S. Barocas, M. Hardt and A. Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023

:: **Leonardo De Pin**

Abstract
This review discusses Fairness and Machine Learning by Barocas, Hardt, and Narayanan, highlighting its comprehensive and interdisciplinary treatment of fairness in algorithmic decision-making. The book is accessible, open access, and enriched by insights from ethics, law, and the social sciences.

Keywords
Automated decision-making; Fairness; Machine Learning

Submitted 01/10/2025; **Accepted** 06/01/2026

How to Cite
Leonardo De Pin. *S. Barocas, M. Hardt and A. Narayanan. Fairness and Machine Learning: Limitations and Opportunities. MIT Press, 2023*. The Reasoner 20(1) 2026. https://doi.org/10.54103/1757-0522/29851

Automated decision-making systems based on machine learning are increasingly deployed in high-stakes areas such as criminal justice (risk assessments), hiring (automated résumé screening), healthcare (predictive diagnostics), and others. These developments raise questions about their legitimacy, advantages, and limitations regarding fairness and discrimination—these are precisely the issues tackled by Barocas, Hardt, and Narayanan in *Fairness and Machine Learning: Limitations and Opportunities*.

The appeal of automating decision-making lies in the promise of more accurate and efficient outcomes. Such systems can follow rules, learn from past decisions, and detect new patterns in data. For example, résumé-screening tools predict job performance and reduce arbitrariness in hiring. Yet they may also perpetuate inequalities encoded in training data or introduce new biases through model design—such as penalizing equally qualified women if historical gender bias is present. Even seemingly relevant predictors like education can raise concerns, since they

reflect structural factors such as family income. In sensitive domains like criminal sentencing, the very legitimacy of delegating decisions to machines is open to question.

Often, these concerns are addressed through technical solutions or formal definitions of fairness, which are frequently very abstract. However, these issues go beyond technical features of the system and involve moral, legal, and domain-specific considerations. Each context has different discrimination mechanisms, penalized categories and objectives. The authors capture this complexity effectively, offering a comprehensive treatment of fairness: problems, risks, formal criteria, connections to moral and legal concepts, analyses of discrimination mechanisms, and more.

The book structure is straightforward, with chapters grouped into four main themes: the first two provide the motivation and an overview of fairness issues; Chapters 3-5 develop the technical background; Chapters 6-7 adopt a more practical stance, illustrating how fairness is addressed in law and tested; the last two chapters offer broader reflections on discrimination and datasets.

The opening chapters introduce the debate by showing why fairness in machine learning is important and outlining the decision-making contexts where fairness concerns arise.

Chapter 1 presents the book's central concern: the impact on group disparities when decision-making is delegated to machine learning. To better explain how demographic disparities can propagate, the authors introduce the "machine learning loop", a conceptual model with four stages: measuring the state of the world in data, learning to turn data into a model, acting on the model's predictions, and feeding back the results, to alter the world. Biases can arise at any stage, and understanding these dynamics and mitigating them is complex.

Chapter 2 examines the legitimacy of automated decision-making systems, asking whether their use is fair. The authors argue that machine learning is used to automate bureaucratic processes designed to limit arbitrariness through formal procedures. They focus on concerns posed by predictive optimization—a particular form of automation where decision rules are learned directly from data— illustrating cases in which the legitimacy of automated decision making can be questioned.

The Reasoner 20(1), 2026. https://doi.org/10.54103/1757-0522/29851

Chapters 3–5 provide technical foundations, presenting measures, definitions, and models of fairness and their trade-offs.

Chapter 3 focuses on classification, the task of predicting which category an input belongs to, based on patterns learned from labeled data. The authors formally characterize classification and introduce three statistical fairness criteria: independence (group membership is independent of the classification's outcome), separation (error rates are the same across groups), and sufficiency (the outcome is independent of group membership given the prediction). These criteria are then shown to be incompatible: satisfying one often means violating the others. The chapter ends with a discussion of the inherent limitations of statistical criteria, one being that they do not reveal the causal mechanisms through which disparities are created.

Chapter 4 explores the moral landscape behind fairness and discrimination, discussing why discrimination is wrong and illustrating three views of equality of opportunity, often conceived as the foundation of fairness. The three views—narrow, middle and broad—are assessed by highlighting the tensions between them, their strengths and implications. Most importantly, the chapter considers whether statistical notions of fairness can be normatively justified by connecting them to views of equality of opportunity. While some alignment is possible, the correspondence is loose: moral notions require a causal understanding of disparities, something statistical approaches lack.

Chapter 5 addresses these limitations by introducing causal reasoning tools. