify the implicit assumptions and scope of these varied uses of algorithmic bias as evidence, showing that they require different (and sometimes conflicting) values about policing and crime. Notably, using enforcement bias to inform the more efficient allocation of police patrols diverges from the use and interpretation of the same evidence by stakeholders with lived experience in marginalised communities, who diagnose the source of the bias as systemic. This divergence reflects long-standing tensions in the criminal justice reform landscape between the abolitionist values of liberation and healing often centered by communities impacted by incarceration and the values of surveillance and deterrence often instantiated in data-driven reform measures. I discuss the implications of these differences in relation to participants' social positions, and I advocate for centering the interests and experiential knowledge of communities impacted by incarceration through participatory methods to ensure that algorithmic bias evidence can function as a device to challenge the status quo.

5.2 Background and Related Work

5.2.1 Algorithmic bias: from an object of social critique to a device for social change

Notwithstanding the widespread claim that algorithmic bias is a mirror of existing disparities (Mayson, 2018), it is rarer for algorithmic bias to be taken seriously as a source of evidence to emancipate and support marginalised communities. Bias has helpfully been identified as a boundary object (Grill et al., 2023), an object which, despite disagreement about its nature, provides a lens for critique and confrontation of different stakeholders' views. In this chapter, I take this characterisation of bias a step further; namely, I want to move from bias as an object of and for critique to a pragmatic device to address the disparities it mirrors.

Skepticism about applying a computational lens to societal problems (Broussard et al., 2019; D'Ignazio, 2024; Eubanks, 2018; Gebru, 2020; Green, 2021a) has led to a reconsideration of the roles of computational tools from optimisation and individual predictions toward systemic interventions to address deeper issues of justice and equity (Barabas et al., 2018; Green, 2021a). Recent work in critical data studies (Abebe et al., 2020; D'Ignazio, 2024; Passi & Barocas, 2019), as well as past work at the intersection of sociology and statistics (Bruno et al., 2014; Desrosières, 1998) has mapped out ways in which practices around the use of computational tools, from the design to the development and deployment of these tools, can be used to address existing disparities or injustices in the status quo.

Computational tools can help unveil or criticise specific aspects of society. Through their reliance on historical and empirical data, tools both in statistics (Bruno et al., 2014) and in Machine Learning (ML) (Abebe et al., 2020) can help disclose the wrongdoings of an institution by appropriating their own numbers. The use of computational tools can also indirectly lead to the creation or the resurfacing of some social categories (Bruno et al., 2014). The categorisation and labeling needed to make data legible can help prove the existence of a neglected or unrecognised category, attempt to redefine an existing category or seek means to defend this category and its rights (e.g., make it 'legible' to the system) (D'Ignazio, 2024). These tools can also help reformulate social problems and reshape public objectives. The exercise of problem formulation and formalisation that they entail could help reason about and reassign an institution's priorities (Bruno et al., 2014), articulate and reason on how societal problems are formulated (Passi & Barocas, 2019) and highlight the limitations presented by computational techniques (Abebe et al., 2020).

The power of numerical tools, however, can work both ways. It can expose as much as it can affirm power (Bruno et al., 2014). The same numbers that can be used by communities to expose their oppression can, for example, be used by the police to justify their intervention. Starting from these considerations, this chapter explores whether algorithmic bias can be used as a device to disclose or address the societal disparities that it mirrors. At the same time, however, it does so by exploring how the evidence of bias enables diverse problem diagnoses and supports the identification of solutions contingent upon individual worldviews. Given its potential to disclose as much as affirm power, it explores how – if at all – this evidence can be applied in practice to serve the interests of affected communities.

I connect these questions to a parallel tension in criminal justice between abolitionist and reformist responses to the crisis of mass incarceration. Abolitionist philosophy calls for an end to the reliance on imprisonment and policing and proposes alternatives that entail

> *'holistic engagement with the structural conditions that give rise to suffering, as well as the interpersonal dynamics involved in violence'* (McLeod, 2018).

Abolitionists often advocate for 'non-reformist reform', which is

> *'conceived not in terms of what is possible within the framework of a given system and administration, but in view of what should be made possible in terms of human needs and demands'* (Gorz, 1968).

Non-reformist reform prioritises decarceration and reinvestment of resources back into communities. This is in contrast to 'reformist reform', which limits its objectives to the maintenance and practicality of the current system. Most computer science applications in the context of criminal justice are reformist reforms (Green, 2019); as it can be seen in the interview results, the predictive policing tool itself turns out to be an instance of a

reformist reform, drawing a sharp line with reform advocates in favor and abolitionists opposed.

5.2.2 The Chicago Crime Prediction Tool

Crime prediction is an increasingly common reform measure, particularly when an institution is experiencing legitimacy problems and a corresponding loss of trust in human decision-makers (T. M. Porter, 1995). In response to public scrutiny toward policing scandals, over the past two decades some of the largest police departments in the US have started turning to 'data-driven' policing as a reform and public relations measure, contracting with companies like PredPol and Palantir to forecast crime 'hot spots' based on past crime incidence data (Brayne, 2020).

Chicago has a history of using algorithms to predict crime. In 2013, the Chicago Police Department (CPD) generated a list of likely perpetrators or victims of violent crime using predictive policing technology made by the Illinois Institute of Technology (Stroud, 2021). The program was terminated in 2020 following investigations by RAND and the Chicago Sun Times, which showed that every person arrested or fingerprinted in Chicago since 2013 was on the list (P. Verma, 2023). In 2018, ShotSpotter acquired HunchLab, a predictive policing company that employs sound-based technology to dispatch police to areas suspected of recent gunfire. This has led to high-intensity investigatory stops, resulting in fatal interactions between the CPD and city residents (Simkin, 2023). In March 2021, ShotSpotter was involved in an incident in which police killed 13-year-old Adam Toledo, intensifying community opposition to the technology. Notwithstanding this opposition, the contract with ShotSpotter was quietly extended for the second time, until it was recently announced that it will cease in September 2024 {Citation}. These events happened in the backdrop of the 2017 consent decree, a court order mandating broad police reform in the City of Chicago.

By contrast, Rotaru et al. (2022) present the crime prediction algorithm they developed as a different approach to predictive policing. It aims not simply to accurately predict hotspots of crime, but also to produce insights on the system of crime and its rules of organisation, as well as how enforcement interacts with, modulates and reinforces crime. Specifically, the authors aimed to demonstrate the dual role of predictive models in enhancing state power through criminal surveillance while also providing a means to scrutinise the state by uncovering systemic biases in crime enforcement (Rotaru et al., 2022). At the moment of writing, the developers of the algorithm are planning to extend the application of the algorithm to also Boston and Philadelphia.

Using crime incidence data from the City of Chicago, they devised a spatiotemporal network that inferred patterns from past event occurrences. This network, known as the Granger network, comprised local estimators forming a communicating network to predict future infractions. The researchers delved into publicly recorded historical event logs, extending their analysis beyond the training sample to events in the subsequent year. The study focused on two overarching categories of reported criminal infractions: violent crimes (e.g., homicide, assault, and battery) and property crimes (e.g., burglary, theft, and motor vehicle theft).

In addition to assessing the algorithm's accuracy in predicting future crime, Rotaru et al. (2022) argued for the presence of socioeconomic enforcement bias. Modeling the number of individuals arrested during each recorded event allowed the researchers to probe the possibility and pattern of this bias by contrasting crime rates with arrest rates. They observed how, in wealthier neighborhoods, elevated crime rates correlated with increased arrests. Conversely, in disadvantaged neighborhoods, arrest rates dropped without a corresponding decrease in crime. The researchers attributed this bias to resource constraints on police, compounded by a biased prioritisation toward wealthier

neighborhoods (Rotaru et al., 2022). The findings aligned with suspected enforcement bias in US cities, particularly bias toward suburbs with high socioeconomic status (Lipton, 1977; W. B. Meyer & Graybill, 2016). Strikingly, this pattern persisted consistently across all the years under analysis. The authors used evidence of socioeconomic enforcement to suggest a need-based reallocation of police patrols across neighborhoods in Chicago.

5.3 Methods

To interrogate the ways in which the police enforcement bias captured in the Chicago crime prediction algorithm can serve as evidence, I conducted semi-structured interviews with 18 stakeholders, primarily from the Chicago metropolitan area.

5.3.1 Recruitment and Demographics

I conducted interviews with stakeholders from various backgrounds, including academia, civil society, and the public sector. After an initial mapping of relevant actors through desk research, I selected the rest through referrals from initial responders. Snowball sampling allowed me to trace a latent network of relevant actors that I otherwise would not have identified (Parker et al., 2020). I continued recruiting and interviewing participants from each background until no new themes or information directly relevant to the project emerged from the interviews (Small, 2009) or my timeframe for data collection passed. In total, I attempted to recruit 62 participants, achieving a response rate of 29%. I made an effort to select participants with variation (Weiss, 1994) across roles (see Table 1) and level of personal exposure to the criminal legal system (see Table 2), ranging from no prior involvement to formerly incarcerated. This variation enabled me to consider the role of individuals' positionality in the results. While 8 participants had direct experience or knowledge of the algorithm, several interviewees had no prior familiarity with it; these latter interviews helped inform my understanding of Chicago-area policing more broadly.

Role	Frequency	%
Community Organizations (C)	7	38.8%
Re-entry associations	4	22.2%
Safety & Community Initiatives	3	16.6%
Researchers (R)	6	33.4%
Tool Developers	1	5.6%
Academic Researchers	5	27.8%
Public Sector (P)	5	27.8%
Public Officials	1	5.6%
Police	2	11.1%
Contractors & Oversight bodies[2]	2	11.1%
Total	18	100%

Table 1. Roles of participant population

Exposure to the CLS	Frequency	%
Justice-impacted individuals	2	11.1%
Working directly with justice-impacted individuals	7	38.9%
Employed by the criminal legal system (e.g., police)	2	11.1%
Working with the criminal legal system	3	16.7%
No direct exposure (e.g., researchers)	4	22.2%
Total	18	100%

Table 2. Participant exposure to the criminal legal system (CLS)

Due to confidentiality concerns, recruiting participants from the public sector was challenging. This is a common issue in interviews with participants in senior or official roles – so-called 'elite' interviews (Richards, 1996). I accounted for this by recruiting two additional participants, one who is a contractor for the criminal legal system and another who works for an oversight body.

5.3.2 Interview Process

Interviews ranged from 30 minutes to 1 hour, with an average length of 39 minutes. Two interviews were carried out in person and the rest via video call. Interviews were semi-structured (Edwards & Holland, 2013), allowing me to focus and dig deeply into an initial set of topics, while also letting participants' expertise guide the interview and allowing new themes to emerge (Roulston & Choi, 2018). Specifically, I asked about participants' understanding of and attitudes toward algorithmic bias generally and police enforcement bias specifically; the challenges and problems they believe to be associated with these biases; and the interventions available or those they envision in the future (see Appendix H for a full list of interview questions). The content and wording of the questions varied slightly depending on the participants' particular knowledge of or experience with the algorithm. For example, for participants with direct knowledge of the algorithm, such as its developers, more technically detailed questions were posed, while for other participants, such as members of community organisations, a stronger emphasis was placed on questions about the Chicago context and community safety.

5.3.3 Qualitative Analysis

I analysed interview transcripts by conducting a thematic analysis, using the qualitative data analysis software Nvivo. I analysed interviews iteratively, identifying recurring themes, reviewing transcriptions 2-3 times, and switching between inductive coding and data collection to clarify and strengthen themes and uncover opposing evidence (Miles et al., 2014). I converged on seven high-level themes (in order of most to least prevalence in the interview data): downsides of AI; upsides of AI; structural bias; lived experience and positionality; abolition; reformism; and controversy around policing (see Appendix I for a code table). I complemented the results from the interviews with a document analysis

(Karppinen & Moe, 2011) of policy proposals and programs produced by the interviewed stakeholders, as well as other active participants in the domain of policing and justice reform in the Chicago area (e.g., policymakers, regulators, community organisations, and think tanks). Through triangulation with findings from the interview data, I was able to elucidate, refute, or expand on the findings from the interviews (Frey, 2018).

5.4 Results

Across the board, interviewed stakeholders strategically used algorithmic bias as evidence to advance interventions that aligned with their own positionality and values about crime. Drawing inspiration from Catherine D'Ignazio's taxonomy of feminicide activists 'refusing and using' data (D'Ignazio, 2024), I see that stakeholders from different backgrounds use evidence of algorithmic bias to *reform* the policies around police patrol allocation; in the case of more powerful actors, *reaffirm* their own authority or existing power structures; *reject* algorithm-based policing interventions; *reveal* counter data on authority figures in an effort to subvert their power; *reframe* crime as a structural rather than interpersonal problem; and *repair* and heal families and communities. I will elaborate on each of these in the next paragraphs.

5.4.1 Reform: Towards '*a legitimate, ethical, fair police force.*' [P013]

5.4.1.1 Policing as an optimisation problem

Making policing more 'fair' is a standard goal of criminal justice reform (Van Craen & Skogan, 2017). Stakeholders from academic and public sector contexts tended to focus on 'over-' or 'under-policing' as the source of the Chicago crime prediction algorithm's bias, emphasising the benefits of allocating policing more optimally, such as increased public safety and police legitimacy. However, these participants had mixed views about whether predictive policing was a viable strategy to achieve this goal.

Three of the academic researchers I interviewed, including one of the tool's developers and members of a Chicago-area research group focused on criminal justice research, framed the problem revealed by algorithmic bias as a police allocation issue. One senior member of the research group [R006] explained the need for improving police presence in marginalised neighborhoods:

> *'People who live in those communities ask for more police because they're facing high levels [of violent crime] ... I mean, the reality is like somehow we are both, like, those communities are both over- and under-policed at the same time.'* [R006]

Another member of the research group [R007] posited that '*police presence alone deters crime,*' even if police make no arrests.

These attitudes aligned with perspectives voiced by two police officers. Echoing the senior academic's perspective, one officer [P003] argued that

> *'The residents in those [Black and Brown] neighborhoods, that's the silent majority. They want us there, they need us.'* [P003]

Another officer [P013] distinguished '*feeling safe'* and '*being safe*', emphasising the significance of the former and highlighting that police presence contributes to a sense of security among citizens. On this shared view, there is intrinsic value to policing, and distributing it optimally is a primary reform goal.

However, some stakeholders who were otherwise proponents of reforming police allocation expressed reservations about the use of predictive policing specifically to achieve those goals. The officer who argued that increased police presence promotes the feeling of safety [P013] warned that the use of predictive policing and other technological systems could be *'preventing cops from doing their job, which is a social job'* by making them too *'comfortable coming into an office and sitting behind a computer all day',* thereby diminishing the physical presence of police in communities and compromising the

feeling of safety. Likewise, one of the tool's developers [R002] did not want the tool to be directly used to inform police allocation decisions.

5.4.1.2 Reducing interpersonal bias

The tool's enforcement bias was also viewed by several public sector actors as a potential result of another kind of unfairness – interpersonal bias, including implicit bias, the bad judgment of a 'few bad apples', or simply limitations in human cognition. Reducing interpersonal bias among police officers through implicit bias training, data-driven policing, or more diverse police forces are standard routes of reform (Spencer et al., 2016; Van Craen & Skogan, 2017).

One county commissioner [P015] in a majority Latino district of Chicago called implicit bias *'very damaging'* to communities and explained that police officers, judges, and *'all of our criminal justice system'* must attend implicit bias training. One police officer [P013] also described what he called declining *'standards of professional behavior'* among police officers, underscoring the need for strong leadership programs, adding that this would improve community trust and legitimacy of police.

Arguing for the benefits of 'research-driven' policing, another police officer [P003] admitted that *'we're not the brightest bulbs'* but that

> *'There are smarter minds analysing all the different mathematical models in the back end of the AI, so they'll be able to inform us and they'll keep all of the AI models accountable.'* [P003]

Using the example of people's tendency to overcount the significance of a stale criminal record, one academic researcher [R006] argues that an algorithm could *'have a sort of more built-in decay mechanism than like a human could',* thus improving on limitations in human cognition.

Participants across all roles also proposed diversity, equity and inclusion (DEI) as a strategy for minimising interpersonal bias, especially integrating diverse perspectives or other stakeholders in the creation of data and algorithms. As one interviewee from a reentry association [C001] put it,

> *'Whose voice should be in the room for the algorithms that are being built, right? And I don't just mean a little bit of, yeah, engaging in participatory research where you sit down with a group of folks and this and that and you go off in your room and you build this thing. No, I mean you actually have to make them or the folks from these communities where these tools are going to be deployed an active creator of the system, right?'* [C001].

Several researchers working with justice-impacted individuals and victims of violence also stressed the importance of incorporating their voices in research.

Another participant from a reentry association [C004], however, highlighted that incorporating diverse stakeholders on its own would not solve the problem:

> *'You could take a Black person or a Brown person, or some other disenfranchised individual, and they could be the author of, you know, some new algorithm. But unless they have a specific training and knowledge around systems of bias – how, how they're constructed, how they operate and how they can be deconstructed – then I don't think that these individuals that are constructing these programs can effectively create solutions for them.'* [C004]

5.4.1.3 Increasing transparency and accountability

Beyond algorithmic bias, some participants associated another reform goal with the predictive policing tool, namely, transparency and accountability. This is a common byline in police reform rhetoric (White et al., 2021) and is often presented as a justification for 'data-driven' policing and surveillance through body-worn cameras.

Referring to the Chicago crime prediction tool, one police officer [P003] said that *'it being open source is a winner'* because of what he pejoratively referred to as the *'two big buzzwords'* around policing following George Floyd's murder in 2020, accountability and transparency. He said,

> *'Here's the open source data, you know. You look at it then I'm not hiding anything. There's no secret sauce to this. There's no ulterior motives.'* [P003]

However, other participants pointed out that the relationship between open-source and accountability is far from a given. One reentry community organiser [C001] argued that

> *'The people that build them [algorithms] should have some accountability for how they're used ... even if you made it open source, right? Even more then, in some ways, right? You have an obligation to have thought about how it could be used for better or for worse, right?'* [C001]

One of the tool's developers [R002] acknowledged the same issue with open-source, noting that *'you have to also be careful when you're giving it away'* because nothing prevents private companies from forking open-source software into a private (e.g., non-transparent) repository. The fact that police officers explicitly reported wanting to acquire the predictive policing tool underscores that even open-source software, once developed, may be used in ways contrary to the desires of their developers (Widder et al., 2022), particularly once it becomes commercialised (Widder & Nafus, 2023).

5.4.2 Reaffirm: '*How can there be a bias if in that triangle is all the same race?'* [P003]

5.4.2.1 Following the data

Participants in public sector positions used evidence of algorithmic bias to reaffirm their own authority and the legitimacy of existing policing practices. This is exemplified well by one officer [P003], who candidly explained why there is no over-policing problem:

> *'Where the crime is happening, the concentration is 99% Black or Brown and the victims are 99% Black or Brown, so that's our baseline. We're there because the data says that's where we need to be.'* [P003]

He adds that

> *'it's not that the AI is biased. The AI is just working with the data, and the data says that this geography, this demographic is, you know, these individuals.'* [P003]

This officer provides a problem diagnosis of the algorithm's behavior that essentialises the policed community – 'these individuals' – as inherently exhibiting more criminal behavior, thereby using the pattern to justify or take as given the status quo. Disparities in crime, in other words, are the factor that drives disparities in policing and in turn the algorithmic non-bias, and the officer laments that *'in policing we've been knocked around'* for simply responding to what *'the data is telling us.'*

5.4.2.2 New and improved technology

The need for more technology is also a response to enforcement bias. Some public sector participants see algorithms as an opportunity to do better, and others as the only way forward. For instance, speaking about the motivation behind the development of the crime prediction tool, one developer explained that *'all of them [previous predictive policing algorithms] were doing it wrong' and the need to 'show what is the right way to do it.'* One police officer [P003] called for the need for policing to be informed by the crime prediction algorithm and to *'try a different approach on the AI.'* These stances all show a reaffirmation of the need for more and better technology as a response to enforcement bias.

5.4.2.3 More policing

As I discussed in section 5.4.1, one response to enforcement bias is the need to increase police presence in areas with higher crime incidence, which effectively affirms the legitimacy of the current policing system. An interviewee from a community association [C011] voiced concern about affected communities similarly endorsing approaches that do not change the existing system and go against their own interests, such as *'more policing,'* and attributed this to a lack of awareness of alternative options. Referring to the race for a major election in a major US city, he said,

> *'Older Black residents who are concerned about public safety, and many of them, their response is "we need more police". And the reason is because it's the only tool they know, right?'* [C011]

This shows that individuals can support goals that are antithetical to those of their own communities. Direct experience or knowledge about disparate consequences of policing or valid alternatives can shape the solutions that individuals come to value.

5.4.3 Reject

5.4.3.1 Garbage in, garbage out

Stakeholders across all groups shared some inherent reservations about AI in the policing context, particularly concerns about its underlying data. In particular, the theme of algorithmic bias being a *'result of the system'* [C001] – that is, reflecting structural bias and systemic racism – recurred frequently. One justice-impacted participant [C005] argued that the system should not be used at all because it does not account for systemic and individual bias:

> *'It just cannot be neutral, and therefore it should not be used because people that look like me are the ones who are going to be impacted.'* [C005]

Another concern voiced by researchers working with justice-impacted individuals was a general distrust of low-quality or manipulated police data:

> *'Their [police] incentives are wrapped up in what is reported. You know what crimes are in the news and what people are saying. It's like they're kind of incentivised to obscure data.'* [R008]

Researcher participants described that documentation of police misconduct is likely to be reduced to the most minor violation; for instance, an illegal search will only be reported even when it also includes sexual assault. Another researcher working with justice-impacted individuals [R009] pointed to several cases where police officers intentionally misreported the race of individuals in order to conceal bias and discrimination in their

practices. He mentioned that in Louisiana, for example, several law enforcement agencies targeted Latino individuals but identified them as white on tickets in order to conceal discrimination (Webster, 2021), and that similar incidents have also taken place in Connecticut (Nierenberg, 2023), Los Angeles, and San Francisco (Barba & Baustin, 2023).

Finally, several participants from community organisations had reservations about the inherently interpretative component of AI, which they said biases its results. A member from a reentry community organisation [C001] argued that AI is an *'interpretive instrument'* rather than a mechanical one:

> *'In an ideal sense, let's say the [AI] system had no politics, right? But the people that will use the results and interpret the information do, and you know, as they always say, people can make statistics tell any story they want it to tell.'* [C001]

These considerations stress that not only data alone '*do not speak for themselves and don't always change hearts and minds or policy'* (Benjamin, 2019, p.206), but also that people's reading and interpretations play a role in changing the nature and narrative of data itself.

5.4.3.2 Unneeded, unconvincing and resource intensive

Other respondents were opposed to the predictive policing tool for more pragmatic reasons. For instance, several participants felt that the algorithm's enforcement bias did not show anything that was not already known. One academic researcher [R006] commented that developing an algorithm is *'a pretty inefficient way'* to inform police reallocation, arguing that *'you wouldn't need to develop an algorithm to reach these same conclusions.'* An interviewee working for a reentry association [C017] expressed similar skepticism about the need for a tool that predicts crime:

> *'He [formerly incarcerated individual] is always going to be discriminated against in one form or fashion, maybe an employment house. He's always going to be there, and that's a predictor of recidivism or returning to be a criminal. Well, yeah. We don't need an algorithm to say that.'* [C017]

The academic researcher [R006] also said that the algorithm was 'not particularly convincing' for other stakeholders:

> *'I barely understand exactly what they're doing and how they reach their conclusion, ... And like, I have a PhD and I do this for a living. So I think, you know, trying to convince other stakeholders is very difficult.'* [R006]

Other participants, including another academic researcher and several community organisations, also reported not understanding the paper or how the tool worked, though in the case of one police officer [P003], who indicated that was a positive feature because it indicated that the tool's developers were *'smarter minds'*.

The same police officer [P003] also commented on the resource-intensiveness, both in terms of computational power and departmental funding, of acquiring and using a predictive policing tool in practice:

> *'Will we have a budget for this? Or do I need to go out and fundraise for it and convince foundations that, you know, this is the way we need to use the limited resources that we have?'* [P003]

Understanding the underlying motivation for developing the tool helps contextualise the above criticisms. One of the tool's developers [R002] explained that the team had been working on a general algorithm and was looking for a real-world test case to demonstrate how the algorithm could be instantiated:

> *'Why crime? No particular reason. We were looking for a problem that would be an impactful thing.'* [R002]

An advantage of working on crime, they add, is that there is *'lots of data on it.'* The research team had access to compute power that a typical community organisation or police force does not, and it is clear that the team did not have the goal of convincing stakeholders or revealing surprising patterns. Rather, the research team simply saw predicting crime as an 'interesting' challenge, seeking the 'technical sweetness' (K. Hill,

2023) that engineers feel when they encounter the solution to a challenging technical puzzle.

5.4.4 Reveal: *'Data that helps us understand how the police see our communities'* [R009]

Enforcement bias was one type of evidence community organisations used to engage in sousveillance, or 'watching from below', in which the traditional framing of panoptic state surveillance is inverted, allowing people to observe powerful state actors (Mann et al., 2003; Newell, 2020). Researchers and community organisers described how they co-opted surveillance methods to watch the watchers. Alternate data sources, or counterdata (D'Ignazio, 2024), reveal issues with policing, remediate problems with how police and victim data are collected and reported, and allow alternate narratives to emerge.

One member of a community safety initiative [C101] lamented that

> *'a lot of times what criminologists will tend to do is they'll just look at the data. And I think what's missing is that there's a story behind it.'* [C101]

To provide this missing context, one human rights researcher [R009] describes working with grassroots groups that do on-the-ground documentation through cop watching and court watching, gathering counterdata on police officers, judges, and prosecutors. They argue that *'you need to record different things than what the courts are'* in order to include context normally absent in court data. The same researcher also describes using data about human rights violations to understand patterns of violence, and even to make statistical estimates of events that are not represented in the data.

As I discuss in section 5.4.3, several researchers working with justice-impacted individuals also reported the existence of low-quality, unreliable or manipulated police data. By collecting data from the Chicago Police Department, the City of Chicago, the

Civilian Office of Police Accountability through Freedom of Information requests, one research group [R008] described how they published an alternative dataset about police misconduct in Chicago. They also recounted collaborating with justice-impacted individuals and victims of police violence to build a basic classifier to predict police misconduct using complaint data generated by community members. This alternate algorithm revealed a culture of abuse within policing:

> *'instead of something like operation and personnel violation, you know, we're able to see clusters of, you know, families who are like ... "this is how police treated me". And what was really fascinating about that project is that we were able to move away from this idea of individual officers, like, back up to what is maybe the culture of policing that exists, what are the kind of tactics that that people are reporting that connect them across space and time?'* [R008]

Interestingly, as mentioned in section 5.2.2, one of the aims explicitly stated by the developers in building the tool was also to provide a means to scrutinise the state by uncovering systemic biases in crime enforcement. However, the kind of data considered in the tool (e.g., violent and property crimes) and their source (public data from the city of Chicago) did not allow for the same kind of revelations made by the research groups above. This stresses the importance of using alternative data and the involvement of multiple actors and institutions in revealing a more complete and less 'biased' picture of injustices.

5.4.5 Reframe and Repair: *'Deconstructing the status quo and creating a new status quo'* [C004]

Countering mainstream understandings of crime among respondents from the public sector and some academic researchers, many participants from community organisations presented alternate data epistemologies (Ricaurte, 2019) focused on abolitionist understandings of the *'root of the problem'* [C016]. In this vein, participants responded to enforcement bias by calling for the need to recognise systemic racism, reframe the

narrative toward an understanding of crime as structural rather than interpersonal, and repair and build healthy communities.

5.4.5.1 Recognition of systemic racism

A common thread in the interview data was the *'tale of two cities'* [C001; P015], or Chicago's history of segregation and systematic disinvestment in its Black and Brown communities. Participants emphasised the significance of Chicago's history of segregation, citing it as a root cause of not only crime but also numerous other social problems, such as disparities in life expectancy:

> *'I want to say 20-30 year difference in life span between those that live in the poor neighbour community.'* [C001]

and in wealth:

> *'You can go to the different parts of the north side and you can just see the wealth, right? ... and then you can go to the south side or the southwest side and see completely different– you can almost think that you're in a different city.'* [P015]

Some justice-impacted individuals and researchers working with them also remarked on the intrinsic suspicion they witness toward policed communities, accompanied by a tendency to trust or believe who does the policing. One of them [C017] argued that criminal legal policies are *'setting people up for failure,'* and that *'when you talk about [criminal] background, it's all racism'*. He used Nixon's 1970s policy on potential delinquents as an example of how criminals are created out of assumptions rather than facts. This policy, which included new statistical systems to supervise suspects released on bail like SEARCH (System for Electronic Analysis and Retrieval of Criminal Histories) allowed officials to establish criminal records and backgrounds which then brought mere suspects into the criminal justice system (Hinton, 2017).

Multiple participants stressed that the first step to resolving the problem presented by enforcement bias is having a clear understanding of what the problem is. One member of a community safety initiative in Chicago [C011] noted that

> *'Before anybody can heal there needs to be a reckoning, there needs to be some kind of identification that the systems that were created actually have caused harm.'* [C011]

Two more participants [C004; C005] echo this sentiment almost verbatim, arguing that

> *'If we're not going to address a system of white supremacy, then we're not going to get to any real hard solutions.'* [C004]

5.4.5.2 Narrative change

Beyond reckoning, participants emphasised reframing the narrative around justice-impacted individuals, police, and community safety. As one community worker [C005] put it,

> *'The delusion that somehow we can scare people out of crimes of poverty is ridiculous. We have been so indoctrinated by current systems to accept the claims, regardless of the reality, right?'* [C005]

She argues that *'One of the biggest things that we do is narrative change work'* in order to push back on dominant narratives around crime and policing.

In particular, participants called for a reorientation toward a structural understanding of the *'root cause'* [C014] of crime, focusing on treating the cause rather than the symptom. Two participants criticised interventions that focus on the interpersonal level, with one participant [C011] comparing the current criminal legal system to a hospital that only contains an emergency room and another [R104] lamenting that interventions like community violence and prevention programs:

> *'Stop at the interpersonal level, but they don't also account for the structural systemic impediments at play.'* [R104]

5.4.5.3 *'Building healthy communities rather than reducing crime.'* [C001]

Grounded in an alternate understanding of crime as structural, participants responded to enforcement bias by advocating for non-reformist reforms such as police alternatives and rethinking how to use data to promote community safety.

One community organiser [C001] asked,

> *'Instead of thinking about how do we reduce violence and crime, what if we pose the question of how this information helps us build healthy communities?'* [C001]

We saw in section 5.4.4 how researchers and activists gather counter data on police and make up for limitations in data. This strategy centers on the needs of affected communities, which community organisation participants agreed got little benefit from predicting crime. Another researcher [R008] explained how AI could instead be used to reduce crime in the sense of supporting survivors of violence, especially police violence:

> *'When we pivot our approach to, you know, engaging with survivors then we're better understanding the implications of policing and the ways that we can use artificial intelligence to support survivors of violence and reduce crime, like, period.'* [R008]

Beyond the use of data, several participants also advocated for non-reformist reforms, including alternatives to police. Drawing on an abolitionist critique of police, one researcher [R008] working with justice-impacted individuals challenged the assumptions underlying policing:

> *'Policing should not be about "we gotta get the bad guys", right? It should be about "how do we keep people safe".'* [R008]

Taking this view a step further, one community organiser [C101] questioned the belief that 'we need more [police]', adding that

> *'If we bring more tools to the community, there's– the solutions are here. We just haven't given you all the equipment to do so.'* [C101]

Notably, the police and academic position on police promoting safety was starkly different from the perspective of some interviewees from community organisations for which police presence was an *'enforcement presence, not a not a deterrent, but an enforcement'* [C001].

Even when there was agreement about the need to strengthen communities, however, different actors have different views on which kind of interventions to put in place to do so. Some reentry organisations advocated for giving the same rights to justice-impacted individuals or covering their past criminal record after some time [C017], while others fought for more police and state accountability [R008]. One individual working as a contractor for the public sector [P016] recognised the structural nature of the problem and yet focused on *'changing the individual'* as a solution, namely, moving individuals out of a crime-heavy environment or alternatively, strategically reclaiming crime-heavy spaces by organising community events (e.g., sports competitions, community barbecues, etc.). Another member of a community organisation [C011] focused on *'changing the conditions'* – namely, focusing interventions on the socioeconomic determinants of crime such as wealth and health.

5.5 Discussion

These results shed light on the socially contested nature of algorithmic bias applied within the realm of predictive policing, emphasising that bias is not understood as a singular challenge but rather as evidence of a multifaceted package of problems with a corresponding broad range of possible interventions. I grouped the interview findings depending on whether stakeholders from different backgrounds use evidence of algorithmic bias to *reform*, *reaffirm*, *reject*, *reveal*, *reframe* and *repair*, showing the ability of enforcement bias to be used both to support and challenge the status quo.

Whether and how algorithmic bias was seen as useful for advancing each of these goals varied according to each stakeholder's broader understanding of the source of police enforcement disparities – structural inequality versus interpersonal bias – and the set of interventions aligned with each respective problem diagnosis. Stakeholders from different professional roles and positionalities tended to articulate problem diagnoses of the bias aligned with their own relationship to the criminal legal system, though I saw a diversity of attitudes and responses within each subgroup and points of convergence between individuals with different positionalities and roles. For instance, I noticed substantial overlap among interviewees from all backgrounds regarding the importance of community safety and potential downsides to the use of AI and data in the context of policing, such as AI's possibility to diminish the role of nuanced human discretion. On the other hand, I saw disagreement between groups about whether police presence promotes community safety, and there were differences within groups relating to whether predictive policing tools would improve policing and whether incorporating affected stakeholders in the development of predictive policing could improve the tools' viability.

Many of the inter-group differences I saw reflect tensions between the assumptions underlying interventions grounded in reform versus abolition – reformist reform that operates within the logic of the current system and non-reformist reform that aims for longer-term systemic change, respectively. Along these lines, I saw other contested concepts emerge, such as the notion of community safety and 'overpoliced' communities, with the connotation and valence of 'overpoliced' changing depending on the socioeconomic status of the overpoliced community. For example, in their paper, developers of the Chicago crime prediction algorithm suggest that the police are significantly more present in wealthier neighborhoods because those areas have more resources. 'Overpoliced' here has a positive connotation – the privilege and safety of well-

resourced communities. At the same time, interviewees from community organisations and from the public sector tended to refer to overpoliced communities as poorer communities with higher rates of crime.

I examine one significant binary that correlated with participants' attitudes toward the criminal legal system – the 'AI hype' spectrum – and end by discussing how different positionalities and attitudes can be reconciled to apply evidence of algorithmic bias in the service of the affected communities.

5.5.1 AI Hype Rhetoric

With stakeholders from all groups, except the academic researchers most closely familiar with the tool's development, attitudes toward predictive policing tended to fall into one of two AI hype narratives: optimism about AI's potential to 'disrupt' the criminal legal system in a beneficial way, and inherent opposition to AI on the basis that it is racist and fundamentally flawed. Techno-optimism mapped onto the extent to which a stakeholder understood the problems the algorithm was capturing as primarily interpersonal rather than structural, which correlated with positionalities in higher positions of power. Some endorsers of structural inequality narratives still saw value in certain uses of AI, such as grassroots classifiers to counter narratives put forward by police.

The clearest example of disruption rhetoric came from officer P003, who repeatedly emphasised his optimism about the *'smarter minds'* developing the tool. Initially, he appears to be endorsing AI as a solution to a structural problem: *'We're using technology to break down the cycles that have been being perpetuated for years.'* [P003]. However, it soon becomes clear that what he means by this is to send more police to crime-afflicted Black and Brown neighborhoods in order to deter crime, rather than put forth any changes to the structure of the current system or 'break down' any cycles. The

officer also remarks repeatedly that AI could help the police force's *'marketing ability to brand us as open as transparent and as accountable,'* what he calls the two salient *'buzzwords'* of the *'2020, BLM, post-Floyd'* era.

AI may yet become a successful marketing tool for police forces, similarly to past cases in which police acquisitions of technologies like body-worn cameras strategically co-opted community calls for accountability to justify larger police budgets (Karakatsanis, 2019; Umansky, 2023).[42] However, the underlying claim that predictive policing promotes transparency and accountability in policing is to be scrutinized. Sociologist Sarah Brayne, for example, provides a compelling argument that predictive policing simply displaces discretion to other, less publicly visible stages of the policing process, thereby reducing both transparency and accountability (Brayne, 2020).

On the other end of the AI hype spectrum, I saw community organisers describe AI as inherently racist. This is exemplified by C005, a formerly incarcerated leader of a reentry organisation, who argued that algorithms should not be used at all in the context of the criminal legal system because *'it's just racist and it needs to be stopped.'* Those in positions of power, they argued, are incentivised to use algorithmic systems because *'it keeps the status quo in place'* which is what gives those individuals power. Like P003, C005's perspective paints a broad and uniform brush across all uses of AI in the context of the criminal legal system.

Not all the participants I interviewed fell into this AI hype dichotomy, but the polarity was widespread enough that I believe it is important to explicitly take into account in future work that aims to use algorithmic bias as evidence to challenge the status quo. In

[42] Police are trained to manipulate and withhold body-worn camera footage to their own benefit (Umansky, 2023).

practice, existing empirical research of evidence-based reforms suggests that they have minimal impact and that these effects wash out over time (Pruss, 2023; Stevenson, 2017). As I explain in the following section, participatory research is crucial to develop responses to algorithmic bias that are attentive to the interests and knowledge of communities impacted by incarceration; at the same time, such work would be strengthened by helping participants resist one-dimensional narratives about the benefits and harms of technology in the criminal legal system.

5.5.2 *'Bias as defined by who, right? So ultimately, who's the judge?'* [C001]

The range of problem diagnoses and uses of algorithmic bias can suggest a 'relativistic' stance toward the issue, where the understanding of algorithmic bias is contingent upon an individual's subjective worldview. Going back to the research question – *'How can bias as evidence be applied in practice?'*, this plurality of views may seem to suggest that bias can be used to motivate any intervention that aligns with a stakeholder's own positionality. However, rather than dismissing these findings as mere relativism, I believe that they provide a valuable lens to understand the conditions under which algorithmic bias becomes a device to reinforce or challenge power dynamics. Drawing on standpoint theory (Harding, 1992), I argue that unravelling the relationship between bias and positionality serves as a critical tool to evaluate shortcomings and mistaken assumptions in potential responses to algorithmic bias coming from actors in relatively high positions of power, such as police officers and even academic researchers.

Standpoint theory (Harding, 1992) posits that individuals' social positions shape their access to knowledge, influencing how they interpret and respond to phenomena such as bias. Marginalised social locations have better awareness of injustices and can more easily question the 'naturalness' of existing power structures (Bowell, n.d.). In the context of algorithmic bias and predictive policing, this means that those in positions of power,

such as public sector actors and academics, may inadvertently reaffirm fundamentally unjust systems due to their distance from the lived experiences of communities impacted by incarceration.

These findings emphasise the importance of centering the needs and experiential understanding of justice-impacted communities through community-based participatory research in designing and evaluating responses to enforcement bias. By doing so, responses to enforcement bias can be designed and evaluated with a more nuanced and contextually informed approach. A critical consideration of positionality and bias may promote interventions that can serve affected communities' stated goals of community investment and healing, contributing to the development of interventions that address the root cause of algorithmic bias in predictive policing.

5.6 Conclusion

In this chapter, I presented a critical, applied study on evidence of enforcement bias within the Chicago crime prediction algorithm, as detailed in Rotaru et al. (2022). The research, based on interviews with various stakeholders, including community organisations, academic researchers, and public sector figures in the Chicago area, reveals that stakeholders offer diverse diagnoses of this bias, strategically leveraging these diagnoses to promote interventions that align with their respective positionality.

These strategic interventions encompass a range of approaches, such as using data to reform police practices that contributed to biased data collection; advocating for increased transparency, accountability, and education regarding the development and utilisation of predictive policing tools; discontinuing the use of such tools altogether; involving a diverse set of stakeholders in the creation and evaluation of these tools; and proposing a shift away from investing in data-driven policing towards more abolitionist

goals, such as community investment. I examine the implicit assumptions and the scope underlying these diverse responses to algorithmic bias, demonstrating that they are grounded in reformist and abolitionist thinking. I conclude by reflecting on the divergent AI hype views reported by participants and the importance of participatory research to enable algorithmic bias to subvert the status quo.

5.7 Ethical Considerations

5.7.1 Researchers' Positionality

As a white woman academic researcher from Europe, I acknowledge that my personal and professional background shaped how I selected research questions and analysed data, as well as the manner in which I interacted with study participants from backgrounds similar and dissimilar to mine. I have a privileged academic affiliation, race, and lack of personal contact with the criminal legal system, which undoubtedly shaped the information that grassroots community organisations and justice-impacted interviewees were comfortable sharing with me. In fact, the theme of exploitation of marginalised communities by academic researchers came up explicitly in one interview, which I discuss below. On the other hand, my privileged position as white academic researchers from well-recognised institutions helped build rapport with other academic researchers in the study, who had similarly received training at elite academic institutions.

5.7.2 Ethical Considerations

5.7.2.1 Privacy and Anonymity

In line with ACM Ethics Guidelines and the requirements of CUREC approval, participants' identities were protected by pseudonymising interview transcripts. The

document matching pseudonym to identifiable information was stored separately from the research data.

5.7.2.2 Engaging non-extractively with affected communities

The primary ethical challenge I encountered had to do with how to involve stakeholders impacted by the criminal legal system in the research in a non-extractive and mutually beneficial manner. Due to resource constraints, I did not compensate interviewees for their time, which meant there was no direct benefit for study participation, and my positionality meant that affected stakeholders had little incentive or reason to place their trust in me. This became particularly salient at the end of an interview with a formerly incarcerated participant, who works with returning citizens in Chicago. In response to a question about referring me to other study participants, they responded,

> *'I'm going to be completely transparent with you. Our community is exploited, often by academia. And, so I don't feel comfortable because of that constant exploitation to put other people in that position. [Academic researchers] take our expertise, they subject us to trauma by asking us to tell our stories, and then they never come back and connect it to us.'*

I acknowledge the respondent's concern and plan to share all research outputs with research participants, crediting them when desired, with the understanding that I cannot give back as much as I have received from their participation. The experience of this power differential demonstrated to me the importance of building longer-lasting community relationships outside the scope of academic projects prior to engaging in research, and increasing communities' ownership and participation in the research, with the gold standard in mind of participatory action research.

5.7.3 Adverse Impacts

By focusing on the Chicago Crime Prediction tool, my research may inadvertently publicise the availability of the tool or promote more intensive policing. Beyond

recognising this limitation, I wish to make explicit that I do not support the development of this (or similar tools) in any form. I hope that, by exposing the tool and the different problems it raises, I can shed light on how and why solutions addressed at tackling bias often do not end up benefiting the affected communities.

Chapter 6:

How can bias as evidence inform policy frameworks to tackle discrimination?

Chapter 6. Evidence of bias: Bridging Information Gaps in The EU AI Liability Directive

Abstract

The proposed European AI Liability Directive (AILD) is an important step towards closing the 'liability gap', i.e., the difficulty in assigning responsibility for harms caused by AI systems. However, if victims are to bring liability claims, they must first have ways of knowing that they have been subject to algorithmic discrimination or other harms caused by AI systems. This 'information gap' must be addressed before a discriminated user can build a claim of algorithmic discrimination and for the AILD to meet its regulatory objectives. Through qualitative analysis of the AILD proposal document (H. Morgan, 2022), this chapter examines its capacity to close the information gap and ultimately, meet its regulatory objectives. I argue that the current version of the AILD reduces legal fragmentation but not legal uncertainty; privileges transparency and disclosure of evidence of high-risk systems over notification of harm and discrimination; and shifts the burden on the claimant from proving fault to accessing and understanding the evidence provided by the defendant. I conclude by providing four recommendations on how to improve the AILD to address the 'information gap' and indirectly the 'liability gap', necessary steps toward using evidence of algorithmic bias to inform the regulatory framework.

Note: This chapter has been published in the European Journal of Law and Technology (see (Ziosi et al., 2023)). The chapter has been expanded in order to ensure coherence with the thesis as a whole. The original article was co-authored with Dr. Jakob Mokander, Dr. Claudio Novelli, Prof. Edoardo Celeste, Prof. Luciano Floridi and Prof. Mariarosaria Taddeo. Please see Appendix A-G for authorship statements.

6.1 The EU AI Liability Directive: background, scope, and purpose

The EU has produced several initiatives to regulate artificial intelligence (AI),[43] foremost amongst which is the Artificial Intelligence Act (AI Act)[44] (AI Act, 2021) proposed in April 2021. The AI Act proposes a comprehensive regulatory framework and an oversight structure to prevent harms from AI-related risks. However, it leaves certain areas less explicitly defined, such as the specific redress mechanisms available when discriminatory outcomes occur. For instance, while the AI Act mandates that AI systems are designed and developed in compliance with existing non-discrimination laws, it does not detail the procedures for individuals to seek redress if they are harmed by AI-driven decisions, such as a credit scoring algorithm that disproportionately disadvantages candidates from certain ethnic backgrounds, thereby perpetuating racial biases in credit.

Since 1999, the European Commission's liability regime relied on the EU Product Liability Directive (PLD) (Product Liability Directive, 2022). However, a 2018 evaluation report (Evaluation of Council Directive 85/374/EEC, 2018) of the PLD identified important shortcomings regarding AI; specifically, whether AI software would count as a product, new types of AI-related risks (e.g., cybersecurity breaches), and specific barriers to proving harm due to specific aspects of AI technologies (e.g., complexity, autonomy and opacity). As a result, the proposed AI Liability Directive (AI Liability Directive, 2022) and a revised Product Liability Directive (Product Liability Directive, 2022) were published in September 2022. While an interinstitutional agreement has been reached concerning the PLD (Bertuzzi, 2023b; Hanks et al., 2024), legislative progress on the

[43] As mentioned in the literature review (Chapter 2), other examples are the revision of the General Product Safety Directive and Machinery Directive.

[44] I shall consider the first Draft (COM/2021/206 final) and the final draft of compromise amendments to the same, (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)). While I cite the first draft, a link to the latest one can be found here: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

AILD is currently stalled. However, there are expectations that it will be taken up again after the coming into force of the AI Act (Novelli et al., 2024).

The two directives were designed to be complementary to each other (Hacker, 2022) The Product Liability Directive (PLD) is concerned with strict liability [45] and it encompasses both physical goods and software, including AI [46]. It is applicable to manufacturers and other entities within the supply chain, such as remanufacturers and businesses that significantly alter products, provided certain conditions are met. It also extends its coverage to defective products leading to instances of physical injury, property damage, and data loss. On the other hand, the AILD is concerned with fault-based liability [47] and it focuses specifically on AI systems. It encompasses claims against manufacturers, but also professional (providers) and non-professional (consumers) users. The AILD may also involve the violation of fundamental rights and primary financial loss. Importantly, it also covers cases of algorithmic discrimination.

If passed, the AILD will transform the legal landscape for companies designing and deploying AI systems, as well as users who have incurred harm. It aims to:

(1) harmonise legal regimes and reduce legal uncertainty;

(2) prevent liability gaps between providers and users of AI systems; and

(3) make the process of compensation for injured parties easier and more effective.

---

[45] A defendant held to strict liability must compensate for damages regardless of whether they adhered to or violated a specified standard of conduct. The action that causes harm is all that is required for liability to be triggered.

[46] Free and open-source software that is developed or supplied outside the course of a commercial activity is excluded from the scope of the directive (De Luca, 2024).

[47] Fault-based liability is based on the notion of ascribing responsibility for a specific wrongdoing. Ascribing responsibility rests on proving that the defendant has been 'at fault,' i.e., having failed to adhere to specific standards of conduct.

The so-called ‘liability gap’ refers to the difficulty of ascribing responsibility for harm caused by AI systems. Sometimes, it may be challenging to determine whether and how to allocate various types of liabilities for an AI system's misconduct among designers, developers, deployers, or users. Consider the case of algorithmic discrimination against protected categories such as sex and race. Referring to the examples from the previous chapters, it may be difficult to explain why or how women are systematically offered less favourable loans than men through credit risk assessment tools or why people of colour are more likely to be labeled at high-risk of crime by risk assessement algorithms. This difficulty derives from those properties of AI systems, like complexity, autonomy and opacity, coupled with the fact that these systems frequently function within intricate socio-technical environments (Novelli et al., 2023). These factors collectively weaken accountability chains and make it difficult to determine who is responsible for discrimination (Z. Porter et al., 2022; Rowland et al., 2022).

Closing this gap is essential, but it is not the only relevant concern. A victim of algorithmic discrimination may be unaware that an AI system has produced that outcome or may not know that they have been discriminated against (Santoni de Sio & Mecacci, 2021). The person of colour who is classified as high-risk, for example, might not be aware that the police in their neighbourhood rely on algorithms for crime prediction. A woman who is not offered the same loan as a man might not even realise the difference. Even when they are aware, they might have difficulty accessing or understanding the evidence needed to bring about a liability claim for algorithmic discrimination. The kind of technical documentation available for AI systems, for instance, might prove inaccessible for a non-technical audience 30/06/2025 09:42:00. I refer to this lack of knowledge as the ‘information gap’. This ‘gap’ must be addressed for a user to successfully build a claim of algorithmic discrimination. Additionally, this information is essential to ascribe

responsibility. Thus, I argue that the information gap is primary and necessary in order to let evidence of algorithmic bias inform the AILD, as well as to fill the accountability gap and achieve the AILD's regulatory objectives.

In theory, the AILD should close both the liability gap by reducing the 'burden of proof' for claimants and the information gap by granting claimants the right to access information about high-risk AI systems to prove faults (AI Liability Directive, 2022). In practice, however, the AILD fails to address the information gap adequately, leaving questions about informational relevance, distribution, and asymmetries unanswered. The central question is whether, and if so how, the AILD grants the right of access to evidence in ways that can successfully inform users' claims for algorithmic bias and discrimination. Especially, I explore whether it shifts the 'burden of evidence' – the challenges related to accessing and understanding relevant evidence - in a way that is sufficient to reduce information asymmetry between claimants (e.g., users) and defendants (e.g., developers). I do so by engaging in a qualitative analysis of the AILD proposal document (H. Morgan, 2022): a structured review of the declared aims of the act, concerning their ability to close the information gap and meet the AILD regulatory objectives.

Overall, I argue that the AILD reduces legal fragmentation but not legal uncertainty for discriminated users; privileges transparency and disclosure of evidence of high-risk systems over knowledge of harm and discrimination; and shifts the burden on the claimant from proving fault to accessing and understanding the evidence provided by the defendant. Therefore, I contend that the AILD – as currently conceived – is unlikely to meet its regulatory objectives and, as mentioned above, effectively reduce information asymmetry between the claimants and defendants.

The next paragraphs will analyse the AILD's aim to harmonise legal regimes and reduce legal uncertainty (section 6.2), prevent liability gaps due to complexity, autonomy

and opacity of Ai systems (section 6.3); and make the process of compensation for injured parties easier and more effective (section 6.4). I conclude with some recommendations for improving the AILD as currently proposed (section 6.5). Overall, this chapter aims to provide an explorative 'deep dive' into the possibility of using evidence of algorithmic discrimination to inform the application of a specific regulatory framework; in this case, the proposed EU AILD. Given that the AILD is at the proposal stage, the claims here presented should not be taken as conclusive findings; rather, as informed considerations that can point EU policymakers towards future directions and provide recommendations.

6.2 Harmonisation and legal (un)certainty

The AILD's first objective is to harmonise non-contractual civil law claims[48] for damages caused by AI across the EU to avoid legal fragmentation and reduce legal uncertainty. Its Explanatory Memorandum states that

> *'In the absence of EU harmonised rules for compensating damage caused by AI systems, injured persons would be faced with 27 different liability regimes.'*

and that, currently,

> *'If a victim brings a claim, national courts […] may adapt the way in which they apply existing rules on an ad hoc basis to come to a just result for the victim'* and that *'That will cause legal uncertainty.'* (AI Liability Directive, 2022, Explanatory Memorandum)

However, the solutions proposed in the AILD still expose victims to uncertainty about legal treatment and the scope of the legal regime.

First, the AILD can achieve harmonisation only regarding procedural aspects of AI systems' liability (Hacker, 2022). It states that

[48] Judicial claims unrelated to contract breaches. They include personal injury, defamation, property damage, and similar issues where one party believes it has been wronged and seeks compensation.

> *'This Directive should not harmonise general aspects of civil liability which are regulated in different ways by national civil liability rules, such as the definition of fault or causality [and] the different types of damage that give rise to claims.'* (AI Liability Directive, 2022, Explanatory Memorandum)

Therefore, national regulators can introduce laws favouring claimants, such as complete reversals of the burden of proof. However, minimum harmonisation could lead to different treatment for claimants depending on which state they are in. It also incentivises tech providers to act in countries with less stringent liability laws, exposing users to higher likelihoods of harm or lower guarantees for compensation.

Second, uncertainty remains regarding the AILD's material scope. Several of the specifications on the scope of application of the AILD rely on the AI Act (Hacker, 2022). At the moment of writing, a political agreement has been reached on the AI Act (Volpicelli, 2024). However, it is important to monitor potential improvements introduced in the next legislative phases. Changes might be brought through delegated acts and implementing acts (Novelli et al., 2024), including changes to the list of high-risk use cases and the definition of AI (Uuk, 2024). This could indirectly affect the material scope of the AILD.

For instance, the AILD requirement for providers to disclose evidence applies only to 'high-risk systems' (AI Liability Directive, 2022). At present, there are open questions regarding whether and under what conditions the AILD will apply to other systems that pose 'systematic risks' and yet, are not classified as 'high-risk', such as to General Purpose AI (GPAI) (e.g., LLMs) (Novelli et al., 2024). In the drafting stage, these systems were proposed to be classified as high-risk by default. However, at the present stage, a separate tiered risk classification has been developed for them, with a set of distinct overarching obligations that apply. These obligations will apply 12 months after the entry into force of the AI Act (Novelli et al., 2024). Even though the entry into force is not expected to happen earlier than 2026, it is unsure whether discussions on the AILD will have been

taken up again by then, and at which stage they will be. Until then, it is unsure how these changes to the AI Act will affect the scope of the directive. This is concerning, especially given the current deployment of GPAI models bringing about new cases of bias and discrimination across sectors in society (Martineau, 2022).

Regarding the definition of an AI system in the AI Act, it adheres to the most recent definition of AI set forth by the Organisation for Economic Co-operation and Development (OECD). In November 2023, EU policymakers opted to suspend discussions on this matter in light of the OECD's decision to revise the definition (Bertuzzi, 2023a). Given that, as previously mentioned, the definition of AI stands as one of the sections subject to potential amendments, similar future developments may carry significant implications for the AILD. On the one hand, this can increase uncertainty for companies claiming to use AI in their business (Hacker, 2022). Companies might not know whether their products can legitimately be marketed as AI systems under evolving definitions. On the other hand, this can generate confusion for users who seek legal action upon having been wronged (e.g., discriminated against). For example, when a user faces biased decisions in loan rejections from a financial institution, they may be unclear whether this experience falls under algorithmic discrimination as per the AI definition, or whether it should be addressed under traditional consumer protection laws related to unfair practices.

6.3 AI systems' complexity, autonomy, and opacity

The AILD's second objective is to close the liability gap created by AI systems' complexity, autonomy, and opacity. Specifically, it enables potential claimants, who previously questioned the AI system provider without success, to request courts to

> *'order disclosure of evidence about specific high-risk AI systems that are suspected of having caused damage.'* (AI Liability Directive, 2022, Article 3(1))

Following disclosure, a presumption of causality for the damage can be triggered by showing a defendant's *'lack of compliance with a duty of care under Union or national law.'* The defendant can rebut this presumption, typically by demonstrating that such lack of compliance could not have resulted in the reported damage (AI Liability Directive, 2022, Article 3(5)).

While the order to disclose evidence applies only to high-risk systems, the AILD extends it to non-high-risk systems that are considered opaque, and it relieves from it high-risk systems on which sufficient evidence is already available through documentation pursuant to the AI Act (AI Liability Directive, 2022, Article 4(5) and (4)). Collectively, these measures incentivise defendants to disclose information about AI systems. However, they still do not address the information gap in cases where individuals do not know that they are interacting with an AI system, let alone where they have been discriminated against.

Article 52 of the AI Act delineates transparency obligations, mandating that AI providers inform users whenever they engage with an AI system (unless this is evident from the context) or are exposed to systems such as emotion recognition, biometric categorisation, GPAI models, or deepfakes (Ferguson et al., 2022; Hupont et al., 2022). However, it remains unclear how users should be notified. This is a critical limitation, given that the effectiveness of such communication largely depends on when and how it takes place. Drawing parallels with cookie consent notices, a study found that despite the legal requirements for explicit and informed consent, the majority of websites place cookie notices at the bottom of the page and allow users to proceed without responding (Utz et al., 2019). Moreover, manipulative tactics are employed to steer users towards consenting, exemplified by the emphasis placed on the 'agree' button (Lomas, 2019). Similarly, the absence of specifications regarding notification methods increases challenges to the

effectiveness of transparency obligations. In turn, it also increases the likelihood that users remain unaware of their interactions with AI systems, or that they consent to this interaction without a clear understanding of the possible implications.

To submit claims, users must either know or reasonably suspect harm and provide adequate facts and evidence to validate the likelihood of a damages claim. However, such knowledge is not necessarily easily obtained. For example, users may not know, nor suspect, whether an AI system's decision results from algorithmic bias that creates unlawful discrimination, and they cannot typically access the required information from the system's output logs. Some cases, such as the refusal of a loan application, might create incentives to enquire. In other cases, the impact of discrimination is more subtle, and the outcome of the AI decision may never be questioned. When discrimination simply amounts to a woman systematically not being shown the same level of job opportunities as a man by automated job ads, discrimination amounts to the absence of an opportunity, rather than a denial of it. How can users be aware of and claim discrimination in such cases? This is especially concerning given that the AILD is meant to also help users in cases beyond strictly economic harm. While the AILD privileges transparency as a solution, it leaves a high degree of opacity on the potential victim's side.

6.4 The burden of evidence

The AILD's third objective is to make the compensation process easier and more effective by granting claimants the right to the disclosure of evidence about AI systems suspected of having caused harm. This right is conditional on the claimant presenting

> *'facts and evidence sufficient to support the plausibility of a claim for damages.'* (AILD, Art. 3 (1))

Further, claimants have the right of access only to evidence deemed *'necessary and proportionate to support a potential claim'* (AILD, Art. 3 (4)). Alongside the presumption

of causality, these conditions shift the problem from proving fault to accessing and interpreting evidence.

Consider, for example, '*plausible* evidence' (my italics), where the term plausible is left unspecified. Indeed, the term might be left intentionally vague to allow courts the flexibility to subjectively evaluate evidence during the AILD's application, fostering adaptability rather than rigid adherence to fixed rules. At the same time, it might be hard for a non-expert user to consider which amount or type of evidence can be considered plausible. On the one hand, this might discourage them from making a claim. On the other hand, this might leave room for a defendant with access to and knowledge of their AI system documentation to further deter them from doing so, arguing for its implausibility. This endangers the process and demands further regulatory guidance.

Moreover, the AILD specifies that courts shall only order defendants to disclose evidence *if* the claimant has made all proportionate attempts to gather it from the defendant (AILD, Art. 3 (2)). It also adds that the refusal of the defendant upon the request of the claimant - thus previous to the request of the court - should not trigger any presumption of non-compliance with the duties of care (AILD, Art. 17). These caveats, while preventing claim overload, presuppose not only a victim's awareness but also confidence and perseverance in front of the incurred harm. However, without such confidence or means to persevere, protections against reckless litigations might curtail the advantages of evidence disclosure and burden reversal.

Once the claimant has established their right to evidence disclosure, they must prove fault resulting from non-compliance with the AI Act's requirements and obligations. For example, a presumption of causality is triggered if a high-risk AI system (i) was not trained, validated, and tested on data sets that meet quality criteria (AI Act, 2021, Article 10(2) to (4)), (ii) does not meet transparency requirements regarding its design and

intended use (AI Act, 2021, Article 13), (iii) does not allow for effective human oversight (AI Act, 2021, Article 14), or (iv) does not achieve an appropriate level of accuracy, robustness and cybersecurity (AI Act, 2021, Article 15).

The AI Act specifies that information about AI systems should be *'relevant, accessible and comprehensible to users'* (AI Act, 2024, Article 13(2)). However, the AI Act's required documentation is likely to be both technical in nature and legal in form. Even when information about black-box algorithms is transparently disclosed, inadequacy in technical and legal literacy may make it difficult for claimants to evaluate defendants' compliance with rules (Nawaz, 2022). Furthermore, there are currently no standards regarding how the elements that the documentation should contain (e.g., quality criteria) ought to be measured across providers (Novelli et al., 2024). While these might be developed at future stages by the establishment of harmonised standards, this currently further complicates the ability to understand the evidence disclosed.

As mentioned, according to the AILD, the disclosure of evidence should be limited to what is *necessary* and *proportionate* to uphold a potential claim. This aims to balance claimants' rights with the protection of third parties' trade secrets and confidential national security information. Regarding such information, the proposal states that national courts are empowered to

> *'take specific measures necessary to preserve confidentiality when that evidence is used or referred to in legal proceedings.'* (AI Liability Directive, 2022, Article 3(4))

However, it opens the door to the possibility of withholding information under vague pretences of confidentiality, which is specifically relevant given the difficulty of establishing a necessary and proportionate threshold for evidence of discrimination before it is proven. This leaves room for abuse and adds to the burden of knowledge on the claimant, a burden of ignorance.

6.5 Recommendations

In light of the aforementioned concerns, I put forth four recommendations addressed to EU policymakers. These recommendations aim to reduce the information gap and enhance the effectiveness of the AILD. To meet the latter objective, recommendations 1 and 2 concentrate on broader facets of the directives' efficient execution and its integration with the other regulatory initiatives. To tackle the former issue, recommendations 3 and 4 directly confront the information gap, thereby positively impacting the liability gap given their interconnected nature.

1. *Establish a real holistic approach in defining the supranational legal framework of AI.*

As seen above, the proposed Directives heavily rely on the content of the AI Act. Considering that, the EU legislator should adopt a holistic approach, fostering the clarification and the correct integration of definitions and requirements enshrined in the AI Act with the PLD and AILD, promoting thus a clearer interaction between the three different pieces of legislation. For example, the PLD and the AI Act exhibit misalignment, as the former necessitates the disclosure of evidence for all AI systems, whereas the latter's proposal imposes record-keeping obligations solely for high-risk systems (AI Act, 2021, Article 12) (Hacker, 2022). Research has shown how promoting regulatory interoperability can lower regulatory burdens and enhance the effectiveness of regulatory initiatives (Roberts et al., 2023). In future iterations of the AILD, interoperability among these initiatives could be enhanced by further specifying the modalities relating to AI systems' documentation (e.g., how to disclose and document transparency requirements) to align liability's needs with those of risk management expressed in the AI Act.

2. *Establish targeted measures to promote coherence and complementarity in the EU-wide application of the AILD and PLD.*

As the full integration of the AILD and PLD seems unlikely (Bertuzzi, 2022; Hacker, 2022), clarificatory guidance and support should be provided to claimants, defendants and courts to harmonise their scope and implementation. For example, identifying contexts where it is advantageous to apply strict liability rules to AI systems, and contexts where it is more cost-effective to prevent harm by focusing on the quality of care of the provider. The AILD directive's next evaluation should thus consider promoting the complementarity of the two Directives through targeted reviews and revisions.

3. *Create incentives to develop and deploy AI systems that are less complex, less autonomous, and less opaque ex-ante, without over-relying on transparency requirements ex-post.*

Research shows that in high-stakes environments (e.g., medicine), inherently interpretable models should be preferred over black-box ones (Rudin, 2019). I recommend that EU policymakers set European-wide standards for the interpretability of high-risk models and incentivise their adoption by, for example, granting a presumption of conformity with the AI Act or simplifying the conformity procedures for providers who implement them.[49] Fundamentally, this strategy recommends that EU policymakers adopt a proactive stance towards accountability. This means implementing measures that anticipate and prevent failures before they occur, rather than adopting a reactive stance that focuses on redressing failures after they have happened (Novelli et al., 2023). This should help lower the entry barrier to understand evidence of discrimination for users and, in turn, facilitate the use of

[49] A general presumption of conformity with the AI Act given adherence to standards is already on the table (Roberts et al., 2023).

this evidence to inform claims for damages about algorithmic discrimination under the AILD.

4. *Strengthen the right of access to evidence for the claimant by incentivising notification of harm.*

As this analysis has shown, access to evidence might become ineffective when users are unaware of having been discriminated against. I recommend that EU policymakers introduce incentives for providers to share information about their high-risk AI systems with trusted third parties (e.g., through independent audits) (Falco et al., 2021; Mökander, 2023) to allow for the notification of specific groups about their higher likelihood of exposure to bias, discrimination or errors. External validation of models by trusted third parties can ensure the reproducibility of results and surface biases (Haibe-Kains et al., 2020). Incentives of this nature could encompass strategies such as facilitating partnerships between providers and reputable third-party entities (e.g., auditing agencies) to enable comprehensive disclosure of information - along with evidence - about adverse effects (Novelli et al., 2024). This point is pivotal because the effectiveness of disclosure mechanisms relies on this prerequisite: users must have a valid basis to suspect harm. Additionally, going back to the aim of this chapter, this can assist users in providing substantial details along with corroborating evidence to inform a potential claim for damages concerning algorithmic discrimination.

6.6 Conclusion

As this chapter has shown, the complexity, autonomy, and opacity of AI systems create challenges in ascribing liability for algorithmic discrimination, exacerbating the difficulty of proving fault. Furthermore, victims of discrimination may be unaware that AI systems have produced unfavorable outcomes, introducing an additional layer here termed the

'information gap'. The proposed European AI Liability Directive (AILD) represents a crucial stride in bridging the 'liability gap' associated with AI-induced harms. However, in this analysis, I argue that the information gap is primary and necessary to let evidence of algorithmic bias successfully inform the AILD, to eventually fill the accountability gap and achieve the AILD's regulatory objectives.

The current iteration of the AILD, though enhancing legal harmonisation, introduces heightened legal intricacies, tilts toward prioritising transparency and evidence disclosure over recognising harm and discrimination, and places a burden on claimants to navigate and comprehend the evidence presented by defendants. To fortify the AILD and effectively close both the 'liability gap' and the 'information gap', I recommend the four key improvements outlined above.

As mentioned at the beginning of the chapter, these improvements are based on an analysis of a proposal that, because of its nature, is subject to change. Additionally, this analysis serves as a deep dive into a specific regulatory initiative rather than a comprehensive examination of Europe's liability landscape. The considerations and recommendations presented in this chapter should thus be taken as direction pointers, rather than as points of arrival. By implementing these recommendations, however, the AILD can hopefully evolve into a more robust framework that not only facilitates liability claims for AI-induced harms but also empowers individuals to navigate the complexities of algorithmic discrimination, thereby fostering the possibility of bringing evidence of bias in support of algorithmic discrimination.

Chapter 7:

How can bias as evidence inform policy frameworks to tackle discrimination?

Chapter 7. Exploring Algorithmic Bias as Evidence in the California Racial Justice Act: Opportunities, Challenges, and Policy Implications

Abstract

In California, the use of risk prediction and predictive policing tools has raised concerns regarding their reinforcement of existing racial disparities in the Criminal Justice Syststem (CJS), especially through algorithmic bias. Despite regulatory efforts like Proposition 25, which aimed to limit the use of RAIs, their deployment persists across California's counties, alongside predictive policing algorithms in major cities. The California Racial Justice Act (CRJA) (2020) (AB 2542) emerges as a ground breaking initiative to actively address racial disparities within the state's CJS. By allowing statistical disparities as evidence of intentional discrimination, the CRJA challenges legal precedents and aims to rectify systemic racial biases ingrained in the justice system. This chapter conducts a qualitative analysis of the CRJA document (H. Morgan, 2022), exploring its potential to incorporate algorithmic bias as evidence to combat racial disparities in charges, convictions, and sentencing. I argue that algorithmic bias, as an inherent feature of these algorithms, represents an alternative form of evidence – as a 'trace' – that can complement more traditional sources used to build discrimination claims in courts. The findings offer insights to policymakers and stakeholders in the CJS, aiming to inform decisions and actions regarding the integration of algorithmic bias in the CRJA and in regulatory initiatives more widely.

Note: This chapter has been submitted as an article to the Big Data & Society Journal and it is currently being considered for review. The chapter has undergone minimal editing in order to ensure coherence with the thesis as a whole.

7.1 Introduction

In the US, risk prediction and predictive policing tools have come under scrutiny for their (re)production of existing racial disparities in the Criminal Justice System (CJS) (Angwin & Larson, 2016; Ensign et al., 2018; Richardson et al., 2019a) . In California, Proposition 25 prevented algorithms such as risk prediction tools from replacing cash bail (McGreevy, 2020). Still, risk prediction tools are being used in almost all counties in California (Movement Alliance Project & MediaJustice, 2019) and the deployment of predictive policing algorithms has been reported in Oakland (Eticas, 2021) and Los Angeles (LA) (Gates, 2018; Lapowsky, 2018), among others. By virtue of (re)producing systemic racial injustices (e.g., historical discrimination, redlining, among others), algorithmic bias in these tools simultaneously perpetrates and reflects existing societal disparities. As mentioned in the previous chapters, algorithmic bias refers to when

> '*predictive performance - however specified - unjustifiably differs across groups along social axes such as race, gender, and class.'* (Mitchell et al., 2021, p.1)

This bias may also be referred to as a 'skewed performance' along one of these social or demographic axes.

Amidst these challenges, the California Racial Justice Act (CRJA) (2020) (AB 2542) represents a landmark legislative response to the growing outcry against racial injustice within the criminal justice system (CJS) in the state. Motivated by a heightened awareness fuelled, in part, by the tragic death of George Floyd (McGreal et al., 2021), the CRJA stands as the most expansive among a trio[50] of major criminal justice reforms. Its

[50] AB 3070 is a jury selection reform which requires clear and convincing evidence that a preemptory strike was not related to a protected class. AB 2512 amends the death penalty disability statute to prevent anyone who satisfies the scientific definition of intellectual disability to be eligible for the death penalty.

declared aim is to address racial disparities embedded in the CJS actively; by focusing on implicit bias and systemic racial disparities.

Differently from existing legal approaches to racism in criminal proceedings, the CRJA takes systematic, racial disparities as discriminatory in themselves. Especially, it allows for statistical disparities as a valid source of evidence to prove the kind of intentional discrimination required for a constitutional violation (California Racial Justice Act, 2020). In doing so, the Act challenges the much-debated 1987 Supreme Court ruling in McClesky v. Kemp (McCleskey v. Kemp, 1987), which dismissed statistical variations as evidence of deliberate discrimination, particularly in the administration of the death penalty.

The Act focuses on four forms of racial discrimination: a) bias or racial animus towards the defendant, b) the use of racially discriminatory language during trial, c) racial disparities in charges and convictions, and d) racial disparities in sentencing. Taking algorithmic bias – and analyses thereof - as a form of statistical evidence, this chapter seeks to explore the criteria and considerations relevant to its potential use to inform the implementation of the CRJA, especially to address the third and fourth forms of racial discrimination identified by the Act – racial disparities in charges and convictions, and racial disparities in sentencing.

To achieve this, I conduct a qualitative analysis of the document (H. Morgan, 2022): a structured review of the CRJA text to offer insights into the latent opportunities (Section 7.2) and challenges (Section 7.3) inherent to this approach and propose recommendations (Section 7.4) to strike a sensible balance. The proximate goal is to provide policymakers and stakeholders in California with insights to inform decisions and actions related to the use of algorithmic bias evidence within the framework of the California Racial Justice Act. The ultimate aim is to spark interest and support the use of

algorithmic bias as a source of evidence in the other existing Racial Justice Acts[51] in the US, and in regulatory policies around bias and discrimination in general.

7.2 Opportunities to use algorithmic bias as evidence

Among the discerned opportunities, data on algorithmic bias can help complement analyses of discriminatory trends by offering information that is more conducive to statistical analysis when contrasted with individual, text-based court or police reports. Analyses of algorithmic bias are also likely to encompass information on sensitive characteristics or proxies thereof, which can help prove discrimination. Moreover, methods utilised in addressing algorithmic fairness, such as counterfactual approaches, can prove instrumental in substantiating evidence for 'similarly situated' individuals as mandated by the law to demonstrate racial discrimination. Algorithmic bias, as an intrinsic aspect of an algorithm, can also provide mechanisms to mitigate the direct risk of external manipulation by the accused parties. Below, I elaborate on each of the above opportunities.

7.2.1 Evidence of Systematic Disparities

To qualify for a hearing under the Civil Rights Justice Act (CRJA), a defendant must initially establish a prima facie[52] case of discrimination. At this juncture, the evidence presented, ranging from testimonies to articles and statistical reports, should sufficiently demonstrate a 'substantial likelihood'[53] of a violation of the Act. Upon successful establishment of the prima facie case, the court proceeds with a hearing. During the hearing, the defendant may file a motion requesting the disclosure of all pertinent evidence,

[51] The only other two existing Racial Justice Acts are in North Carolina and Kentucky. The one in North Carolina, however, has been repealed in 2013. The CRJA is much more expansive than the other acts in its scope and requirements for evidence. More info on the other acts can be found in (Lowe, 2022)

[52] In both civil and criminal law, the term is used to denote that, upon initial examination, a legal claim has sufficient evidence to proceed to trial or judgment (Kenton, 2021).

[53] 'Substantial likelihood' is defined as 'more than a mere possibility, but less than a standard of more likely than not'–that the Racial Justice Act has been violated (California racial Justice Act, 2020, subd. (h)(2).)

including statistical evidence, relevant to a potential violation and under the possession or control of the state to substantiate claims of racial discrimination (California Racial Justice Act, 2020, subd. (d)). While it might be difficult to extract systematic patterns of discrimination from the type of data offered by courts, analyses of algorithmic bias in risk prediction and predictive policing tools might help reconstruct patterns of racial disparities.

The data that is typically in the hands of the state is likely to be administrative records or reports (e.g., police reports, lists of court cases, data on charges and sentencing enhancements, etc.). These documents are often text-based and unstructured and, when in digital format, are likely to be on a thumb drive with gigabytes of data (Weld, 2021). Additionally, in the absence of stringent reporting requirements (Zahn, 2021), these reports might be unkept or missing relevant data to prove discrimination (Knox et al., 2020). For example, in reporting a 'spotting operation', police might write that they caught someone because they have an 'orange jacket' or because 'they know' they are a criminal or that they are part of a gang. In turn, this might complicate the process of proving discrimination.

While results from risk prediction and predictive policing tools are often used as 'evidence' to evaluate people, the bias they contain could conversely be used toward evidence of their discrimination. Some attorneys in California (Weld, 2021) have suggested that the mere use of such tools in a jurisdiction can suggest racism. Additionally, most of the risk prediction tools in use in California counties participate in the Pretrial Pilot Program[54]. This program requires the publication of periodic reports including specific information regarding the tool, such as information on country-specific data used in it. While full datasets are not released, it reports which features are used in which county. It has been shown that some of the risk prediction tools used in counties in

[54] https://www.courts.ca.gov/sb36.htm

California (e.g., ORAS-PAT, PSA, VPRAI) rely on variables beyond criminal history, and include some demographic information which can indirectly be linked to race (e.g., housing status, history of substance abuse, and employment history) (Mapping Pretrial Injustice, 2019). As these have been argued to act as proxies for racial biases (Datta et al., 2017; B. E. Harcourt, 2010), these reports can provide a site to further investigate whether racial disparities are present.

Furthermore, according to the law, a defendant's history of convictions may constitute a pertinent factor in determining the severity of charges, convictions, or sentences in evaluating discrimination claims. Especially, it states that if

> '*the defense produces evidence that the conviction history may have been impacted by racial profiling or historical patterns of racially biased policing, the court shall consider the evidence.'* (California Racial Justice Act, 2020, subd. (h)(1))

Police reports or other forms of 'classic' evidence might be useful to capture cases of racial profiling where race-related language (e.g., racial slurs) is present. The Act, for example, provides a novel approach to addressing police racism as, in some cases[55], it includes racist statements and conduct by police officers among their prohibitions (Lowe, 2022). However, research shows how police administrative data fail to capture the invisible selective process through which police come to engage civilians (Knox et al., 2020), a process that prior work shows to be a function of a civilian's race (Gelman et al., 2007). It is important to capture police activity that seems not to be explicitly targeted toward one race and yet, it systematically is. Algorithmic bias, by tracking the systematic disadvantage imposed on racialised groups, can help address that. The Pretrial Pilot Program, for example, requires risk prediction tools tools not to use race or gender in making their inferences. Yet, it requires these tools to be periodically validated and tested

[55] Racial Offenses type a) and b)

for racial and gender bias in their results. This can provide publicly available checks of disparate impact on racialised groups.

7.2.2 Counterfactual evidence

To build their claim, the defense must show a significant difference [56] in charges, convictions or sentences between people of different races, comparing individuals who have engaged in similar conduct. The comparison considers individuals who are similarly situated based on legally relevant factors and within the same county (California Racial Justice Act, 2020). I argue that the criteria required to conduct this comparison are fitting to what fairness metrics for algorithmic bias can offer, specifically counterfactual approaches to algorithmic fairness.

According to counterfactual fairness, a decision is fair if, for a particular individual or group, its prediction in the real world is the same as that in a counterfactual world where the individual(s) had belonged to a different demographic group (Kusner et al., 2017; Wu et al., 2019). Counterfactual methods can allow one to check whether a protected characteristic (e.g., race or gender) was the cause of a specific discriminatory outcome for two individuals who are similar according to factors that are relevant in charging and sentencing but differ in their race. Such an approach could be used, for example, to explore whether individuals who are similar along criminal record and other 'relevant factors' according to the California Rules of Court[57] (e.g., nature of the offense) but for their race, were generally assigned a 'higher' risk level by algorithms.

---

[56] The 'statistical significance' of the data is a factor the court may consider but is not necessary to establish a significant difference (Hurtado & Yan, 2023).

[57] A full list of relevant factors according to the California Rules of Court can be found here: https://www.courts.ca.gov/cms/rules/index.cfm?title=four&linkid=rule4_421#:~:text=The%20California%20Supreme%20Court%20has,3d%20410%2C%20425.

As presented in Chapter 4, some counterfactual methods can also be used to develop counterfactual explanations, which can explain why that outcome, rather than an opposite one happened (Karimi et al., 2021). These methods can further support the defendant by providing an accessible form of evidence in support of their claim, one that can help to bring *'cold numbers convincingly to life'* (Zahn, 2021, p.231). Furthermore, they can also be used to provide recommendations on how to change the output (Karimi et al., 2021; Ziosi et al., 2023). Thus, they could also help courts to reason on or consider practical changes to actively combat racism in their system.

While counterfactual approaches can help compare individuals who have engaged in similar conduct, alternative fairness metrics can help show a difference in charges, convictions or sentences across races. For example, False Positive and False Negative rates can offer insights into statistical disparities across demographics across demographics. Metrics for demographic parity can help assess whether racialised groups are classified as 'high-risk' at similar rates[58]. Additionally, given that there is currently only little case law on the subject of the CRJA (Zahn, 2021), reference to recognised metrics can provide a preliminary standard backed by extensive research on so-called 'algorithmic fairness' (See e.g., Barocas et al., 2022; Berk et al., 2017; Calders & Verwer, 2010; Dwork et al., 2011; Feldman et al., 2015; Hajian & Domingo-Ferrer, 2013; Kamiran & Calders, 2012).

7.2.3 Less 'biased' evidence

The Act mandates the disclosure of relevant evidence by the court or the police (California Racial Justice Act, 2020). However, it does not mandate that this evidence be compliant with certain standards of, for example, completeness or validity (California Racial Justice Act, 2020). Additionally, it is not clear that the state is in any way compelled to collect,

[58] For a full list and explanation of metrics see (Barocas et al., 2022)

synthetise or publish data (Zahn, 2021). This introduces the risk that data is manipulated or missing. I posit that algorithmic bias, as an inherent feature of algorithms, represents an alternative form of evidence – as a 'trace' - and offers mechanisms to mitigate the risk of manipulation.

Queally (2018) reported that California police departments uphold less than one in ten police misconduct file complaints. Additionally, records may be intentionally destroyed or altered by those responsible for their disclosure, such as law enforcement agencies or the judiciary (Richardson et al., 2019b). For instance, in 2018, Senate Bill 1421 was enacted to make police disciplinary records accessible to the public (California Senate Bill 1421, 2018). Before the law's effective date in January, police departments destroyed their records (Bond Graham, 2019). Instances have also been documented where law enforcement deliberately misreported individuals' race during police traffic stops in Los Angeles and San Francisco (Barba & Baustin, 2023).

Risk prediction and predictive policing tools draw from a variety of evidence (Richardson et al., 2019b); not only do they rely on a variety of variables, but they also pool together different public records from multiple agencies (Mapping Pretrial Injustice, 2019; Rotaru et al., 2022), This type of evidence is allowed by the Act which, together with statistical evidence, allows for any other form of aggregate data (California Racial Justice Act, 2020). Arguably, this pooling of data is likely to reduce the effect of misreporting from an individual source. Additionally, the accuracy of these tools is dependent upon the quality of the data (Gomede, 2024), thereby diminishing the incentive to compromise it. This minimises the risk that an altered police record compromises the results.

Moreover, many of these tools are proprietary (Carlson, 2017; Elizabeth, 2022). This poses ethical challenges in their deployment in the public sector (McKay, 2020).

However, it also creates a separation between the system provider and the user, potentially mitigating the risk of relevant evidence regarding biases being directly manipulated by parties seeking to conceal their actions. Given the generally adversarial character of US law (Benefiel, 2022), this might help navigate the complex power dynamics that ensue when challenging an institution, as it introduces a middleman between the institution and the defendant. By virtue of offering a service to these institutions, however, relying on private companies might provide a cover rather than a solution to this conflict of interest.

At the same time, as it has been mentioned in section 7.2.1, the Pretrial Pilot Program requires the Judicial Council to publish publicly funded periodic reports on gender and racial bias in the risk prediction tools in use. Even though most validation studies and analyses are conducted within institutions, the public nature of their funding increases the value of accountability requests. For example, even in cases where these studies periodically report a suspicious absence of bias, members of the public could be more easily entitled to receive more information on it under the Freedom of Request Information Act (FOIA). Even though not a solution, publicly funded validation reports might represent a more reliable and neutral ground between the two parties and reduce potential conflicts of interest.

7.3 Challenges of using bias as statistical evidence

While statistical evidence from algorithms might help provide more structured data, there are open questions regarding data access, opacity and lack of AI experts to do so. Beyond that, the evidence they provide might not fit the conventional canons of evidence to which courts are used or might offer evidence beyond the county level, which would be inapplicable. Finally, evidence from algorithmic bias might help accused actors defer their responsibility to the ‘error’ or ‘bias’ of the machine, rather than the CJS.

7.3.1 Access, opacity, and complexity of AI

As previously noted, several risk prediction and predictive policing tools are proprietary, a characteristic that may alleviate certain aspects but also complicates the accessibility of evidence (Carlson, 2017; Elizabeth, 2022). Differently from the Pretrial Pilot Program involving public reports on RAIs, for example, the cases of predictive policing tools reported in California were linked to private companies such as Palantir (Gates, 2018). This complicates access to evidence, and further challenges attempts to assess bias in the case of opaque and complex AI systems.

The CRJA allows a defendant to obtain access to the evidence in the possession or control of the state [59] (California Racial Justice Act, 2020). As what counts as in 'possession' and 'control' remains vague, it seems unlikely that the law would be sufficient for private data to be disclosed. Additionally, even where access was to be granted, standard technical or legal practices such as data anonymisation or omission of protected categories might impact the ability to reliably assess racial disparities. For example, Xu & Zhang (2020) show how some techniques for data anonymisation tend to produce more False Positives than False Negatives and thus, impact the identifiability of statistical disparities. This might complicate the process of finding 'similarly situated individuals' which is necessary to build a claim.

Furthermore, predictive policing tools are becoming increasingly sophisticated, with some employing techniques based on neural networks (Flaxman et al., 2019; Kang & Kang, 2017; Mohler & Porter, 2018; Rotaru et al., 2022). Even when algorithmic bias can be estimated, discerning the degree to which it signifies evidence of racial disparities versus technical biases (e.g., sampling problems) may pose challenges. As Chapter 5 has

[59] The Act takes the 'State' to mean: 'Attorney General, a district attorney, or a city prosecutor.' (California Racial Justice Act, 2020, subd. (h)(4))

shown, however, recent studies have expressed optimism regarding this prospect by estimating police enforcement bias within a complex predictive policing tool developed at the University of Chicago (Rotaru et al., 2022). Nevertheless, the tool has not yet been implemented in a real-world context, leaving uncertainties regarding the computational and economic feasibility of such an endeavor in practice.

7.3.2 AI expertise, convenience & scope

Not only the CRJA introduces an innovative legal approach to discrimination but also a novel approach to what counts as evidence of discrimination (California Racial Justice Act, 2020). Given the lack of resources and lack of technical training among public defenders (Weld, 2021) – who handle 80% of felony cases in California (Tafoya & Nguyen, 2015) - reliance on algorithmic bias might exacerbate existing challenges in the application of the new law (Hurtado & Yan, 2023) and further undermine the feasibility of such an approach. Especially, algorithmic bias may present an unconventional and thus, unattractive and challenging option for them.

The role of statistical evidence in US courtrooms has been an object of debate (Gastwirth, 2020), let alone that of algorithmic bias. In the case *Gill v. Whitford* (2018), several Supreme Court judges expressed apprehension about the use of a statistical metric to measure a map's bipartisan bias in establishing evidence of gerrymandering (Roeder, 2017). Some voiced doubts about the reliability of such complex calculations, while others expressed concerns about the public understanding of such metrics fearing that it could undermine the legitimacy of the court. As mentioned in the introduction, in *McCleskey v. Kemp* (1987), the court also dismissed evidence of state-wide racial disparities in Georgia

capital sentencing[60] as insufficient to prove discrimination without proof of intent (Zahn, 2021). Notwithstanding CRJA's attempt to tackle precisely this problem, defendants may prioritise more conventional forms of evidence, especially ones more likely to express discrimination more directly.

The Judicial Council of California Criminal Jury draws an important distinction between 'direct evidence', which can prove a fact by itself, and 'circumstantial evidence'. which contributes to proving a fact only indirectly and which needs additional sources of evidence (Judicial Council of California, 2023). Direct testimonies or video recordings from police body-worn cameras are more likely to be perceived by courts as a form of direct evidence than statistical reports. Arguably, in being less decisive, indirect evidence can also be more easily manipulated by the respondent in their favor. In *Finley vs Superior Court* (2023), one of the few CRJA cases existing at the time of writing, the concept of 'high-crime area' was used both by the defendant to claim racial discrimination, as well as by the police officer as an appropriate consideration in determining the reasonableness of his investigative detention following a traffic stop.

Furthermore, the CRJA mandates that disparities must be demonstrated at the county level, rather than statewide (California Racial Justice Act, 2020). Even though risk prediction tools in California are validated on county populations, many of them are trained on extensive datasets sourced from other US states or counties. The Virginia Pretrial Risk Assessment Instrument (VPRAI-R) has been used in at least 14 counties in California and the Ohio Risk Assessment Tool (ORAS-PAT) in at least 20 (Mapping Pretrial Injustice, 2019.). In such instances, algorithmic bias could only serve as evidence of disparities in deployment—highlighting how the algorithm may discriminate against

[60] The research showed that black defendants in Georgia were 4 times more likely than their white counterparts to be sentenced to death.

certain racialised groups in its outputs and how this discrimination influences the decisions made by law enforcement or judicial authorities. However, it cannot serve as evidence of existing or past racial disparities within the specific county under scrutiny.

7.3.3 The burden of responsibility

Despite extensive reporting on racial bias in risk prediction tools (Angwin & Larson, 2016; Berk et al., 2017; B. E. Harcourt, 2010) and predictive policing tools (Ensign et al., 2018; Richardson et al., 2019a; Rotaru et al., 2022), research indicates that some judges, for instance, prefer not to utilise them or even ignore them (Pruss, 2023). Moreover, even when these tools are considered, it is challenging to establish whether and how their outputs influence judicial decisions (Skeem et al., 2020). Positive decision bias may lead a judge to place undue trust in algorithmic outputs (Gravett, 2024), whereas averse attitudes toward technology might prompt them to diverge from algorithmic suggestions (Fine et al., 2023). Furthermore, the notion that algorithmic bias merely influences rather than dictates decision-making may undermine the value of such evidence, leading to a reliance on more precise or relevant forms of evidence regarding racial disparities, such as county-level, single-prosecutor statistical studies (Zahn, 2021).

Conversely, relying on algorithmic bias as evidentiary support may incentivise judges or law enforcement officials to shift responsibility onto AI systems or the liable providers of such tools. They may argue that racial disparities within their courts or police departments are attributable to the AI system, thus absolving themselves of responsibility. Research shows how AI systems can act as ‘moral buffers’ with humans relinquishing their responsibility over their decisions (Cummings, 2006). This issue is particularly pertinent given that if a disparity is substantiated, the prosecution has the option to account for it using race-neutral justifications (California Racial Justice Act, 2020). Algorithmic bias may provide a convenient scapegoat in this regard.

However, the law's focus on addressing systemic racial disparities suggests that its objective is not primarily to penalise individual actors but rather to rectify broader injustices within the system. The Act explicitly states,

> '*The intent of the Legislature is not to punish this type of bias, but rather to remedy the harm to the defendant's case and to the integrity of the judicial system.*' (California Racial Justice Act, 2020, Subd. (2)(i))

If successful, the Act offers remedies both at the trial court level (e.g., declare a mistrial, dismiss enhancements, reduce charges) and post-conviction (e.g., vacate conviction, resentence). While these measures may reflect poorly on the involved courts or police departments, they do not necessarily target them directly and thus, they lower the weight of their potential burden of responsibility. Hence, these considerations do not entirely preclude the use of algorithmic bias for such purposes.

7.4 Implications and policy recommendations

The CRJA allows for an expansive reading of evidence of systematic racism (Lowe, 2022). This opens the door for sources of evidence that shed light on racism in the CJS more in general and thus, for a strategic use of algorithmic bias. However, whether we should or is convenient to do so depends on how the opportunities and challenges above will be balanced. The following policy recommendations aim to guide California policymakers and stakeholders in the CJS to the integration of algorithmic bias as evidence, focusing on its strategic use and potential impact on systemic reforms.

1. *Implement Targeted Measures to Enhance Transparency and Accountability*

In order to fulfill the objectives of the CRJA effectively, it is imperative to establish targeted measures aimed at incentivising reporting requirements while prioritising accessibility and transparency of information. To achieve this, I propose the following policy recommendations:

a. *Address Challenges in Expertise and Access:*

Recognise the necessity of technical skills and tools to support defendants in legal proceedings. Allocate resources and strategic planning to overcome challenges such as the scarcity of technical experts, data unavailability, and accessibility barriers. For instance, initiatives like tech training for public defenders can enhance vigilance regarding the inaccurate and unfair use of technologies (Weld, 2021). Most importantly, it can also help turn their discriminatory application in support of the discriminated. For example, work by Abebe et al. (2022) proposes the introduction of adversarial testing techniques in court to support the defense in scrutinising and cross-examining the validity of evidence from statistical software (such as RAIs) utilised by the CJS to incriminate people. This can help assess the role of algorithmic bias in court decisions and enhance the accountability of courts.

b. Enhance Transparency and Accountability through Reporting Requirements:

Ensure facilitated access to data on algorithmic bias by incentivising reporting requirements tied to algorithm deployment with an eye for ensuring validity and accessibility of evidence. Where tools are proprietary, implement measures such as mandating the disclosure of third-party audit results for AI developers or deployers (Falco et al., 2021; Mökander, 2023). Regarding the creation of incentives, examples from cities like Amsterdam [61] illustrate the effectiveness of algorithm registries in enhancing algorithmic transparency. In cases where direct access to data is not feasible through courts, utilise mechanisms like the Freedom of Information Act (FOIA) requests to obtain information from additional sources like fusion centers in California[62].

[61] https://algoritmeregister.amsterdam.nl/en/ai-register/

[62] https://ncric.ca.gov/ncric/#:~:text=The%20NCRIC%20was%20established%20by,sharing%20nexus%20in%20Northern%20California.

2. *Strategic integration of algorithmic evidence with conventional forms of evidence*

To effectively leverage evidence in legal proceedings, it is crucial to strategically complement traditional forms of evidence with 'algorithmic' evidence. I propose to:

a. *Utilise Evidence Sources in a Complementary Fashion:*

The CRJA allows for aggregate data and multiple forms of evidence to build a claim. Incorporate anecdotal and expert witness declarations with statistical information to construct comprehensive case narratives (Zahn, 2021). In instances where conventional reports are inadequate or compromised, insights from algorithmic reports can enhance understanding of discrimination. As previously mentioned, public reports from risk prediction tools in California can provide valuable data on the features employed in assessment and on the potential disparate impact of these tools, supplementing traditional evidence sources. Furthermore, the application of counterfactual fairness metrics can help to establish evidence of 'similarly situated' individuals. Fairness metrics measuring error rates or positive prediction rates can help assess specific aspects in the frequency of convictions, charges, or sentencing disparities.

b. Timing Considerations:

Consider the relevance of algorithmic bias in different stages of legal proceedings. As mentioned by Weld (2021), the mere use of risk prediction and predictive policing tools can raise concerns about racism. Reports on their use could be leveraged to build a prima facie case under the CRJA. Classic forms of evidence such as testimonies or video recordings can play a significant role when the CRJA is invoked during a present trial, where there is a need for immediate evidence on the present case. Algorithmic bias might be more relevant when the CRJA is invoked on past cases, where past evidence might be lacking or incomplete, and algorithmic bias can supply to that lack by shedding light on past discriminatory trends. Specifically, in cases where risk prediction and predictive

policing significantly influence decisions, algorithmic bias should be admissible as evidence.

c. Geographical considerations:

Under the CRJA, racial disparities must be shown at the county level. Zahn (2021) showed how county-level statistics can be useful in building more effective equal protection claims in claiming for racial discrimination in death sentence cases. Similarly, when risk prediction tools are validated in a specific county in California (Movement Alliance Project & MediaJustice, 2019), evidence of their use of proxies for race or analyses of their biases can be directly leveraged to build a case. When these tools were not validated on county-specific information or not validated at all, proof of their disregard for accuracy standards or potential constitutional restrictions coupled with reports on their biases (e.g., disparate impact) could be brought as evidence to the case.

3. Driving Systemic Reforms Through Algorithmic Bias:

Remedies for racial disparities allow the court to look to other available remedies under the California Penal Code (California Racial Justice Act, 2020). Notwithstanding this, the offered remedies mostly include individual rather than systemic changes. To address racial disparities and promote broader systemic reforms within the California CJS, I propose the following recommendations:

a. *Structural Diagnosis:* The CRJA introduces important individual measures such as vacating a conviction, resentencing or declaring a mistrial. However, for the Act to live up to its aim to actively tackle systemic racial bias in California's criminal legal system, emphasis should also be placed on identifying and addressing deeper structural issues in the CJS. Algorithmic bias, by reproducing systemic racial disparities, could serve as evidence to substantiate these systemic issues, empowering public prosecutors to better understand and point courts toward these

problems. As mentioned in Chapter 3, metrics for algorithmic bias can serve important diagnostic functions to gather evidence on racial disparities, as well as formalise differential opportunities for fairness interventions.

b. *Informing Policymaking:* Utilise evidence of algorithmic bias to guide discussions and decisions regarding systemic changes in justice and police reforms. Scholarly proposals advocate for interventions that extend beyond the CJS itself, such as new federal legislation to limit the power of local police (Hasbrouck, 2020) and shifting power dynamics towards policed populations (Simonson, 2021). For the present case, data revealing the perpetuation of historical redlining patterns through predictive policing should prompt a reassessment of law enforcement strategies. This reassessment should prioritise community support and healing initiatives over punitive measures.

7.5 Concluding remarks

The analysis and recommendations here presented aspire to facilitate a nuanced and strategic deployment of algorithmic bias evidence within the criminal justice system in California, thereby paving the way for profound systemic reforms and enhanced fairness. It is important to note, however, that the considerations outlined in this chapter are not exhaustive nor definitive. As discussed in Chapter 4, there exist significant technical and conceptual limitations to the practical application of algorithmic bias and technical tools in addressing and representing real-world discrimination. Furthermore, the CRJA is still in its early stages, with minimal case law to draw upon and no reported successful cases as of the time of writing.

Looking ahead, I hope that future research endeavours and the evolution of legal precedents will contribute to refining the contours of the considerations presented here. By remaining attentive to emerging insights and developments, policymakers and stakeholders

can continue to refine their approaches, ensuring that the use of algorithmic bias evidence remains informed, effective, and aligned with the pursuit of justice and equity within the criminal justice system.

Chapter 8:

Conclusion

8. Conclusion

The overarching question guiding this thesis research was, '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*' This thesis argued that algorithmic bias in risk prediction tools, presented by the literature as a problem concerning its disparate effects on different societal groups in areas as varied as credit, health, and criminal justice, can also be conceived of as a carrier of evidence of existing societal disparities. Given the presence of unfair differences in the way risk prediction tools perform towards different demographic groups, I enquired whether I could qualify and quantify these differences in performance in a way that is informative for policy frameworks around bias and discrimination. This did not entail testing for bias, but rather inquiring whether I could *conceive*, *measure,* and *use* these unfair differences in performance to *inform* policies that address existing societal disparities. Consequently, this research unfolded across four distinct levels of inquiry: a *conceptual*, a *methodological*, an *applied*, and a *policy* level.

On a conceptual level, sub-question 1 asked, *'How can bias as evidence be conceptualised?'* On the methodological level, sub-question 2 asked, *'How can bias as evidence be measured?'* On the applied level, sub-question 3 inquired, *'How can bias as evidence be applied in practice?* Finally, on the policy level, sub-question 4 asked, *'How can bias as evidence be used to inform policy frameworks to tackle discrimination?'* The sub-research questions above remain overly broad. For this reason, they were asked within the framework of an integrated thesis, where each chapter was an inquiry into a specific, yet relevant topic for the research question. In order, the chapters dealt with algorithmic bias as evidence (Chapter 3), a set of proposed tools to measure it (Chapter 4), its application in the context of predictive policing (Chapter 5), and its possible integration

into specific policy frameworks to inform claims to tackle discrimination (Chapters 6 and 7).

The next paragraphs will present a synthesis of contributions and findings from each chapter. Following, I will elaborate on the overall implications of these findings for the main research question, as well as on the specific lessons for the distinct target audiences of this thesis drawn from the specific chapters. As anticipated in Chapter 1, these lessons will take the form of a series of policy recommendations tailored to researchers, practitioners and policymakers working at the intersection of algorithmic fairness and policymaking. I will then outline the limitations of the research and conclude with some concluding remarks.

8.1 Synthesis of contributions and findings

Each chapter of the thesis contributed distinct insights toward answering the respective sub-questions. Especially, their findings substantiated the contributions set out at the beginning of this thesis. Below, I summarise the contributions and findings of each section.

Conceptually, a shift from framing bias as a problem to framing it as evidence of problems (e.g., societal disparities) can re-orientate some efforts of the research community from mitigating its effects, towards looking for the conditions that give rise to it. Toward that end, Chapter 3 brought a 'thick' conception (Williams, 1985) of bias: one that ascribes both a descriptive and a prescriptive aspect to algorithmic bias. It argued that one should preserve the evidence of societal disparities carried in algorithmic bias and, through this evidence, inform interventions to address these disparities. It showed how bias preserving metrics, such as calibration, can be used as a diagnostic tool to measure societal disparities and gain a deeper understanding of them. Additionally, in combination with competing metrics, calibration can help reframe trade-offs between fairness metrics as

different opportunities for fairness. On the one hand, the disagreement between two metrics can help formalise the different opportunities they respectively offer, and to help justify different choices regarding bias mitigation. On the other hand, it can work as evidence of the multiple, irreconcilable costs involved for marginalised populations and thus, as an argument against the use of risk tools. These informative differences, captured through bias preserving metrics, could be used to inform the application of bias transforming metrics toward more equitable, rather than strictly equal, fairness interventions. Still, the practical legal aspects of these propositions should be informed by their specific context and area of application. Overall, these efforts can contribute to addressing the root causes of unfair predictions, rather than solely tackling their consequences.

Methodologically, this shift in framing algorithmic bias from just a 'problem' to also 'evidence' can prompt a redirection of bias detection tools toward uncovering its root causes and thus, provide insights into its origins and mechanisms. To that end, Chapter 4 offered an analytical framework for XAI tools to approach algorithmic bias 'genealogically'. This approach encourages the explanation of bias based on the conditions leading to its occurrence, rather than focusing on its consequences. It advocated understanding bias not as a result of a single cause but as an accumulation of evidence related to the conditions that give rise to it. The chapter demonstrated how observational XAI approaches, such as marginal and conditional Shapley values, enable the assessment of whether a protected characteristic, such as race or gender, is unfairly linked to an outcome. This holds relevance for ensuring procedural fairness in decision-making processes (Grgić-Hlača et al., 2018), particularly in tasks like ethics-based auditing (EBA) (Mökander et al., 2021b; Mokander & Floridi, 2021). However, given their assumption about feature independence, I suggested that observational approaches should not be used

to explain or determine which role a protected characteristic played in a discriminatory outcome. Instead, I proposed referring to explanations derived from these approaches as 'evidential observations', which may prompt further investigations into the conditions leading to discriminatory outcomes.

The chapter also contended that both interventional and counterfactual XAI approaches can contribute to enhancing the agency of discriminated users. Interventional approaches can provide consequential recommendations to improve the user's status. However, instead of suggesting changes for the user to improve their discriminated position, I claimed that these approaches should focus on developing constructive recommendations for developers on how to intervene to prevent algorithmic discrimination in the first place. Counterfactual approaches can be useful for seeking recourse in cases of algorithmic discrimination. Still, rather than determining which feature caused a discriminatory outcome, they should be employed to shed light on the conditions constituting discrimination; namely, to name the conditions by which similar individuals are treated differently on the basis of protected characteristics. Going back to the goal of this chapter, these efforts were aimed at approaching bias *ex ante* by focusing on the conditions that constitute it and thus, by providing insights into its origins and mechanisms.

At the *applied* level, trying to capture what bias stands as evidence of could help rethink the skew in predictive performance of risk prediction tools as a mirror of historical conditions of oppression (Kleinberg et al., 2017; Mayson, 2019), rather than just a future predicament. Accordingly, Chapter 5 offered an analysis that helped recognise whether the evidence of algorithmic bias in a crime prediction tool motivated criminal justice interventions that reinforced or challenged existing power structures. This contributed to seeing themes around data-driven policies as mirroring long-standing tensions and debates in contexts such as policing. Through interviews with Chicago-area academic researchers,

civil society organisations, and public sector employees, I found that evidence of algorithmic bias has the potential to challenge as much as to affirm the status quo. I showed that stakeholders from different groups articulate different problem diagnoses of the algorithm's bias, strategically using these diagnoses to advance criminal justice interventions that align with individuals' political ends. These strategies include using data to reform the police that generated the biased data in the first place, such as by allocating police patrols more equitably; reaffirm or strengthen existing models of policing; reject the use of predictive policing tools; and reframe the narrative away from reforms like data-driven policing toward abolitionist aims, such as repairing communities. Overall, the findings emphasised the importance of centering the needs and experiential understanding of justice-impacted communities through community-based participatory research in designing and evaluating responses to algorithmic bias.

Finally, if what bias stands as evidence of was to *inform policy*, this could broaden the scope of policies around AI from how to best automate solutions, toward how to tackle existing societal disparities and power asymmetries (Abebe et al., 2020). Chapter 6 provided strategies to alleviate the burden of evidence on end-users for proving algorithmic discrimination under the EU AI Liability Directive (AILD). The AILD seeks to close the 'liability gap' related to harms caused by AI systems. The complexity, autonomy, and opacity of AI systems pose challenges in attributing liability for algorithmic discrimination, complicating the proof of fault. Concurrently, victims of discrimination may not be aware that AI systems have generated unfavorable outcomes, contributing to an additional layer known as the 'information gap'. My analysis exposed significant deficiencies, particularly in addressing this critical latter gap. The current iteration of the AILD, though enhancing legal harmonisation, introduces heightened legal intricacies, tilts towards prioritising transparency and evidence disclosure over notification

of harm and discrimination, and places a burden on claimants to navigate and comprehend the technical or general evidence presented by defendants. To fortify the AILD and effectively close both the 'liability gap' and the 'information gap', I recommended some key improvements, among which the creation of incentives to develop and deploy AI systems that are less complex, less autonomous, and less opaque ex-ante and strengthening the right of access to evidence for the claimant by incentivising notification of harm. By implementing these recommendations, the AILD can evolve into a more robust framework that not only facilitates liability claims for AI-induced harms but also empowers individuals to navigate the knowledge asymmetries entailed in instances of algorithmic discrimination, thereby fostering a more just and accountable AI landscape.

Chapter 7 shed light on the opportunities and challenges of using evidence of algorithmic bias to prove the existence of historical, systemic racial disparities in the CJS through the California Racial Justice Act. Among the discerned opportunities lies the prospect of algorithmic bias providing structured data that is more amenable to analysing structural discriminatory trends compared to individual, text-based police reports. This data is also likely to contain information on sensitive characteristics or proxies thereof. Furthermore, metrics employed in addressing algorithmic fairness, such as counterfactual approaches, can prove valuable in establishing evidence for ‘similarly situated’ individuals as required by the law to prove racial discrimination. Algorithmic bias, as an inherent feature of an algorithm, also offers mechanisms to minimise the direct risk of external manipulation by the accused actors. Notwithstanding its advantages, however, there are open questions regarding data access, opacity, and lack of AI experts posing challenges to its effective use. Beyond that, the evidence that algorithmic bias can provide might not fit the conventional canons of evidence to which courts are used or might offer evidence beyond the county level, which would be inapplicable according to the Act. Finally,

evidence from algorithmic bias might help accused actors defer their responsibility to the 'error' or 'bias' of the machine, rather than systemic racial disparities in the CJS. Overall, the chapter highlighted criteria and considerations for policymakers to leverage the use of algorithmic bias as evidence to inform remedies to structural racism throughout the CJS in California.

8.2 Implications and policy recommendations

The findings derived from the specific chapters represent distinct answers to the thesis' sub-research questions. At the same time, they collectively contribute to addressing the overarching research question. In the next paragraphs, I will outline the overarching implications for the thesis and provide a summary of the policy recommendations articulated in each chapter, tailored to their respective target audience.

8.2.1 Overall Implications

After having considered the specific contributions and findings of each chapter, this section outlines the overarching implications of their results for the main research question, '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*'

1. *Algorithmic bias, as a carrier of evidence, can help develop hypotheses on societal disparities.*

Algorithmic bias can provide a signal or a piece of evidence to develop hypotheses on the presence of societal disparities in specific contexts.

To do so, established approaches to algorithmic bias such as fairness metrics and explainability methods should be re-evaluated in their ability to gather evidence – note, not proof – of societal disparities. Regarding fairness metrics, one should reconfigure their use within a paradigm of substantive fairness (Green, 2021b); namely, by shifting their goal

from achieving 'fair(er)' predictions toward supporting a broader set of interventions to address existing disparities. As previously mentioned, Green (2023) has drawn an important difference between formal algorithmic fairness, where mathematical representations of fairness are applied to algorithmic decision-making, and substantive algorithmic fairness, where algorithms are incorporated into broader agendas for equity reform. Of course, fairness metrics alone cannot provide reliable evidence or proof of injustices. However, research has shown how metrics in general can help get a fuller and more nuanced picture of a problem, specifically when combined with qualitative accounts (Thomas & Uminsky, 2022). Fairness metrics can thus provide a signal or a piece of evidence to develop hypotheses on the presence of societal disparities and, in turn, support broader equity efforts. Abebe et al. (2020) show how computational tools in general (e.g., algorithms) can play a variety of roles in this sense. Given the exercise of formalisation that they entail, for example, they can help articulate and reason on how societal problems are formulated in algorithms and highlight the limitations presented by computational techniques. At the same time, given their use or reliance on historical data, they can help diagnose emerging injustices and let historical ones resurface. Chapter 3 showed how, under this reconfiguration, fairness metrics can both help preserve evidence of injustices, as well as contribute to equity-driven fairness interventions.

Additionally, Chapter 4 showed how methods such as XAI approaches can help surface the conditions that enable algorithmic bias. Specifically, these approaches should be used to highlight the characteristics by which similar individuals are treated differently. For example, given two similar individuals with different credit outcomes, such an approach would entail naming the features by which these individuals are considered similar (e.g., income, zip code) precisely as what makes their (different) treatment unfair. On the one hand, differential treatment based on identical predictors of default risk but

varying gender or race suggests a potential for discrimination. On the other hand, using historically relevant predictors of default such as income or zip codes as a basis for this comparison, can itself add a discriminatory layer, which deserves further scrutiny. Having considered the cases of credit, health, and crime helped highlight important contextual and historical conditions by which this bias is perpetuated. Historical segregation in US neighbourhoods, for instance, profoundly affected the residents' access to credit, health insurance, and education (Agyeman, 2021; Perrino, 2020). In turn, this created the conditions of poverty, unemployment, and past default history by which residents in these communities are considered 'not worthy' of credit when zip codes are used to calculate the risk of default.

Precisely because of the historical and contextual character of these conditions, however, evidence of bias must be corroborated and complemented by qualitative accounts and the participation of multiple stakeholders. Chapter 3 has shown how model cards (M. Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2021) can be used to corroborate or refute evidence of bias identified through these metrics. Both approaches advocate for the thorough documentation of a model or dataset, encompassing a broad range of information, including qualitative dimensions such as intended applications, ethical implications, underlying assumptions, limitations, and additional contextual nuances. Chapter 5 stressed the importance of community-based participatory research in designing and evaluating responses to enforcement bias. By doing so, responses to enforcement bias can be designed and evaluated with a more nuanced and contextually informed approach that puts affected communities at the center.

2. *Evidence of algorithmic bias can be used to challenge as well as reinforce existing power structures.*

The use of algorithmic bias as evidence to inform policy interventions has the potential to challenge as much as to affirm the status quo.

Chapter 3 showed how algorithmic bias itself reproduces as well as mirrors existing societal disparities. Similarly, once applied in action, algorithmic bias can be used to motivate policies that reinforce as much as challenge existing power structures. Chapter 5 showed how, in the case of bias in predictive policing, stakeholders from different groups in Chicago articulated diverse problem diagnoses of the tool's algorithmic bias, strategically using it as evidence to advance criminal justice interventions that aligned with stakeholders' positionality and political ends.

The range of problem diagnoses and uses of algorithmic bias can suggest a 'relativistic' stance toward the issue, where the understanding of algorithmic bias is contingent upon an individual's subjective worldview. However, rather than dismissing these differences as mere relativism, I believe that they provide a valuable lens to understand the conditions under which algorithmic bias becomes a device to reinforce or challenge power dynamics. I think that unraveling the relationship between bias and positionality can serve as a critical tool to evaluate shortcomings and mistaken assumptions in potential responses to algorithmic bias coming from actors in relatively high positions of power, such as police officers and even academic researchers.

In the context of algorithmic bias and predictive policing, this means that those in positions of power, such as public sector actors and academics, may inadvertently reaffirm fundamentally unjust systems due to their distance from the lived experiences of communities impacted by incarceration. A critical consideration of positionality and bias may promote interventions that can serve affected communities' stated goals of

community investment and healing, contributing to the development of interventions that address the root cause of algorithmic bias in predictive policing.

Chapter 6 highlighted another crucial power asymmetry in addressing bias, that in knowledge about discrimination. A victim of algorithmic discrimination may be unaware that an AI system has produced that outcome or may not know that they have been discriminated against (Santoni de Sio & Mecacci, 2021). As previously mentioned, a person of colour who is denied a loan, for example, might not be aware that the bank relied on algorithms for credit scoring. A woman who is not shown the same job opportunity as a man might not even realise the difference. As this information is essential to ascribe responsibility, its ignorance might prevent a discriminated user from claiming discrimination. Regulatory initiatives that aim to tackle algorithmic discrimination without addressing this 'information gap' risk failing the people who need it the most. Importantly, there is a risk that a lack of awareness about being discriminated, lack of technical knowledge to understand AI's documentation, and an intricated legal landscape can help protect existing power structures (e.g., AI providers) and prevent discriminated users from seeking justice. In Chapter 6 I recommended some key improvements, which I elaborate on in the section below.

3. *Regulatory policies can help create recognised mechanisms for algorithmic bias to be considered as valid evidence of discrimination.*

For evidence of algorithmic bias to inform policy frameworks, existing regulations must recognise it as a valid source of evidence and build mechanisms for its effectiveness.

An attempt to use algorithmic bias as evidence of existing disparities can encounter legal challenges. In Chapter 3, it was suggested that the evidence of existing disparities captured in algorithmic bias can serve as a basis for informing fairness interventions aimed

at addressing historical inequalities. I argued that the evidence of bias, as detected by bias preserving metrics, could inform the application of bias transforming metrics toward achieving fairness outcomes driven by equity. Given evidence of historical credit disparities captured through a bias preserving metric, for example, I suggested that bias transforming metrics could be used to ensure that the predicted probability of being granted a loan is not strictly equal across groups but *equitable*: that it is responsive or informed by those historical disparities.

However, Chapter 3 nuanced this assertion by highlighting the reluctance within U.S. law to explicitly consider race or other protected characteristics when shaping policies, even with the intention of rectifying past discrimination. For instance, in 2023, the U.S. Supreme Court's ruling deemed race-conscious university admissions as a violation of the equal protection clause in the U.S. Constitution (K. Meyer, 2023). Therefore, the proposition to repurpose these metrics and advocate for a more equitable treatment of a protected group risks being construed as discrimination. Within existing legal frameworks, leveraging algorithmic bias to address historical discrimination must navigate through constrained legal allowances. Nevertheless, as I noted, scholars such as Hellman (2019) observe that the mere collection of information on racial categories does not inherently constitute racial discrimination, leaving room for utilising bias preserving metrics to gather evidence of racial discrimination. Moreover, rather than outright prohibiting the utilisation of such information, legal scrutiny is directed towards the specific manner in which it is employed (Hellman, 2019). Thus, it remains unclear whether, for example, using such information to foster better predictive outcomes for separate protected populations would be truly precluded.

Additionally, the role of statistical evidence in courtrooms has been an object of debate (Gastwirth, 2020; Post, 2021), complicating the use of algorithmic bias toward what

counts as relevant and sufficient evidence to prove discrimination. The case of the CRJA presented in Chapter 7, for example, is unique in countering the standards for evidence previously set by cases such as *McCleskey vs Kemp* (Neklason, 2019). As mentioned in Chapter 7, this case had ruled that evidence of 'racially disproportionate impact' in the Georgia death penalty was insufficient to mitigate a death penalty sentence without showing a racially discriminatory purpose (Post, 2021). In turn, the relevance of statistical evidence to prove discrimination in courts without evidence of discriminatory intent had been weakened. This jeopardised attempts to tackle those cases of discrimination that are of a systemic rather than an interpersonal nature. Regulatory initiatives such as the CRJA are crucial to set precedents for recognised mechanisms for statistical evidence and thus, algorithmic bias, to be recognised as a source of evidence that can count towards proving cases of implicit bias and systemic racism in the CJS. This shows that it is important for courts to update their norms and standards regarding the validity and relevance given to statistical evidence.

On the one hand, Chapter 6 has shown how evidence from AI systems such as technical documentation can help build a case of algorithmic discrimination. On the other hand, Chapter 7 showed how algorithmic bias itself can work as evidence of racial discrimination under the CRJA. As mentioned in the latter chapter, this can go beyond individual retribution, towards more substantial and systemic changes. Algorithmic bias could help recognise the structural diagnosis of systematic racism in the CJS. While individual remedies are essential, the focus should extend to systemic changes. Algorithmic bias can corroborate systemic issues, aiding public prosecutors in advocating for broader reforms. Additionally, algorithmic bias can be used as evidence to inform discussions and decisions on systemic changes in justice and police reforms. Scholars have proposed a series of interventions to tackle racism in the CJS that act outside of it.

Hasbrouck (2020) argued for new federal legislation to reduce the power of local police. Simonson (2021) argued to shift power away from the police towards policed populations. In the case at hand, for instance, data revealing the reproduction of historical redlining patterns through predictive policing could prompt a re-evaluation of law enforcement strategies, emphasising community support and healing over punitive measures.

8.2.2 Synthesis of policy recommendations

Section 1.6 of Chapter 1 has delineated the target audiences of this thesis. In this section, I present the policy recommendations formulated in each substantive chapter, summarising them for their relevant audience. Importantly, these audiences are interdisciplinary and thus, not mutually exclusive. Consequently, the presence of repetitions is there to ensure that each audience receives pertinent information. Additionally, given their interrelated nature, some overlap with the considerations above is inevitable. The following section is intended to function as an executive summary, allowing interested audiences to extract actionable points from individual chapters.

8.2.2.1 Stakeholders in and around (algorithmic) fairness

This set of policy recommendations is derived from Chapter 3. It addresses the interdisciplinary audience of AI developers and designers in companies, academic researchers, and decision-makers dealing with questions around algorithmic fairness. While the Chapter directly addresses the Fairness, Accountability, and Transparency (FAacT) in AI community, this section also extends beyond it to AI practitioners more in general. For the FAccT community, which is directly engaged in the study of algorithmic fairness, these recommendations can provide a framework for applying fairness metrics in a way that emphasizes substantive fairness—ensuring that fairness extends beyond formal metrics to address deeper societal concerns. For the broader AI practitioner community,

these recommendations offer practical guidance for integrating fairness metrics into their systems while navigating the intricate landscape of legal and ethical implications.

1. *Evidence validation through multiple methods*: Bias preserving metrics offer valuable insights into societal disparities but should not be solely relied upon. Validation through qualitative and participatory methods, along with documentation such as model cards (Mitchell et al., 2019) and datasheets for datasets (Gebru et al., 2021) should be considered crucial in guiding the work of AI developers, designers and academic researchers. For decision-makers, especially those involved in policy development, this recommendation underscores the importance of involving stakeholders impacted by algorithmic bias in intervention planning, alongside considering non-computational interventions like social welfare policies.

2. *Caution in equity interventions:* While bias preserving metrics can reveal disparities, using them to advocate for specific treatments based on protected characteristics will raise legal concerns. The collection of information on racial categories is distinct from its utilisation (Hellman, 2019); therefore, employing bias preserving metrics to gather evidence of discrimination may not inherently constitute bias. AI developers and designers should distinguish between the collection of data on protected characteristics and its subsequent use in decision-making processes, ensuring that their actions align with legal obligations. AI researchers should investigate the nuanced role that calibration metrics can play in understanding disparities and reframing trade-offs inherent in the choice of fairness metrics. Decision-makers must be aware of the legal implications associated with equity interventions that may rely on bias-preserving metrics, seeking advice from legal experts.

3. *Monitoring and reasoning on outcomes:* Bias transforming metrics can help monitor and reason on differential outcomes along task-relevant features for protected populations. Utilising race information to determine which features are predictive for different groups can indirectly contribute to reducing disparities. However, the legality of such practices is context-dependent and requires further research and testing. AI developers and designers are encouraged to use bias transforming metrics as suggested here to gain insights into how specific technical choices in their work can differentially impact protected groups. AI researchers can further explore the legal and technical feasibility of the approach described here. Decision-makers should consider integrating the ability hereby afforded by such metrics to monitor and reason on outcomes to inform policy interventions with fairness in mind.

8.2.2.2 Stakeholders at the intersection of explainability and fairness

These policy recommendations are derived from Chapter 4. While they are broadly relevant to the Fairness, Accountability and Transparency community, they specifically target explainable AI (XAI) practitioners: AI developers and designers in industry, academic researchers and decision-makers who aim to guide the development and regulation of AI systems with a focus on fairness and explainability.

1. *Clarification of the relevance of explainability for fairness*: XAI approaches should explicitly address fairness-related questions and solutions. XAI developers and designers should clarify the capabilities of their tools in tackling different aspects of algorithmic bias and provide clear instructions for their use. Academic researchers should better explore and clarify the analytical link between explainability and fairness. Decision-makers involved in AI policy developments

should ensure that regulatory initiatives recognise and promote the coherence and complementarity of XAI methods for fairness to avoid misuse and enhance their potential benefits.

2. *Reconsidering responsibility in addressing discrimination*: XAI approaches should not solely rely on discriminated users seeking recourse after discrimination. XAI developers and designers should leverage XAI methods to help notify instances of discrimination and provide constructive explanations for addressing its negative impact. Academic researchers should explore how XAI methods can be harnessed to prevent discrimination during deployment besides focusing on bias mitigation techniques. Decision-makers involved in policy development should incentivize external validation mechanisms, such as third-party audits which include XAI methods, to ensure that AI systems comply with fairness regulations and prevent future liability claims.

3. *Comprehensive explanations of discriminatory outcomes:* Explanations of discriminatory outcomes should identify not only the main cause but also the underlying conditions contributing to discrimination. Causal graphs (Hu, 2019) and counterfactual approaches (Karimi et al., 2020) can help provide a richer understanding of systemic discrimination, aiding in compliance with emerging regulatory initiatives. By naming the features contributing to differential treatment, XAI approaches can help address systemic biases and promote fairness in AI systems. Such methods should be leveraged by academic researchers to develop a more nuanced understanding of discriminatory mechanisms and outcomes, such as gaining insights into intersectional forms of discrimination. AI developers and designers in industry could utilize such methods to surface systemic biases in their systems. Decision-makers involved in policy development should use insights from

these comprehensive explanations to inform real-world fairness interventions aimed at tackling the conditions that perpetuate discrimination.

8.2.2.3 Policy practitioners and stakeholders in community safety

This set of policy recommendations is derived from Chapter 5. By addressing the socially contested nature of algorithmic bias within predictive policing, it offers insights for policymakers and stakeholders involved in community safety in Chicago specifically and in the US more generally.

1. *Recognise and address inherent biases in predictive policing:* Policy responses must acknowledge the multifaceted nature of algorithmic bias and its connection to systemic issues related to policing in the US. A nuanced understanding of bias should inform policy decisions, recognising the tension between reformist reforms operating within the current system and non-reformist reforms aiming for systemic change.

2. *Center marginalised communities:* Policy responses to crime should prioritise the knowledge, experience, and needs of marginalised, affected communities. Community-based participatory methods offer a way to involve these communities directly in the design and evaluation of responses to enforcement bias, ensuring that interventions are contextually informed and aligned with community goals.

3. *Critically examine bias and positionality:* Policies should critically examine the relationship between bias and positionality to ensure that responses prioritise equity and justice over reinforcing existing power dynamics. Standpoint theory can help recognise that those in positions of power may inadvertently uphold unjust systems,

further erasing the experiences and needs of marginalised communities in policy responses.

8.2.2.4 EU Policymakers

The following recommendations are derived from Chapter 6. Being addressed to EU policymakers, they aim to reduce the information gap for victims of algorithmic discrimination and enhance the effectiveness of the Artificial Intelligence Liability Directive (AILD) for AI users.

1. *Establish a holistic approach:* EU policymakers should adopt a comprehensive approach to defining the supranational legal framework of AI. This involves promoting a smooth integration of definitions and requirements outlined in the AI Act within the PLD and AILD, to ensure coherence across different pieces of legislation. Enhancing regulatory interoperability can reduce burdens and improve the effectiveness of these regulatory initiatives (Roberts et al., 2023).

2. *Promote coherence and complementarity:* Targeted measures should be implemented to harmonise the scope and implementation of the AILD and Product Liability Directive (PLD). This includes providing clarificatory guidance and support to stakeholders and courts, identifying examples of contexts for applying liability rules, and promoting complementarity through reviews and revisions.

3. *Incentivise less complex AI systems:* Policymakers should incentivise the development and deployment of AI systems that are less complex, less autonomous, and less opaque *ex ante*. Establishing European-wide standards for model interpretability and granting presumptions of conformity can encourage the

adoption of inherently interpretable models, promoting proactive accountability measures.

4. *Strengthen access to evidence:* Policymakers should strengthen the right of access to evidence for claimants by incentivising notification of harm. Introducing incentives for providers to share information about high-risk AI systems with trusted third parties can facilitate the notification of specific groups about potential biases or discrimination.

8.2.2.5 US (California) Policymakers

These recommendations are derived from Chapter 7. They are addressed to policymakers and stakeholders in the California criminal justice system (CJS) more specifically and within the US more generally. The following recommendations aim to strategically integrate algorithmic bias as evidence in regulatory initiatives to drive systemic reforms, specifically within the framework of the California Racial Justice Act:

1. *Implement Targeted Measures for Transparency and Accountability:* To effectively utilise the California Racial Justice Act (CRJA), policymakers must enhance reporting requirements to ensure transparency and accessibility of evidence. Initiatives like tech training for public defenders can bolster vigilance against discriminatory use of technologies, while mandates for algorithmic transparency can be enforced through mechanisms like third-party audit disclosures.

2. *Strategically Integrate Algorithmic Evidence:* To complement traditional forms of evidence, policymakers should strategically utilise algorithmic evidence in building discrimination claims. This includes utilising evidence sources in a complementary fashion by integrating anecdotal and expert witness declarations with statistical

information. Consideration of timing and geographical factors is crucial, with algorithmic bias being particularly relevant in cases where risk prediction and predictive policing tools influenced a court's decision.

3. *Drive Systemic Reforms Through Algorithmic Bias:* Policymakers should leverage evidence of algorithmic bias to address racial disparities and promote broader systemic reforms. This involves diagnosing structural issues within the CJS and using algorithmic bias as evidence to substantiate these issues, empowering policymakers to advocate for overarching reforms. Additionally, data on algorithmic bias can inform policymaking discussions regarding justice and police reforms, prompting a reassessment of law enforcement strategies toward community support and healing initiatives.

Audience	**Recommendations**
Stakeholders in algorithmic fairness	*Evidence captured through bias preserving metrics should be corroborated through qualitative and participatory methods.*
	Bias preserving metrics should be used to collect evidence of - rather than prescribe – the differential treatment of groups.
	Bias transforming metrics could be used to monitor and reason on – rather than enact – differential outcomes along task-relevant features.
Stakeholders in explainability	*The relevance of explainability for fairness should be explicitly articulated and integrated in AI development and regulation.*
	The responsibility to act on discriminatory outcomes should not lie exclusively with discriminated users.
	Explanations of discriminatory outcomes should name conditions, rather

	than just the main cause(s) of discrimination.
Policy practitioners and stakeholders in crime risk prediction	*Recognise and address the inherent biases within predictive policing through a nuanced understanding of algorithmic bias as evidence of broader systemic issues, connecting policy responses to existing debates between abolitionist and reformist reforms.*
	Center policy responses to crime around marginalised, affected communities, prioritising their knowledge, experience and needs through community-based participatory methods
	Advocate for policies that critically examine the relationship between bias and positionality to ensure responses to algorithmic bias in policing prioritise equity and justice over reinforcing existing power dynamics
EU Policy makers	*Establish a real holistic approach in defining the supranational legal framework of AI.*
	Establish targeted measures to promote coherence and complementarity in the EU-wide application of the AILD and PLD.
	Create incentives to develop and deploy AI systems that are less complex, less autonomous, and less opaque ex-ante, without over-relying on transparency requirements ex-post.
	Strengthen the right of access to evidence for the claimant by incentivising notification of harm
US (California) Policy Makers	*Implement Targeted Measures to Enhance Reporting Requirements for Transparency and Accountability as well as addressing challenges in expertise and access*
	Strategically integrate algorithmic evidence with conventional forms of evidence
	Drive systemic reforms through evidence for algorithmic bias by leveraging its potential for structural diagnosis of societal disparities

Table 3. Summary of policy recommendations for targeted audiences

8.3 Limitations and avenues for future research

In this concluding section, I reason on the limitations of the thesis. Specifically, in the following paragraphs, I elaborate on methodological limitations and limitations in scope and directions for future research. Even though the methodology section in the introduction has already addressed potential initial concerns with the research methodological set-up, this section aims to reflect on the limitations resulting from its execution.

8.3.1 Methodological limitations and reflections

The choice of an integrated thesis has allowed me to conceive this thesis in a modular fashion: as a system of components that can be separated and recombined without the meaning being lost. This strategy, as noted by Dunleavy (2003), provides advantages such as offering an agile read and catering to diverse audiences with varying interests or backgrounds. However, it inherently carries limitations, particularly in terms of topical depth. Unlike a monograph, which delves deeply into a specific aspect of a research topic, an integrated thesis like this one may just touch upon multiple aspects of it (Fridlund, 2010). The topic of algorithmic bias, however, belongs to a rapidly evolving field with multiple interdisciplinary dimensions. Conferences such as FaccT, for example, are increasingly incorporating diverse disciplinary perspectives (including policy more recently[63]) to address the multifaceted and rapidly evolving nature of algorithmic bias. In line with the research topic, this thesis has endeavoured to analyse algorithmic bias across various levels while respecting its ongoing evolution. Despite the inherent drawbacks of

[63] https://facctconference.org/2023/schedule

the chosen approach, thus, I believe that an integrated thesis framework has offered a suitable structure for addressing the research topic.

Furthermore, the pragmatic research approach adopted in this study prioritises the practical applicability of its findings within a specific context rather than focusing on their empirical generalisability (Morgan, 2007). A notable critique of this approach points to the external validity of the results, highlighting the challenge of extrapolating lessons to different situations or contexts. However, consistent with its pragmatic orientation, the primary objective of this thesis was not to prove some specific findings but rather to explore a range of considerations derived from the critical analyses conducted in each chapter. Consequently, the form of generalisation afforded by this approach is not empirical in nature. Instead, pragmatic research traditionally employs 'analytical generalisation', wherein researchers generalise by comparing results from particular studies to established theory and by positing empirical, but also logical connections between them (Yin, 1984). Adhering to this tradition, the various chapters of the thesis served to illustrate, rather than definitively prove, broader points, the validity of which depends not only on the presented data but also on the quality of reasoning employed in extrapolation. This approach is reflected in the conclusion, where general considerations are drawn while specific findings remain tied to individual substantive chapters.

Within the framework of an overarching critical pragmatist research paradigm, each chapter of this thesis has employed a distinct methodological approach. While Chapters 3 and 4 have primarily adopted theoretical perspectives, Chapter 5 has incorporated a structured empirical methodology. However, Chapters 6 and 7, characterised as ‘deep dives’, could have introduced some potential methodological concerns given their ‘lighter’ methodological approach. Both chapters rely on a document analysis (H. Morgan, 2022) of specific regulatory initiatives and thus, run the risk of

producing results that suffer from selectivity bias (H. Morgan, 2022) and lack sufficient detail (Bowen, 2009) to answer their driving research sub-question. According to selectivity bias, a limited collection of documents runs the risk of reflecting the limited goals and emphases of a particular institution (Yin, 1984). Additionally, basing research on a few documents might also provide insufficient material to craft answers to one's inquiry (Bowen, 2009). On the one hand, however, these chapters aimed to only explore the possibility of using bias as evidence in precisely the regulatory initiatives that they have analysed. As 'deep dives', they are meant as mere pointers into potential policy directions for the thesis. On the other hand, they are indirectly guided and motivated by the considerations drawn in the previous paragraphs. Thus, even if they stand as modular units, they have been informed by the results of more methodologically solid chapters.

8.3.2 Limitations in scope and directions for future research

A main limitation concerns the scope of the research and thus, the representativeness of its results. Predominantly, the focus lay on cases within the United States, with only one chapter dedicated to the European context (Chapter 6). Moreover, the reliance on English-speaking resources and the collection of English-language data further restricted the scope, tying the findings and implications of this research to a Western-centric perspective. This limitation is partly to be attributed to my positionality as a White European woman, as acknowledged in section 1.5 of Chapter 1, where my language and positionality were deemed most suitable for engaging with such resources. However, resource constraints also contributed to this limitation, as numerous resources and data concerning the deployment of risk prediction algorithms are predominantly available in the United States. Additionally, given the policy focus of the thesis, significant policy initiatives are underway in both the United States and the European Union, which are considered 'early movers' in this field (Roberts et al., 2023).

Despite these constraints, the thesis did not aim to 'prove' the existence of bias in new, unexplored contexts. Rather, it operated from the premise of its presence to investigate its potential policy implications. Therefore, it was essential to build upon an existing empirical and theoretical foundation. However, now that the research has presumably achieved its objective, subsequent studies should expand to unexplored contexts. At present, I am conducting follow-up policy research collaborating with researchers from various countries, including Japan, Brasil, Singapore, and South Korea, to survey their respective AI policies and to put to a test the alleged categorisation of Western countries as 'early movers'. This aims to address this limitation and broaden the geographical scope of the investigation.

Finally, regarding its focus, the thesis has failed to approach bias from an explicitly intersectional and lived standpoint. Intersectionality holds that there are multiple aspects of a person's identity that combine into unique forms of discrimination (Crenshaw, 1989, 1990). As an example, biases experienced by Black women are not necessarily the same as those experienced by White women or by Black men (Lepori, 2020). Additionally, this thesis has mostly focused on racial bias – occasionally gender bias – and in contexts that, in some chapters (e.g., Chapters 3 and 4) have only been considered through indirect sources. These choices can negatively affect the ability of the research to bring justice to the lived experience of marginalised communities. Especially for Chapter 5, for which this concern was most pressing, I have added a dedicated section for addressing this consideration in section 5.7. Still, future research could explore the role that algorithmic bias can play in addressing historical disparities at the intersection of multiple identities. Hoffmann (2019) and Xenidis 2018) provide a good starting point for such research.

8.4 Concluding remarks

In conclusion, this thesis has navigated the complex terrain of algorithmic bias within risk prediction algorithms in contexts as varied as credit, health, and criminal justice. It has grounded its research on an understanding of algorithmic bias as a manifestation of existing societal disparities rather than an isolated technical issue. The overarching goal was to understand how algorithmic bias could serve as evidence of these disparities to inform policies aimed at addressing the underlying societal conditions that enable them. The main research question asked, '*How, if at all, can algorithmic bias in risk prediction tools work as evidence for policy?*' To address this question, the research unfolded across four levels of inquiry: conceptual, methodological, applied, and policy.

Conceptually, bias was reframed as evidence of systemic problems, prompting a shift from merely mitigating its effects to identifying and addressing its root causes. Methodologically, this reorientation stimulated a redirection of bias detection tools toward uncovering the origins and mechanisms of bias. In practice, trying to grasp what bias can signify for different stakeholders led to a reevaluation of it as a device that can challenge as well as reinforce existing power dynamics. Ultimately, leveraging bias as evidence to inform policy frameworks contributed to broadening the scope of AI policies, shifting the focus from automation to addressing existing societal challenges.

Each chapter of the thesis delved into specific aspects of this inquiry, culminating in a series of policy recommendations tailored to different audiences and accompanied by overarching considerations on the role of algorithmic bias in policy frameworks to tackle discrimination. By reframing bias as evidence and harnessing its insights, this research has hopefully contributed to a more nuanced understanding of risk prediction tools and their societal implications.

Bibliography

A Pro-Innovation Approach to AI Regulation (2023). https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, *298*, 103502. https://doi.org/10.1016/j.artint.2021.103502

AB 2542. (2020). *AB 256*. ACLU California Action. https://aclucalaction.org/bill/ab-256/

Abdollahi, B., & Nasraoui, O. (2018). Transparency in Fair Machine Learning: The Case of Explainable Recommender Systems. In J. Zhou & F. Chen (Eds.), *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent* (pp. 21–35). Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0_2

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., & Robinson, D. G. (2020). Roles for computing in social change. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260. https://doi.org/10.1145/3351095.3372871

Abebe, R., Hardt, M., Jin, A., Miller, J., Schmidt, L., & Wexler, R. (2022). Adversarial Scrutiny of Evidentiary Statistical Software. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1733–1746. https://doi.org/10.1145/3531146.3533228

Abels, G., & Behrens, M. (2009). Interviewing Experts in Political Science: A Reflection on Gender and Policy Effects Based on Secondary Analysis. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing Experts* (pp. 138–156). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244276_7

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2870052

Adler, N. E., & Newman, K. (2002). Socioeconomic disparities in health: Pathways and policies. *Health Affairs (Project Hope)*, *21*(2), 60–76. https://doi.org/10.1377/hlthaff.21.2.60

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2016). Auditing Black-box Models for Indirect Influence. *arXiv:1602.07043 [Cs, Stat]*. http://arxiv.org/abs/1602.07043

Agre, P. (1997). *Toward a Critical Technical Practice*. https://pages.gseis.ucla.edu/faculty/agre/critical.html

Agyeman, J. (2021, March 9). *How urban planning and housing policy helped create 'food apartheid' in US cities*. The Conversation. http://theconversation.com/how-urban-planning-and-housing-policy-helped-create-food-apartheid-in-us-cities-154433

AI Bill of Rights. (2022). + *Algorithmic Discrimination Protections* | *OSTP*. The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/algorithmic-discrimination-protections-2/

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. *Proceedings of the 36th International Conference on Machine Learning*, 161–170. https://proceedings.mlr.press/v97/aivodji19a.html

All Info - S.2892 - 118th Congress (2023-2024): Algorithmic Accountability Act of 2023 (2023). https://www.congress.gov/bill/118th-congress/senate-bill/2892/all-info

Alvesson, M., & Sandberg, J. (2011). Generating Research Questions Through Problematization. *The Academy of Management Review*, *36*(2), 247–271.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Andrews, D. A., Bonta, J., & Wormith, J. S. (2006). The Recent Past and Near Future of Risk and/or Need Assessment. *Crime & Delinquency*, *52*(1), 7–27. https://doi.org/10.1177/0011128705281756

Angwin, J., & Larson, J. (2016). *Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say*. ProPublica. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say?token=IbTR1_gOPXDL2hhwhYHyPJPJxViQU3QU

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

Arvind Narayanan (Director). (2018, March 1). *Tutorial: 21 fairness definitions and their politics* [Video recording]. https://www.youtube.com/watch?v=jIXIuYdnyyk

Ascher, W. (1986). The evolution of the policy sciences: Understanding the rise and avoiding the fall. *Journal of Policy Analysis and Management*, *5*(2), 365–373.

Baeza-Yates, R. (2013). *Big data or right data? 1087*.

Balayn, A., & Gürses, S. (2021). Beyond Debiasing: Regulating AI and its inequalities. *EDRi Report. Https://Edri. Org/Wp-Content/Uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online. Pdf*.

Barabas, C. (2019). *Beyond Bias: Re-Imagining the Terms of 'Ethical AI' in Criminal Law* (SSRN Scholarly Paper ID 3377921). Social Science Research Network. https://doi.org/10.2139/ssrn.3377921

Barabas, C., Dinakar, K., Ito, J., Virza, M., & Zittrain, J. (2018). Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. *arXiv:1712.08238 [Cs, Stat]*. http://arxiv.org/abs/1712.08238

Barba, M., & Baustin, N. (2023, September 13). *Are San Francisco Police Officers Misreporting the Races of People They Stop?* The San Francisco Standard. https://sfstandard.com/2023/09/13/san-francisco-police-officer-misrepresented-race-bias-investigation/

Barocas, S., Hardt, M., & Narayanan, A. (2022). *Fairness and Machine Learning*. 312.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 80–89. https://doi.org/10.1145/3351095.3372830

Bartlett, R. P., Morse, A., Stanton, R., & Wallace, N. (2017). *Consumer Lending Discrimination in the FinTech Era* (SSRN Scholarly Paper 3063448). Social Science Research Network. https://doi.org/10.2139/ssrn.3063448

Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. *arXiv:2010.07389 [Cs, Stat]*. http://arxiv.org/abs/2010.07389

Benefiel, R. (2022). *Adversarial System of Justice*. Obo. https://www.oxfordbibliographies.com/display/document/obo-9780195396607/obo-9780195396607-0320.xml

Benjamin, R. (2019a). Assessing risk, automating racism. *Science*, *366*(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Benjamin, R. (2019b). Assessing risk, automating racism. *Science*, *366*(6464), 421–422. https://doi.org/10.1126/science.aaz3873

Benjamin, R. (2019c). *Race After Technology*. https://www.ruhabenjamin.com/race-after-technology

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. *arXiv:1703.09207 [Stat]*. http://arxiv.org/abs/1703.09207

Bertuzzi, L. (2022, September 30). *The new liability rules for AI*. Www.Euractiv.Com. https://www.euractiv.com/section/digital/podcast/the-new-liability-rules-for-ai/

Bertuzzi, L. (2023a, November 9). *OECD updates definition of Artificial Intelligence 'to inform EU's AI Act.'* Www.Euractiv.Com. https://www.euractiv.com/section/artificial-intelligence/news/oecd-updates-definition-of-artificial-intelligence-to-inform-eus-ai-act/

Bertuzzi, L. (2023b, December 14). *EU updates product liability regime to include software, Artificial Intelligence*. Www.Euractiv.Com. https://www.euractiv.com/section/digital/news/eu-updates-product-liability-regime-to-include-software-artificial-intelligence/

Bhat, A. (2018, August 3). Conceptual Research: Definition, Framework, Example and Advantages. *QuestionPro*. https://www.questionpro.com/blog/conceptual-research/

Bickman, L., & Rog, D. (1998). *The Sage Handbook of Applied Social Research Methods*. https://doi.org/10.4135/9781483348858

Biswas, S., & Rajan, H. (2021). Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline. *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 981–993. https://doi.org/10.1145/3468264.3468536

Blattner, L., & Nelson, S. (2021). How Costly is Noise? Data and Disparities in Consumer Credit. *arXiv:2105.07554 [Cs, Econ, q-Fin]*. http://arxiv.org/abs/2105.07554

Bond Graham, D. (2019, April 17). *California Cities Have Shredded Decades of Police Misconduct Records*. https://theappeal.org/california-cities-have-shredded-decades-of-police-misconduct-records/

Boucher, A. (2017). Power in elite interviewing: Lessons from feminist studies for political science. *Women's Studies International Forum*, *62*, 99–106. https://doi.org/10.1016/j.wsif.2017.05.003

Bowell, T. (n.d.). *Feminist Standpoint Theory \textbar Internet Encyclopedia of Philosophy*. Retrieved January 23, 2024, from https://iep.utm.edu/fem-stan/

Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal*, *9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Brayne, S. (2020). *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press.

Broussard, M., Diakopoulos, N., Guzman, A. L., Abebe, R., Dupagne, M., & Chuan, C.-H. (2019). Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly*, *96*(3), 673–695. https://doi.org/10.1177/1077699019859901

Bruno, I., Didier, E., & Vitale, T. (2014). *Statactivism: Forms of Action between Disclosure and Affirmation* (SSRN Scholarly Paper 2466882). https://papers.ssrn.com/abstract=2466882

Bryman, A. (2013). *Social Research Methods*. Oxford University Press.

Burton, S. L. (1992). "Thick" Concepts Revised. *Analysis*, *52*(1), 28–32. https://doi.org/10.2307/3328878

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292. https://doi.org/10.1007/s10618-010-0190-x

California Racial Justice Act (2020). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB2542&search_keywords=racial+justice+act

California Senate Bill 1421 (2018). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1421

Campbell, D. T. (1975). III. "Degrees of Freedom" and the Case Study. *Comparative Political Studies*, *8*(2), 178–193. https://doi.org/10.1177/001041407500800204

Carlson, A. M. (2017). *The Need for Transparency in the Age of Predictive Sentencing Algorithms | Iowa Law Review—The University of Iowa*. https://ilr.law.uiowa.edu/print/volume-103-issue-1/the-need-for-transparency-in-the-age-of-predictive-sentencing-algorithms

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), Article 8. https://doi.org/10.3390/electronics8080832

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, *12*(1), 4209. https://doi.org/10.1038/s41598-022-07939-1

Chander, A. (2017). The Racist Algorithm? *Michigan Law Review*, *115*(6), 1023–1045.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056 [Cs, Stat]*. http://arxiv.org/abs/1703.00056

Christiaen, L. L.-A. M. (2022, September 13). *AIDA: The Unfinished Third Act*. Lexology. https://www.lexology.com/library/detail.aspx?g=8bbb4183-aa8a-498e-8bef-e6d76c5eddd7

Christman, J. (2020). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

Citron, D. K. (2008). Technological Due Process. *Washington University Law Review*, *85*(6), 1249–1313.

Citron, D. K., & Pasquale, F. A. (2014). *The Scored Society: Due Process for Automated Predictions* (SSRN Scholarly Paper ID 2376209). Social Science Research Network. https://papers.ssrn.com/abstract=2376209

Collins, C. S., & Stockton, C. M. (2018). The Central Role of Theory in Qualitative Research. *International Journal of Qualitative Methods*, *17*(1), 1609406918797475. https://doi.org/10.1177/1609406918797475

COMMISSION STAFF WORKING DOCUMENT Evaluation of Council Directive 85/374/EEC of 25 July 1985 on the Approximation of the Laws, Regulations and Administrative Provisions of the Member States Concerning Liability for Defective Products Accompanying the Document Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the Application of the Council Directive on the Approximation of the Laws, Regulations, and Administrative Provisions of the Member States Concerning Liability for Defective Products (85/374/EEC) (2018). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018SC0157

Corbett-Davies, S., & Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv:1808.00023 [Cs]*. http://arxiv.org/abs/1808.00023

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. https://doi.org/10.1145/3097983.3098095

Crawford, K. (2013, April 1). The Hidden Biases in Big Data. *Harvard Business Review*. https://hbr.org/2013/04/the-hidden-biases-in-big-data

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., Raji, D., Rankin, J. L., Richardson, R., Schultz, J., West, S. M., & Whittaker, M. (2019). *AI Now Report*. 100.

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics [1989]. In *Feminist Legal Theory*. Routledge.

Crenshaw, K. (1990). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*, 1241.

Cullen, F. T., & Gendreau, P. (2001). From Nothing Works to What Works: Changing Professional Ideology in the 21st Century. *The Prison Journal*, *81*(3), 313–338. https://doi.org/10.1177/0032885501081003002

Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies*, *32*(1).

Danks, D., & London, A. (2017). *Algorithmic Bias in Autonomous Systems* (p. 4697). https://doi.org/10.24963/ijcai.2017/654

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy Non-Discrimination in Data-Driven Systems. *arXiv:1707.08120 [Cs]*. http://arxiv.org/abs/1707.08120

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/10.1109/SP.2016.42

Daumé, H. (2017). *8 | BIAS AND FAIRNESS*. 12.

De Luca, S. (2024). *New Product Liability Directive | Legislative Train Schedule*. European Parliament. https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-new-product-liability-directive

Deem, R. (1994). Researching the locally powerful: A study of school governance. In *Researching The Powerful In Education*. Routledge.

Denzin, N., & Lincoln, Y. (2024, February 9). *The SAGE Handbook of Qualitative Research*. SAGE Publications Inc. https://us.sagepub.com/en-us/nam/the-sage-handbook-of-qualitative-research/book242504

Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing Language Model Deployment with Risk Cards* (arXiv:2303.18190). arXiv. https://doi.org/10.48550/arXiv.2303.18190

Desrosières, A. (1998). *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard University Press.

Dewey, J. (1920). *Reconstruction in Philosophy*. H. Holt.

D'Ignazio, C. (2024). *Counting Feminicide: Data Feminism in Action*. The MIT Press.

DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive) (2022). https://commission.europa.eu/system/files/2022-09/1_1_197605_prop_dir_ai_en.pdf

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [Cs, Stat]*. http://arxiv.org/abs/1702.08608

Dunleavy, P. (2003). Planning an Integrated Thesis: The Macro-Structure. In P. Dunleavy (Ed.), *Authoring a PhD: How to plan, draft, write and finish a doctoral thesis or dissertation* (pp. 43–75). Macmillan Education UK. https://doi.org/10.1007/978-0-230-80208-7_3

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. *arXiv:1104.3913 [Cs]*. http://arxiv.org/abs/1104.3913

Edwards, R., & Holland, J. (2013). *What is Qualitative Interviewing?* Bloomsbury Academic. https://doi.org/10.5040/9781472545244

Elizabeth, C. (2022). Policing by Proxy: Interrogating Big Tech's Role in Law Enforcement. *University Honors Theses*. https://doi.org/10.15760/honors.1208

Ensign, D., Friedler, S., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research*, *81*. https://scholarship.haverford.edu/compsci_facpubs/113

Epp, C. R., Maynard-Moody, S., & Haider-Markel, D. P. (2014). *Pulled Over: How Police Stops Define Race and Citizenship*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/P/bo17322831.html

Eticas. (2021, September 13). Predictive Policing Oakland, California. *Eticas Foundation*. https://eticasfoundation.org/predicting-drug-related-crimes/

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Publishing Group.

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

European Parliament. (2023). *Compromise Amendments on the Draft Report—Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts* (Draft of Compromise Amendments on the Draft Report (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))).

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (2023). https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Express Analytics. (2022, April 1). Everything You Need To Know About Risk Prediction Models. *Express Analytics*. https://www.expressanalytics.com/blog/everything-you-need-to-know-about-risk-prediction-models/

Fahse, T., Huber, V., & van Giffen, B. (2021). Managing Bias in Machine Learning Projects. In F. Ahlemann, R. Schütte, & S. Stieglitz (Eds.), *Innovation Through Information Systems* (pp. 94–109). Springer International Publishing. https://doi.org/10.1007/978-3-030-86797-3_7

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Johnson, H., LaPointe, C., Llorens, A. J., Mackworth, A.

K., Maple, C., Pálsson, S. E., Pasquale, F., Winfield, A., & Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*(7), Article 7. https://doi.org/10.1038/s42256-021-00370-7

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8), e12760. https://doi.org/10.1111/phc3.12760

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *arXiv:1412.3756 [Cs, Stat]*. http://arxiv.org/abs/1412.3756

Ferguson, C., P'ng, J., Whiteside, H., & McAdams-Roy, K. (2022). *The Regulation of Artificial Intelligence in Canada and Abroad: Comparing the Proposed AIDA and EU AI Act*. https://www.fasken.com/en/knowledge/2022/10/18-the-regulation-of-artificial-intelligence-in-canada-and-abroad

Fernandez, F. C. (2023). *Bias Mitigation Strategies and Techniques for Classification Tasks*. https://www.holisticai.com/blog/bias-mitigation-strategies-techniques-for-classification-tasks

Fine, A., Le, S., & Miller, M. K. (2023). Content Analysis of Judges' Sentiments Toward Artificial Intelligence Risk Assessment Tools. *Criminology, Criminal Justice, Law & Society*, *24*(2), 31–46. https://doi.org/10.54555/CCJLS.8169.84869

Fiscella, K., Franks, P., Gold, M. R., & Clancy, C. M. (2000). Inequality in quality: Addressing socioeconomic, racial, and ethnic disparities in health care. *JAMA*, *283*(19), 2579–2584. https://doi.org/10.1001/jama.283.19.2579

Flaxman, S., Chirico, M., Pereira, P., & Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "Real-Time Crime Forecasting Challenge." *The Annals of Applied Statistics*, *13*(4), 2564–2585. https://doi.org/10.1214/19-AOAS1284

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to Design AI for Social Good: Seven Essential Factors. *Science and Engineering Ethics*, *26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fogliato, R., Xiang, A., Lipton, Z., Nagin, D., & Chouldechova, A. (2021). On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes. *arXiv:2105.04953 [Stat]*. http://arxiv.org/abs/2105.04953

Foryciarz, A., Pfohl, S. R., Patel, B., & Shah, N. (2022). Evaluating algorithmic fairness in the presence of clinical guidelines: The case of atherosclerotic cardiovascular disease risk estimation. *BMJ Health & Care Informatics*, *29*(1), e100460. https://doi.org/10.1136/bmjhci-2021-100460

Franchi, M., Zamfirescu-Pereira, J. D., Ju, W., & Pierson, E. (2023). Detecting disparities in police deployments using dashcam data. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 534–544. https://doi.org/10.1145/3593013.3594020

Frey, B. B. (2018). *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Fridlund, B. (2010). *The dissertation book; Should it be a monograph or a compilation thesis?* https://watermark.silverchair.com/eurjcn0144.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAA1wwggNYBgkqhkiG9w0BBwagggNJMIIDRQIBADCCAz4GCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQM7uKnLE3fMUO7cN9IAgEQgIIDD5pnc5DT3vRvsEg2lS_kpNlLc9nfRMN_tXE2uB3KYlfmT2I2WOghMSl6_99iHNrvtfNgI8WC3Fpzx4UxposhAm2-hYcWvD95yz_WE-pLSNeJhsV10q2w9-NraSBNms_L9lE2zbqFjGpc8LVmOdBEjtKHokcjeFnt66IQ2EJxivMM-FRUbEtO1fwnqPnJdr4Io9MNNxdfDkbgHEOpnwvyZ-

86ZKFkjbQLYyPbT2lWuNUlJMNZpmpueAR1XWeY4OolvCVNcrAhU7GUU0zidAJleFuzlRfO20uACYC6AxZ--kJh8aPqRPs19ppNo5_NsfiNNj9JLJPLXhkp1hzWJD59ezJdzjJmTwlEgrDl9MwS4VaixyQyux9I754sjv3VL34Z9lhbqt0PSSDvK6ZTGl7PvIP6S9iKpBEZyQIUzU7-8z1Rg9D4wv2D-6A_f3EkMCGr3JjsJzpd577EQ_MmdJyL8yv7U3oLrHQ486qRB9H0iq-zynyyXqPar260cOWScMt5VCm_8cuM0LImtV78isAp56_RrGTIk0OQJlnlC_Oxo4ePlhsQi86oi4SkMG85XGNg2BSGogumwGyQKFe_LZ7LnPvUetEZ9xqDYezTd-0L3hMl6o6kHpJdeMqeyKUY63WKEPB4eSAMCRu7iI1zjykjqvBJSWEujxqxcszbCY8VXVJY0bh1bnyaHYGAxz7sW59PXuVtBWisad5XNUOJM2_aqR5QKybxV5FFhLD1cBx2q07tUtdvPOQCRA6EaQ_soF7FBOtSPp3JbeV74Jk4we4Jvfflss9nJPumL-MtKq0YTYJrlXuzCLt8v-_lCx3N931OwAiJhVCVupx3nxaYqtIpgDFmijsK6pSsMmQtPlF7XvXVGa4ihqW7XXoO3RKn9xIeAfaBwc8Ief8q8IUeyvNCPMCJmcmeEpkaidz04lm9Ku0p8A_nW-4V7pbz2KCQcTiPLnPSD2fpAEYCYie8ri7A1EDCFMD-OFCt57Y6RU6_Qf4xuxDRGgy9svE5PJHvh10zWQn2vTJ5pTFUT1Wf-y4qTg

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Frye, C., Rowat, C., & Feige, I. (2020). Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, *33*, 1229–1239. https://proceedings.neurips.cc/paper/2020/hash/0d770c496aa3da6d2c3f2bd19e7b9d6b-Abstract.html

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2020). *Predictably Unequal? The Effects of Machine Learning on Credit Markets* (SSRN Scholarly Paper ID 3072038). Social Science Research Network. https://doi.org/10.2139/ssrn.3072038

Future of Life Institute. (2017). *AI Principles*. Future of Life Institute. https://futureoflife.org/ai-principles/

Galhotra, S., Pradhan, R., & Salimi, B. (2021). *Explaining Black-Box Algorithms Using Probabilistic Contrastive Counterfactuals* (arXiv:2103.11972). arXiv. https://doi.org/10.48550/arXiv.2103.11972

Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness Metrics: A Comparative Analysis. *arXiv:2001.07864 [Cs]*. http://arxiv.org/abs/2001.07864

Gastwirth, J. L. (2020). The Role of Statistical Evidence in Civil Cases. *Annual Review of Statistics and Its Application*, *7*(1), 39–60. https://doi.org/10.1146/annurev-statistics-031219-041238

Gates, E. (2018, May 12). Predictive Policing in LA: LAPD Employs Palantir for Surveillance. *American Judicial System*. https://www.ajs.org/predictive-policing-in-la-lapd-employs-palantir-for-surveillance/

Gebru, T. (2020, July 9). *Race and Gender*. The Oxford Handbook of Ethics of AI. https://doi.org/10.1093/oxfordhb/9780190067397.013.16

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). *Datasheets for Datasets* (arXiv:1803.09010). arXiv. https://doi.org/10.48550/arXiv.1803.09010

Gelman, A., Fagan, J., & Kiss, A. (2007). An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias. *Journal of the American Statistical Association*, *102*(479), 813–823. https://doi.org/10.1198/016214506000001040

Goel, S., Shroff, R., Skeem, J. L., & Slobogin, C. (2018). *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment* (SSRN Scholarly Paper ID 3306723). Social Science Research Network. https://doi.org/10.2139/ssrn.3306723

Golub, G. (2022, March 6). Primary Source Analysis – What to Expect. *Students@LSE*. https://blogs.lse.ac.uk/studentsatlse/2022/03/06/primary-source-analysis-what-to-expect/

Gomede, E. (2024, March 7). *Data Quality vs. Data Quantity: The Crucial Balance for Artificial Intelligence*. Medium. https://python.plainenglish.io/data-quality-vs-data-quantity-the-crucial-balance-for-artificial-intelligence-faed8b0eaea4

Gomez, O., Holter, S., Yuan, J., & Bertini, E. (2020). *ViCE: Visual Counterfactual Explanations for Machine Learning Models* (arXiv:2003.02428). arXiv. https://doi.org/10.48550/arXiv.2003.02428

Gorz, A. (1968). Strategy for labor. A radical proposal. *Science and Society*, *32*(4).

Gravett, W. H. (2024). Judicial Decision-Making in the Age of Artificial Intelligence. In H. Sousa Antunes, P. M. Freitas, A. L. Oliveira, C. Martins Pereira, E. Vaz de Sequeira, & L. Barreto Xavier (Eds.), *Multidisciplinary Perspectives on Artificial Intelligence and the Law* (pp. 281–297). Springer International Publishing. https://doi.org/10.1007/978-3-031-41264-6_15

Green, B. (2019). Good" isn't good enough. *Proceedings of the AI for Social Good Workshop at NeurIPS*, *17*.

Green, B. (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 594–606. https://doi.org/10.1145/3351095.3372869

Green, B. (2021a). Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*, *2*(3), 249–265. https://doi.org/10.23919/JSC.2021.0029

Green, B. (2021b). *Escaping the "Impossibility of Fairness": From Formal to Substantive Algorithmic Fairness* (SSRN Scholarly Paper ID 3883649). Social Science Research Network. https://doi.org/10.2139/ssrn.3883649

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 19–31. https://doi.org/10.1145/3351095.3372840

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Greiner, D. J. (2008). Casual Inference in Civil Rights Litigation. *Harvard Law Review*, *122*, 533.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), Article 1. https://doi.org/10.1609/aaai.v32i1.11296

Grill, G., Fischer, F., & Cech, F. (2023). Bias as Boundary Object: Unpacking The Politics Of An Austerity Algorithm Using Bias Frameworks. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1838–1849. https://doi.org/10.1145/3593013.3594120

Hacker, P. (2022). The European AI Liability Directives – Critique of a Half-Hearted Approach and Lessons for the Future. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4279796

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., Kundaje, A., Greene, C. S., Broderick, T., Hoffman, M. M., Leek, J. T., Korthauer, K., Huber, W., Brazma, A., Pineau, J., Tibshirani, R., Hastie, T., … Aerts, H. J. W. L. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), Article 7829. https://doi.org/10.1038/s41586-020-2766-y

Hajian, S., & Domingo-Ferrer, J. (2013). A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459. IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2012.72

Hampson, T., & McKinley, J. (2023). Problems posing as solutions: Criticising pragmatism as a paradigm for mixed research. *Research in Education*, *116*(1), 124–138. https://doi.org/10.1177/00345237231160085

Hanks, H., Austin, A., & Franck-Gwinnell, E. (2024, March 1). *Text of the new EU Product Liability Directive is agreed—What happens next?* Lexology. https://www.lexology.com/library/detail.aspx?g=a18557f4-ae63-46b3-8e8b-7c1120e6b529

Hannah-Moffat, K. (2013). Actuarial Sentencing: An "Unsettled" Proposition. *Justice Quarterly*, *30*(2), 270–296. https://doi.org/10.1080/07418825.2012.682603

Hannah-Moffat, K. (2015). The Uncertainties of Risk AssessmentPartiality, Transparency, and Just Decisions. *Federal Sentencing Reporter*, *27*(4), 244–247. https://doi.org/10.1525/fsr.2015.27.4.244

Hao, K., & Stray, J. (2019). *Can you make AI fairer than a judge? Play our courtroom algorithm game*. MIT Technology Review. https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

Harcourt, B. (2015a). Risk as a Proxy for Race: The Dangers of Risk Assessment. *Fed. Sent'g Rep.*, *27*, 237.

Harcourt, B. (2015b). Risk as a Proxy for Race: The Dangers of Risk Assessment. *Fed. Sent'g Rep.*, *27*, 237.

Harcourt, B. E. (2008). *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. University of Chicago Press.

Harcourt, B. E. (2010). *Risk as a Proxy for Race* (SSRN Scholarly Paper ID 1677654). Social Science Research Network. https://papers.ssrn.com/abstract=1677654

Harding, S. (1992). Rethinking Standpoint Epistemology: What is "Strong Objectivity?". *The Centennial Review*, *36*(3), 437–470.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [Cs]*. http://arxiv.org/abs/1610.02413

Hasbrouck, B. (2020). Abolishing Racist Policing With the Thirteenth Amendment. *UCLA Law Review*. https://scholarlycommons.law.wlu.edu/wlufac/631

Hellman, D. (2019). *Measuring Algorithmic Fairness* (SSRN Scholarly Paper ID 3418528). Social Science Research Network. https://papers.ssrn.com/abstract=3418528

Heskes, T., Sijben, E., Bucur, I. G., & Claassen, T. (2020). *Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models* (arXiv:2011.01625). arXiv. https://doi.org/10.48550/arXiv.2011.01625

Hill, K. (2023). *Your Face Belongs to Us: The Secretive Startup Dismantling Your Privacy*. Simon and Schuster.

Hill, R. K. (2016). Genealogy. In *Routledge Encyclopedia of Philosophy* (1st ed.). Routledge. https://doi.org/10.4324/9780415249126-DE024-1

Hinton, E. (2017). *From the War on Poverty to the War on Crime*. Harvard University Press. https://www.hup.harvard.edu/books/9780674979826

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Hu, L. (2019, October 11). Disparate Causes, pt. I. *Phenomenal World*. https://www.phenomenalworld.org/analysis/disparate-causes-i/

Hupont, I., Tolan, S., Gunes, H., & Gómez, E. (2022). The landscape of facial processing applications in the context of the European AI Act and the development of trustworthy systems. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-14981-6

Hurley, M., & Adebayo, J. (2017). CREDIT SCORING IN THE ERA OF BIG DATA. *Big Data*, *18*, 69.

Hurtado, A., & Yan, B. (2023). *The California Racial Justice Act 2020 Landmark Legislation 2022 Clarifying Changes Introduction and Overview*. https://www.acgov.org/board/bos_calendar/documents/DocsAgendaReg_10_26_23/PUBLIC%20PROTECTION/Regular%20Calendar/Item_1_Racial_Just_Act_overview.pdf

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences*, *12*(3), Article 3. https://doi.org/10.3390/app12031353

James, W. (1907). William James: Pragmatism: A New Name for Some Old Ways of Thinking. In *Central Works of Philosophy: Volume 4: The Twentieth Century: Moore to Popper* (Vol. 4, pp. 54–70). Acumen Publishing. https://doi.org/10.1017/UPO9781844653614.005

Jasanoff, S. (2006). *"Technology as a Site and Object of Politics.* https://www.hks.harvard.edu/publications/technology-site-and-object-politics

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2

Johnson, C. (2022a, January 26). Flaws plague a tool meant to help low-risk federal prisoners win early release. *NPR*. https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act

Johnson, C. (2022b, April 19). Justice Department works to curb racial bias in deciding who's released from prison. *NPR*. https://www.npr.org/2022/04/19/1093538706/justice-department-works-to-curb-racial-bias-in-deciding-whos-released-from-pris

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Bloebaum, P., & Bareinboim, E. (2022). On Measuring Causal Contributions via do-interventions. *Proceedings of the 39th International Conference on Machine Learning*, 10476–10501. https://proceedings.mlr.press/v162/jung22a.html

Kaminski, M. E. (2022). *Regulating the Risks of AI* (SSRN Scholarly Paper 4195066). https://doi.org/10.2139/ssrn.4195066

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kang, H.-W., & Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *PLOS ONE*, *12*(4), e0176244. https://doi.org/10.1371/journal.pone.0176244

Karakatsanis, A. (2019). *Usual cruelty: The complicity of lawyers in the criminal injustice system*. The New Press.

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2021). *A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects* (arXiv:2010.04050). arXiv. http://arxiv.org/abs/2010.04050

Karimi, A.-H., Schölkopf, B., & Valera, I. (2020). *Algorithmic Recourse: From Counterfactual Explanations to Interventions* (arXiv:2002.06278). arXiv. https://doi.org/10.48550/arXiv.2002.06278

Karppinen, K., & Moe, H. (2011). What we talk about when we talk about document analysis. In *Trends in Communication Policy Research*. Intellect.

Kaushik, V., & Walsh, C. A. (2019). Pragmatism as a Research Paradigm and Its Implications for Social Work Research. *Social Sciences*, *8*(9), Article 9. https://doi.org/10.3390/socsci8090255

Kennedy, L. W., Caplan, J. M., & Piza, E. L. (2018). *Risk-Based Policing: Evidence-Based Crime Prevention with Big Data and Spatial Analytics*.

Kenton, B. (2021). *Prima Facie: Legal Definition and Examples*. Investopedia. https://www.investopedia.com/terms/p/prima-facie.asp

Kilonzo, S. M., & Ojebode, A. (2023). Research Methods for Public Policy. In E. R. Aiyede & B. Muganda (Eds.), *Public Policy and Research in Africa* (pp. 63–85). Springer International Publishing. https://doi.org/10.1007/978-3-030-99724-3_4

Kleinberg, J. (2018). Inherent Trade-Offs in Algorithmic Fairness. *ACM SIGMETRICS Performance Evaluation Review*, *46*(1), 40. https://doi.org/10.1145/3292040.3219634

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). *Human Decisions and Machine Predictions* (w23180). National Bureau of Economic Research. https://doi.org/10.3386/w23180

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/jla/laz001

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [Cs, Stat]*. http://arxiv.org/abs/1609.05807

Knox, D., Lowe, W., & Mummolo, J. (2020). Administrative Records Mask Racially Biased Policing. *American Political Science Review*, *114*(3), 619–637. https://doi.org/10.1017/S0003055420000039

Kohler-Hausmann, I. (2019). EDDIE MURPHY AND THE DANGERS OF. *NORTHWESTERN UNIVERSITY LAW REVIEW*.

König, P. D., & Krafft, T. D. (2021). Evaluating the evidence in algorithmic evidence-based decision-making: The case of US pretrial risk assessment tools. *Current Issues in Criminal Justice*, 1–23. https://doi.org/10.1080/10345329.2020.1849932

Köse, Ö. D., & Shen, Y. (2021). *Fairness-Aware Node Representation Learning* (arXiv:2106.05391). arXiv. https://doi.org/10.48550/arXiv.2106.05391

Kumar, A. (2018, November 20). *Machine Learning Models: Bias Mitigation Strategies - DZone AI*. Dzone.Com. https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Lapowsky, I. (2018). How the LAPD Uses Data to Predict Crime. *Wired*. https://www.wired.com/story/los-angeles-police-department-predictive-policing/

Leben, D. (2023). Explainable AI as evidence of fair decisions. *Frontiers in Psychology*, *14*. https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1069426

Lee, M. S. A., Floridi, L., & Singh, J. (2020). *Formalising trade-offs beyond algorithmic fairness: Lessons from ethical philosophy and welfare economics* (SSRN Scholarly Paper ID 3679975). Social Science Research Network. https://doi.org/10.2139/ssrn.3679975

Lepori, M. A. (2020). Unequal Representations: Analyzing Intersectional Biases in Word Embeddings Using Representational Similarity Analysis. *arXiv:2011.12086 [Cs]*. http://arxiv.org/abs/2011.12086

Lessig, L. (1999). *Code and Other Laws of Cyberspace*. Basic Books, Inc.

Lewin, C. (2007). *The Development of Actuarial Science: From Early Insurance and Pensions to Establishment of the Actuarial Profession | PDF | Actuarial Science | Actuary*. https://www.scribd.com/document/337889563/Seminar-2007-Lewin

Lin, Z. "Jerry," Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*(7), eaaz0652. https://doi.org/10.1126/sciadv.aaz0652

Lipton, M. (1977). *Why poor people stay poor: A study of urban bias in world development*. Temple Smith ; Australian National University Press. https://openresearch-repository.anu.edu.au/handle/1885/114902

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, R. (2018). *Bias Mitigation Post-processing for Individual and Group Fairness* (arXiv:1812.06135). arXiv. https://doi.org/10.48550/arXiv.1812.06135

Lomas, N. (2019, August 10). Most EU cookie "consent" notices are meaningless or manipulative, study finds. *TechCrunch*. https://techcrunch.com/2019/08/10/most-eu-cookie-consent-notices-are-meaningless-or-manipulative-study-finds/

Lowe, M. (2022). *Fruit of the Racist Tree: A Super-Exclusionary Rule for Racist Policing Under California's Racial Justice Act*. https://www.yalelawjournal.org/comment/fruit-of-the-racist-tree

Lucas D. Introna, H. N. (2000). Shaping the Web: Why the Politics of Search Engines Matters. *The Information Society*, *16*(3), 169–185. https://doi.org/10.1080/01972240050133634

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Lundberg, S. (2018). *Explaining quantitative measures of fairness—SHAP latest documentation*. https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Explaining%20quantitative%20measures%20of%20fairness.html

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Mackrael, K., & Schechner, S. (2023). *European Lawmakers Pass AI Act, World's First Comprehensive AI Law*. WSJ. https://www.wsj.com/tech/ai/ai-act-passes-european-union-law-regulation-e04ec251

Malik, M. M. (2020). A Hierarchy of Limitations in Machine Learning. *arXiv:2002.05193 [Cs, Econ, Math, Stat]*. http://arxiv.org/abs/2002.05193

Mann, S., Nolan, J., & Wellman, B. (2003). Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments. *Surveillance & Society*, *1*(3), 331–355.

Mapping Pretrial Injustice. (n.d.). Common Pretrial Risk Assessments. *Mapping Pretrial Risk*. Retrieved February 12, 2024, from https://pretrialrisk.com/the-basics/common-prai/

Martineau, K. (2022). *Debugging foundation models for bias*. IBM Research Blog. https://research.ibm.com/blog/debugging-AI-bias

Mashiat, T., Gitiaux, X., Rangwala, H., Fowler, P. J., & Das, S. (2022). *Trade-offs between Group Fairness Metrics in Societal Resource Allocation* (arXiv:2202.12334). arXiv. https://doi.org/10.48550/arXiv.2202.12334

Maurutto, P., & Hannah-Moffat, K. (2006). Assembling Risk and the Restructuring of Penal Control. *The British Journal of Criminology*, *46*(3), 438–454. https://doi.org/10.1093/bjc/azi073

Mayson, S. G. (2018). *Bias In, Bias Out* (SSRN Scholarly Paper ID 3257004). Social Science Research Network. https://papers.ssrn.com/abstract=3257004

McCleskey v. Kemp, 481 U.S. 279 (1987) (1987). https://supreme.justia.com/cases/federal/us/481/279/

McGreal, C., Beckett, L., Laughland, O., & Ajasa, A. (2021, April 21). Derek Chauvin found guilty of murder of George Floyd. *The Guardian*. https://www.theguardian.com/us-news/2021/apr/20/derek-chauvin-verdict-guilty-murder-george-floyd

McGreevy, P. (2020, November 4). *Prop. 25, which would have abolished California's cash bail system, is rejected by voters*. Los Angeles Times. https://www.latimes.com/california/story/2020-11-03/2020-california-election-prop-25-results

McKay, C. (2020). Predicting risk in criminal procedure: Actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice*, *32*(1), 22–39. https://doi.org/10.1080/10345329.2019.1658694

McLeod, A. M. (2018). Envisioning abolition democracy. *Harv. L. Rev.*, *132*, 1613.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [Cs]*. http://arxiv.org/abs/1908.09635

Meyer, K. (2023). *The end of race-conscious admissions*. Brookings. https://www.brookings.edu/articles/the-end-of-race-conscious-admissions/

Meyer, W. B., & Graybill, J. K. (2016). The suburban bias of American society? *Urban Geography*, *37*(6), 863–882. https://doi.org/10.1080/02723638.2015.1118990

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). Qualitative Data Analysis: A Methods Sourcebook. Third Edition. In *SAGE Publications Ltd (CA)*. SAGE Publications.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, *8*(1), annurev-statistics-042720-125902. https://doi.org/10.1146/annurev-statistics-042720-125902

Mohler, G., & Porter, M. D. (2018). Rotational grid, PAI-maximizing crime forecasts. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *11*(5), 227–236. https://doi.org/10.1002/sam.11389

Mökander, J. (2023). Auditing of AI: Legal, Ethical and Technical Approaches. *Digital Society*, *2*(3), 49. https://doi.org/10.1007/s44206-023-00074-y

Mokander, J., & Floridi, L. (2021). *Ethics-based Auditing to Develop Trustworthy AI* (SSRN Scholarly Paper ID 3788841). Social Science Research Network. https://papers.ssrn.com/abstract=3788841

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021a). Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, *27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021b). Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, *27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4

Monahan, J., & Skeem, J. L. (2015). *Risk Assessment in Criminal Sentencing* (SSRN Scholarly Paper ID 2662082). Social Science Research Network. https://papers.ssrn.com/abstract=2662082

Morgan, D. L. (2007). Paradigms Lost and Pragmatism Regained: Methodological Implications of Combining Qualitative and Quantitative Methods. *Journal of Mixed Methods Research*, *1*(1), 48–76. https://doi.org/10.1177/2345678906292462

Morgan, H. (2022). Conducting a Qualitative Document Analysis. *The Qualitative Report*, *27*(1), 64–77. https://doi.org/10.46743/2160-3715/2022.5044

Morley, J., & Floridi, L. (2020). The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. *Science and Engineering Ethics*, *26*(3), 1159–1183. https://doi.org/10.1007/s11948-019-00115-1

Movement Alliance Project, & MediaJustice. (2019). *Mapping Pretrial Injustice: A Community-Driven Database*. Mapping Pretrial Injustice. https://pretrialrisk.com/

Mullaney, T. S., Peters, B., Hicks, M., & Philip, K. (2021). *Your Computer Is on Fire*. MIT Press.

Nabi, R., & Shpitser, I. (2018). *Fair Inference On Outcomes* (arXiv:1705.10378). arXiv. http://arxiv.org/abs/1705.10378

Nannini, L., Balayn, A., & Smith, A. L. (2023). *Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK* (arXiv:2304.11218). arXiv. https://doi.org/10.48550/arXiv.2304.11218

National Institute of Justice. (2021). *2021 Review and Revalidation of the First Step Act Risk Assessment Tool*. https://www.ojp.gov/pdffiles1/nij/303859.pdf

Nawaz, S. A. (2022, November 7). The Proposed EU AI Liability Rules: Ease or Burden? *European Law Blog*. https://europeanlawblog.eu/2022/11/07/the-proposed-eu-ai-liability-rules-ease-or-burden/

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6). https://doi.org/10.1371/journal.pdig.0000278

Neklason, A. (2019, June 14). The ‘Death Penalty’s Dred Scott’ Lives On. *The Atlantic*. https://www.theatlantic.com/politics/archive/2019/06/legacy-mccleskey-v-kemp/591424/

Newell, B. (2020). Introduction: The State of Sousveillance. *Surveillance & Society*, *18*(2), 257–261. https://doi.org/10.24908/ss.v18i2.14013

Nierenberg, A. (2023, September 4). Over 100 Connecticut State Troopers Accused of Faking Traffic Stops. *The New York Times*. https://www.nytimes.com/2023/09/04/nyregion/connecticut-false-tickets.html

Novelli, C., Casolari, F., Hacker, P., Spedicato, G., & Floridi, L. (2024). *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity* (SSRN Scholarly Paper 4694565). https://doi.org/10.2139/ssrn.4694565

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: What it is and how it works. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01635-y

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic Analysis: Striving to Meet the Trustworthiness Criteria. *International Journal of Qualitative Methods*, *16*(1), 1609406917733847. https://doi.org/10.1177/1609406917733847

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

OECD. (2021). *Regulatory policy—OECD*. https://www.oecd.org/regreform/regulatory-policy/

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*. https://doi.org/10.3389/fdata.2019.00013

Ortbals, C. D., & Rincker, M. E. (2009). Fieldwork, Identities, and Intersectionality: Negotiating Gender, Race, Class, Religion, Nationality, and Age in the Research Field Abroad: Editors’ Introduction. *PS: Political Science & Politics*, *42*(2), 287–290. https://doi.org/10.1017/S104909650909057X

Packt, R. G. (2018). Machine Learning Explainability vs Interpretability: Two concepts that could help restore trust in AI. *KDnuggets*. https://www.kdnuggets.com/machine-learning-explainability-vs-interpretability-two-concepts-that-could-help-restore-trust-in-ai.html

Park, Y., Hu, J., Singh, M., Sylla, I., Dankwa-Mullan, I., Koski, E., & Das, A. K. (2021). Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. *JAMA Network Open*, *4*(4), e213909. https://doi.org/10.1001/jamanetworkopen.2021.3909

Parker, C., Scott, S., & Geddes, A. (2020). Snowball Sampling. In *SAGE Research Methods Foundations*. SAGE Publications Ltd. https://doi.org/10.4135/9781526421036831710

Passi, S., & Barocas, S. (2019). Problem Formulation and Fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. https://doi.org/10.1145/3287560.3287567

Pastaltzidis, I., Dimitriou, N., Quezada-Tavarez, K., Aidinlis, S., Marquenie, T., Gurzawska, A., & Tzovaras, D. (2022). Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2302–2314. https://doi.org/10.1145/3531146.3534644

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearsall, B. (2010). *Archived | Predictive Policing: The Future of Law Enforcement? | National Institute of Justice*. https://nij.ojp.gov/topics/articles/predictive-policing-future-law-enforcement

Peirce, C. S. (1903). *Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism*. SUNY Press.

Perrino, J. (2020, July 2). *"Redlining" and health indicators: Decisions made 80 years ago have health consequences today » NCRC*. https://ncrc.org/redlining-and-health-indicators-decisions-made-80-years-ago-have-health-consequences-today/

Porter, T. M. (1995). *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.

Porter, Z., Zimmermann, A., Morgan, P., McDermid, J., Lawton, T., & Habli, I. (2022). Distinguishing two features of accountability for AI technologies. *Nature Machine Intelligence*, *4*(9), Article 9. https://doi.org/10.1038/s42256-022-00533-0

Post, D. (2021, October 23). Opinion | Is the Supreme Court allergic to math? *Washington Post*. https://www.washingtonpost.com/news/volokh-conspiracy/wp/2017/10/19/is-the-supreme-court-allergic-to-math/

Prince, A. E. R., & Schwarcz, D. (2019). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*, *105*, 1257.

Prins, S. J., & Reich, A. (2017). Can we avoid reductionism in risk reduction?: *Theoretical Criminology*. https://doi.org/10.1177/1362480617707948

Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Liability for Defective Products (2022). https://single-market-economy.ec.europa.eu/system/files/2022-09/COM_2022_495_1_EN_ACT_part1_v6.pdf

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Pruss, D. (2023). Ghosting the Machine: Judicial Resistance to a Recidivism Risk Assessment Instrument. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 312–323. https://doi.org/10.1145/3593013.3593999

Puwar, N. (1997). Reflections on Interviewing Women MPs. *Sociological Research Online*, *2*(1), 82–91. https://doi.org/10.5153/sro.19

Queally, J. (2018, September 23). *California police uphold few complaints of officer misconduct and investigations stay secret*. Los Angeles Times.

https://www.latimes.com/local/lanow/la-me-police-misconduct-complaints-20180923-story.html

Reidenberg, J. (1997). Lex Informatica: The Formulation of Information Policy Rules through Technology. *Tex. L. Rev.*, *76*, 553.

Ricaurte, P. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, *20*(4), 350–365.

Richards, D. (1996). Elite Interviewing: Approaches and Pitfalls. *Politics*, *16*(3), 199–204. https://doi.org/10.1111/j.1467-9256.1996.tb00039.x

Richardson, R., Schultz, J., & Crawford, K. (2019a). *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice* (SSRN Scholarly Paper ID 3333423). Social Science Research Network. https://papers.ssrn.com/abstract=3333423

Richardson, R., Schultz, J., & Crawford, K. (2019b). *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice* (SSRN Scholarly Paper 3333423). https://papers.ssrn.com/abstract=3333423

Rios, E. (2023, June 29). What was affirmative action designed to do – and what has it achieved? *The Guardian*. https://www.theguardian.com/law/2023/jun/22/what-is-affirmative-action-supreme-court-explainer

Roberts, H., Ziosi, M., Osborne, C., Saouma, L., Belias, A., Buchser, M., Casovan, A., Kerry, C., Meltzer, J., Mohit, S., Ouimette, M.-E., Renda, A., Stix, C., Teather, E., Woodhouse, R., & Zeng, Y. (2023). *A Comparative Framework for AI Regulatory Policy*. https://ceimia.org/wp-content/uploads/2023/02/Comparative-Framework-for-AI-Regulatory-Policy.pdf

Robinson Yu. (2014). *Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace*. https://perma.cc/C5PJ-3H3G

Roeder, O. (2017, October 18). *The Supreme Court Is Allergic To Math | FiveThirtyEight*. https://web.archive.org/web/20171018073546/https://fivethirtyeight.com/features/the-supreme-court-is-allergic-to-math/?ex_cid=538fb

Roh, T., & Nam, J. E. (2023). *South Korea: Legislation on Artificial Intelligence to Make Significant Progress - Kim & Chang*. https://www.kimchang.com/en/insights/detail.kc?sch_section=4&idx=26935

Rotaru, V., Huang, Y., Li, T., Evans, J., & Chattopadhyay, I. (2022). Event-level prediction of urban crime reveals a signature of enforcement bias in US cities. *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-022-01372-0

Roulston, K., & Choi, M. (2018). *The SAGE Handbook of Qualitative Data Collection*. SAGE Publications Ltd. https://doi.org/10.4135/9781526416070

Rovatsos, M., Mittelstadt, B., & Koene, A. (2019). *Landscape Summary: Bias in Algorithmic Decision-Making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it?* https://www.research.ed.ac.uk/en/publications/landscape-summary-bias-in-algorithmic-decision-making-what-is-bia

Rowland, S. P., Fitzgerald, J. E., Lungren, M., Lee, E. (Hsieh), Harned, Z., & McGregor, A. H. (2022). Digital health technology-specific risks for medical malpractice liability. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00698-3

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Rueda, J., Delgado, J., Parra Jounou, I., Hortal Carmona, J., Ausín, T., & Rodríguez-Arias, D. (2022). *"Just" accuracy? Procedural fairness demands explainability in AI-based medical resource allocations*.

Sakiotis, E., Oberschelp de Meneses, A., Shepherd, N., & Van Quathem, K. (2023, January 27). *Brazil's Senate Committee Publishes AI Report and Draft AI Law*. Inside Privacy. https://www.insideprivacy.com/emerging-technologies/brazils-senate-committee-publishes-ai-report-and-draft-ai-law/

Salkind, N. (2010). *Encyclopedia of Research Design*. SAGE Publications, Inc. https://doi.org/10.4135/9781412961288

Sandberg, J., & Alvesson, M. (2011). Ways of constructing research questions: Gap-spotting or problematization? *Organization*, *18*(1), 23–44. https://doi.org/10.1177/1350508410372151

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

Selbst, A. D. (2017). Disparate Impact in Big Data Policing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2819182

Shapiro, A. (2019). Predictive Policing for Reform? Indeterminacy and Intervention in Big Data Policing. *Surveillance & Society*, *17*(3/4), 456–472. https://doi.org/10.24908/ss.v17i3/4.10410

Shapley, L. S. (1951). *A Value for N-Person Games*. RAND Corporation. https://www.rand.org/pubs/papers/P295.html

Silberg, J., & Manyika, J. (2019). *Notes from the AI frontier: Tackling bias in AI (and in humans)*. McKinsey Global Institute. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf

Simkin, M. (2023, May 10). Looking Forward to the End of Chicago's Contract with ShotSpotter. *Chicago Appleseed Center for Fair Courts*. https://www.chicagoappleseed.org/2023/05/10/ending-chicago-contract-with-shotspotter/

Simonite, T. (2019). Algorithms Should've Made Courts More Fair. What Went Wrong? *Wired*. https://www.wired.com/story/algorithms-shouldve-made-courts-more-fair-what-went-wrong/

Simonson, J. (2021). *Police Reform Through a Power Lens*. https://www.yalelawjournal.org/article/police-reform-through-a-power-lens

Skeem, J., Scurich, N., & Monahan, J. (2020). Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants. *Law and Human Behavior*, *44*. https://doi.org/10.1037/lhb0000360

Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*, *5*(8), 873–883. https://doi.org/10.1038/s42256-023-00692-8

Small, M. L. (2009). `How many cases do I need?': On science and the logic of case selection in field-based research. *Ethnography*, *10*(1), 5–38. https://doi.org/10.1177/1466138108099586

Solly, M. (2020). *158 Resources for Understanding Systemic Racism in America*. Smithsonian Magazine. https://www.smithsonianmag.com/history/158-resources-understanding-systemic-racism-america-180975029/

Solon Barocas (Director). (2022, August 19). *SRA22 Day 3—Keynote Talk with Solon Barocas* [Video recording]. https://www.youtube.com/watch?v=Ft5rK1tTYyw

Spencer, K. B., Charbonneau, A. K., & Glaser, J. (2016). Implicit bias and policing. *Social and Personality Psychology Compass*, *10*(1), 50–63.

Stevenson, M. T. (2017). Assessing Risk Assessment in Action. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3016088

Stratton, S. J. (2019). Literature Reviews: Methods and Applications. *Prehospital and Disaster Medicine*, *34*(4), 347–349. https://doi.org/10.1017/S1049023X19004588

Stroud, M. (2021, May). An automated policing program got this man shot twice. In *The Verge*. https://www.theverge.com/c/22444020/heat-listed-csk-entry

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, *41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x

Student Borrower Protection Center. (2020). Educational Redlining. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3759925

Sundararajan, M., & Najmi, A. (2020). The Many Shapley Values for Model Explanation. *Proceedings of the 37th International Conference on Machine Learning*, 9269–9278. https://proceedings.mlr.press/v119/sundararajan20b.html

Tafoya, S., & Nguyen, V. (2015). *California's Criminal Courts*. Public Policy Institute of California. https://www.ppic.org/publication/californias-criminal-courts/

Tappolet, C. (2004). Through thick and thin: Good and its determinates. *Dialectica*, *58*(2), 207–221. https://doi.org/10.1111/j.1746-8361.2004.tb00297.x

Tashakkori, A., & Teddlie, C. (2003). *Handbook of Mixed Methods in Social & Behavioral Research*. SAGE.

Tellis, W. (1997). Introduction to Case Study. *The Qualitative Report*, *3*(2), 1–14. https://doi.org/10.46743/2160-3715/1997.2024

The Federal Reserve Board. (2007). *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*. 304.

Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. *Patterns*, *3*(5), 100476. https://doi.org/10.1016/j.patter.2022.100476

Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022 (2022). https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/

Translation: Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment) – April 2023 (2023). https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/

Umansky, E. (2023). The Failed Promise of Police Body Cameras. *The New York Times*. https://www.nytimes.com/2023/12/13/magazine/police-body-cameras-miguel-richards.html

UNESCO. (2020). *Toward a draft text of a Recommendation on the Ethics of Artificial Intelligence: Working document—UNESCO Digital Library*.

https://unesdoc.unesco.org/ark:/48223/pf0000373199?posInSet=2&queryId=N-EXPLORE-f810189f-29ee-44f5-9eee-5b94da449648

US Census Bureau. (n.d.). *About the Topic of Race*. Census.Gov. Retrieved February 29, 2024, from https://www.census.gov/topics/population/race/about.html

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable Recourse in Linear Classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. https://doi.org/10.1145/3287560.3287566

Utz, C., Degeling, M., Fahl, S., Schaub, F., & Holz, T. (2019). *(Un)informed Consent | Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. https://dl.acm.org/doi/10.1145/3319535.3354212

Uuk, R. (2024). *AI Act Implementation: Timelines & Next steps | EU Artificial Intelligence Act*. https://artificialintelligenceact.eu/ai-act-implementation-next-steps/

Van Craen, M., & Skogan, W. G. (2017). Achieving fairness in policing: The link between internal and external procedural justice. *Police Quarterly*, *20*(1), 3–23.

VanNostrand, P. M., Hofmann, D. M., Ma, L., & Rundensteiner, E. A. (2024). Actionable Recourse for Automated Decisions: Examining the Effects of Counterfactual Explanation Type and Presentation on Lay User Understanding. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1682–1700. https://doi.org/10.1145/3630106.3658997

Väyrynen, P. (2021). Thick Ethical Concepts. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entrieshick-ethical-concepts/

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 284–293. https://doi.org/10.1145/3351095.3372876

Verma, P. (2023). Heat Listed. *The Washington Post*. https://www.washingtonpost.com/technology/2022/07/15/predictive-policing-algorithms-fail/

Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual Explanations for Machine Learning: A Review. *arXiv:2010.10596 [Cs, Stat]*. http://arxiv.org/abs/2010.10596

Volpicelli, G. (2024, February 2). *EU countries strike deal on landmark AI rulebook*. POLITICO. https://www.politico.eu/article/eu-countries-strike-deal-ai-law-act-technology/

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)*, *31*(2), 841–888.

Wachter, S., Mittelstadt, B., & Russell, C. (2021a). *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law* (SSRN Scholarly Paper ID 3792772). Social Science Research Network. https://doi.org/10.2139/ssrn.3792772

Wachter, S., Mittelstadt, B., & Russell, C. (2021b). *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law* (SSRN Scholarly Paper ID 3792772). Social Science Research Network. https://doi.org/10.2139/ssrn.3792772

Wallin, D. E. (1992). Legal Recourse and the Demand for Auditing. *The Accounting Review*, *67*(1), 121–147.

Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans. Knowl. Discov. Data*, *17*(3), 35:1-35:27. https://doi.org/10.1145/3551390

Wang, J., Wiens, J., & Lundberg, S. (2021). *Shapley Flow: A Graph-based Approach to Interpreting Model Predictions* (arXiv:2010.14592). arXiv. https://doi.org/10.48550/arXiv.2010.14592

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, *200*(2), 65. https://doi.org/10.1007/s11229-022-03485-5

Webster, R. A. (2021, November 22). *"If Everybody's White, There Can't Be Any Racial Bias": The Disappearance of Hispanic Drivers From Traffic Records*. ProPublica. https://www.propublica.org/article/if-everybodys-white-there-cant-be-any-racial-bias-the-disappearance-of-hispanic-drivers-from-traffic-records

Weiss, R. S. (1994). *Learning from Strangers: The Art and Method of Qualitative Interview Studies*. Free Press.

Weld, S. (2021, December 22). *Samuelson Clinic a leader in teaching public defenders about technology*. Berkeley Law. https://www.law.berkeley.edu/article/teaching-tech-berkeley-law-clinic-arms-public-defenders-with-digital-tools/

White, M. D., Fradella, H. F., & Flippin, M. (2021). How can we achieve accountability in policing? The (not-so-secret) ingredients to effective police reform. *Lewis & Clark L. Rev.*, *25*, 405.

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society*, *10*(1), 20539517231177620.

Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022). Limits and possibilities for "Ethical AI" in open source: A study of deepfakes. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2035–2046.

Wikström, P.-O. H., & Treiber, K. (2017). Beyond Risk Factors: An Analytical Approach to Crime Prevention. In B. Teasdale & M. S. Bradley (Eds.), *Preventing Crime and Violence* (pp. 73–87). Springer International Publishing. https://doi.org/10.1007/978-3-319-44124-5_8

Williams, B. (1985). *Ethics and the Limits of Philosophy—Bernard Williams—Google Books*. https://books.google.it/books/about/Ethics_and_the_Limits_of_Philosophy.html?id=CNCvqVQon9kC&redir_esc=y

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136.

Wiseman, J. P. (1979). Toward a Theory of Policy Intervention in Social Problems. *Social Problems*, *27*(1), 3–18. https://doi.org/10.2307/800013

Wu, Y., Zhang, L., & Wu, X. (2019). *Counterfactual Fairness: Unidentification, Bound and Algorithm*. 1438–1444.

Xenidis, R. (2018). *Multiple discrimination in EU anti-discrimination law: Towards redressing complex inequality?* https://doi.org/10.5040/9781509915033.ch-002

Xiang, A. (2020). *Intel's Artificial Intelligence Podcast*. Intel. https://www.intel.com/content/www/us/en/artificial-intelligence/podcast-episodes/artificial-intelligence-podcast-episode-12.html

Xu, H., & Zhang, N. (2020). *Implications of Data Anonymization on the Statistical Evidence of Disparity* (SSRN Scholarly Paper 3662612). https://doi.org/10.2139/ssrn.3662612

Yeung, D., Khan, I., Kalra, N., & Osoba, O. A. (2021). *Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications*. RAND Corporation. https://www.rand.org/pubs/perspectives/PEA862-1.html

Yin, R. K. (1984). *Case Study Research: Design and Methods*. SAGE Publications.

Zahn, E. M. T. (2021). A Local Solution: Utilizing County-Level Statistics to Redress Racial Bias in Capital Sentencing. *Ohio State Journal of Criminal Law*, *19*, 219.

Zajko, M. (2021). Conservative AI and social inequality: Conceptualizing alternatives to bias through social theory. *AI & SOCIETY*, *36*(3), 1047–1056. https://doi.org/10.1007/s00146-021-01153-9

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning* (arXiv:1801.07593). arXiv. https://doi.org/10.48550/arXiv.1801.07593

Zhou, J., Chen, F., & Holzinger, A. (2022). Towards Explainability for AI Fairness. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 375–386). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_18

Zilka, M., Fogliato, R., Hron, J., Butcher, B., Ashurst, C., & Weller, A. (2023). The Progression of Disparities within the Criminal Justice System: Differential Enforcement and Risk Assessment Instruments. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1553–1569. https://doi.org/10.1145/3593013.3594099

Ziosi, M., Mökander, J., Novelli, C., Casolari, F., Taddeo, M., & Floridi, L. (2023). The EU AI Liability Directive (AILD): Bridging Information Gaps. *European Journal of Law and Technology*, *14*(3). https://ejlt.org/index.php/ejlt/article/view/962

Ziosi, M., Watson, D., & Floridi, L. (2024). *A Genealogical Approach to Algorithmic Bias* (SSRN Scholarly Paper 4734082). https://doi.org/10.2139/ssrn.4734082

APPENDIX A - DISCLAIMER FROM LUCIANO FLORIDI

07 March 2024

Subject: Collaboration with Marta Ziosi

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Marta was the sole first author of the following two articles:

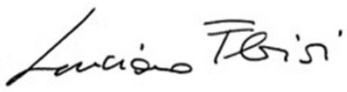
Ziosi, M., Mökander, J., Novelli, C., Casolari, F., Taddeo, M., & Floridi, L. (2023). The EU AI Liability Directive (AILD): Bridging Information Gaps. *European Journal of Law and Technology*, *14*(3).

Ziosi, Marta and Watson, David and Floridi, Luciano, A Genealogical Approach to Algorithmic Bias (February 21, 2024). Available at SSRN: https://ssrn.com/abstract=

As her supervisor, I gave input to Marta's ideas, gave feedback on the presentation of the arguments, and edited the manuscripts. Marta had the idea for the studies, conducted the research, and drafted the manuscripts. I am happy for Marta to include each of the above articles as chapters in her thesis and present the material as her own ideas.

Sincerely,

Luciano Floridi



Founding Director of the Digital Ethics Center

Professor in the Practice in the Cognitive Science Program

Faculty of Arts and Sciences

Yale University

Email: luciano.floridi@yale.edu

APPENDIX B - DISCLAIMER FROM DAVID WATSON

APPENDIX – DISCLAIMER FROM DAVID WATSON

London, 27 February 2024

Subject: Collaboration with Marta Ziosi

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Marta Ziosi was the sole first author of the following article:

Ziosi, Marta and Watson, David and Floridi, Luciano, A Genealogical Approach to Algorithmic Bias (February 21, 2024). Available at SSRN, abstract ID 4734082.

Marta formulated the research question, conducted the research for the study, produced the first draft of the manuscript, and coordinated the project. As one of the co-authors, my main contribution was to validate the findings. Given my background in ML and statistics, I focused primarily on the technical aspects of the article. I also proofread the final manuscript before it was submitted for publication. That said, Marta assumed most of the workload and I have no objection to her including a lightly edited version of the article as a chapter in his DPhil thesis.

Sincerely,

David Watson

David Watson
Lecturer in Artificial Intelligence
Department of Informatics
King's College London
Bush House, 30 Aldwych
London WC2B 4BG

Email: david.watson@kcl.ac.uk

APPENDIX C – DISCLAIMER FROM DASHA PRUSS

Oxford, 23 February 2024

Subject: Collaboration with Marta Ziosi

Dear Oxford Graduate Studies Committee/Examiners,

I, Dasha Pruss, was one of the co-authors of the article:

"*Evidence of What, for Whom? The Socially Contested Role of Algorithmic Bias in a Predictive Policing Tool*"

I hereby confirm that Marta is the first author, given her contribution to the conceptualization and design of the study, collecting and analyzing the data, and producing a first draft. I am happy for Marta to include this paper as part of her DPhil thesis.

Sincerely,



Dasha Pruss, PhD

Fellow, Berkman Klein Center for Internet & Society

Postdoctoral Fellow, Embedded EthiCS Program

Harvard University

Email: dpruss@fas.harvard.edu

APPENDIX D - DISCLAIMER FROM MARIAROSARIA TADDEO

APPENDIX – DISCLAIMER FROM MARIAROSARIA TADDEO

Oxford, 21 February 2024

Subject: Collaboration with Marta Ziosi

Dear Oxford Graduate Studies Committee/Examiners,

I, Mariarosaria Taddeo, was one of the co-authors of the article:

Ziosi, M., Mökander, J., Novelli, C., Casolari, F., Taddeo, M., & Floridi, L. (2023). The EU AI Liability Directive (AILD): Bridging Information Gaps. *European Journal of Law and Technology*, *14*(3).

My main contributions to the paper were providing contextual background and revising the final draft of the manuscript. Marta had the idea for the study, provided its central framing, and was its lead author. I am happy for Marta to include this paper as part of her DPhil thesis.

Sincerely,



Mariarosaria Taddeo

Professor of Digital Ethics and Defence Technologies
ICSS DPhil Programme Director
Oxford Internet Institute
University of Oxford

41 St. Giles', Oxford OX1 3LW

Email: mariarosaria.taddeo@oii.ox.ac.uk

APPENDIX E - DISCLAIMER FROM JAKOB MOKANDER

Oxford, 21 February 2024

Subject: Collaboration with Marta Ziosi

Dear Oxford Graduate Studies Committee/Examiners,

This letter is to certify that Marta was the sole first author of the following article:

Ziosi, M., Mökander, J., Novelli, C., Casolari, F., Taddeo, M., & Floridi, L. (2023). The EU AI Liability Directive (AILD): Bridging Information Gaps. *European Journal of Law and Technology*, *14*(3).

As one of the co-authors of this article, I gave input to Marta's ideas, gave feedback on the presentation of the arguments, and edited the manuscript. Marta had the idea for the study, provided its central framing, wrote the first draft of the paper, and was its lead author. I am happy for Marta to include this paper as part of her DPhil thesis.

Sincerely,

Jakob Mökander (signed by email)

Risk & Resilience Specialist at McKinsey
International Fellow at Yale Digital Ethics Center
Yale University

Email: jakob.mokander@outlook.com

APPENDIX F - DISCLAIMER FROM CLAUDIO NOVELLI

Oxford, 21 February 2024

Subject: Collaboration with Marta Ziosi

Dear Oxford Graduate Studies Committee/Examiners,

I, Claudio Novelli, was one of the co-authors of the article:

Ziosi, M., Mökander, J., Novelli, C., Casolari, F., Taddeo, M., & Floridi, L. (2023). The EU AI Liability Directive (AILD): Bridging Information Gaps. *European Journal of Law and Technology*, *14*(3).

Given my legal background, my main contributions to the paper were checking the legal accuracy, providing legal background, and revising the final draft of the manuscript. Marta had the idea for the study, provided its central framing, wrote the paper and was its lead author. I am happy for Marta to include this paper as part of her DPhil thesis.

Sincerely,

Claudio Novelli (signed by email)

Adjunct Professor
Department of Legal Studies
University of Bologna

Email: claudio.novelli@unibo.it

APPENDIX G - DISCLAIMER FROM FEDERICO CASOLARI

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA | DEPARTMENT
OF LEGAL STUDIES

APPENDIX – DISCLAIMER FROM FEDERICO CASOLARI

Bologna, 21 February 2024

Subject: Collaboration with Marta Ziosi

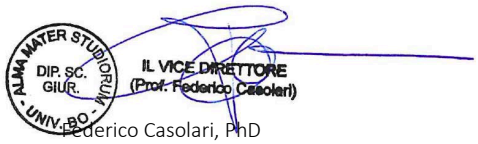
Dear Oxford Graduate Studies Committee/Examiners,

I, Federico Casolari, was one of the co-authors of the article:

Ziosi, M., Mökander, J., Novelli, C., Casolari, F., Taddeo, M., & Floridi, L. (2023). The EU AI Liability Directive (AILD): Bridging Information Gaps. European Journal of Law and Technology, 14(3).

Given my legal background, my main contribution was to validate the accuracy of the arguments from a legal standpoint, adding legal references and proofreading the manuscript. Marta had the idea for the study, gathered and analysed the data, produced a first draft, and was its lead author. I am happy for Marta to include this paper as part of her DPhil thesis.

Sincerely,



Federico Casolari, PhD
Full Professor of European Union Law
Deputy Head, Department of Legal Studies
Alma Mater Studiorum - Università di Bologna
via Zamboni, 27/29 | I - 40126 Bologna
tel. + 39.051.2099651 | fax + 39.051.2086038
WEB: https://www.unibo.it/sitoweb/federico.casolari

DIREZIONE
Via Zamboni 22 | 40126 Bologna | Italia | Tel. + 39 051 2099651 | federico.casolari@unibo.it

APPENDIX H – QUESTION SHEET INTERVIEWS

1. **Context and positionality**
   Probe:
   a. Can you introduce yourself? Can you tell me more about your work?
   b. How are you related to the predictive policing project? How have you heard of it, if at all?
   c. And, if not, are you involved with other initiatives around public safety in Chicago?
2. **Bias definition and utility**
   Probe:
   a. What is algorithmic bias to you in general? And enforcement bias specifically?
   b. What is the problem with it? Why do you think it is there?
   c. What is its specific relevance to your work? And the context of Chicago?
   d. What is your role in tackling bias and problems faced in doing so (if applicable) at work?
3. **Perceived feasibility and desiderability using evidence of bias to tackle societal disparities**
   Probe:
   a. How should it be tackled? What would an ideal solution look like to you (especially with relevance to your role/work)?
   b. What are the problems in solving it? How can it be realistically tackled in your opinion?
4. **Snowball sampling**
   Probe:
   a. Do you know other people who might be relevant for me to interview?

APPENDIX I – CODE TABLE

Abolition

Alternatives to current system/police
Building healthy communities
Calls for changing the narrative/status quo
The need to acknowledge harms of current system
Importance of history and education
Importance of hope and belief in change

Co-option and exploitation

Academic exploitation
Co-opting evidence-based label
Commercial interest in predicting crime
Disruption rhetoric
Open source as a marketing strategy
Responsibilities of developers

Downsides of AI

AI is racist/not objective
Inherent reservations about AI
Garbage in garbage out/distrust of data
Human discretion is important
Nuance not captured by algorithm
Technology isn't going to fix policing

Lived experience and positionality

Affected people are best suited to solve problems
Barriers to entering research
Impacted groups lack resources
Different stakeholders care about different forms of bias
People not affected don't have the same understanding
Who generated the data/designed the tool matters
Importance of trust

Reformism

Interpersonal bias rather than structural
Resource allocation
Safety is the priority
Police promotes safety

Structural bias

Importance of addressing root cause of crime
Racial profiling and segregation
Structural racism requires a structural solution
The system is biased

Upsides of AI

Algorithm can be more objective/humans are biased
Algorithms increase efficiency
Open source
Using data to understand authority figures (sousveillance)
Information has multiple uses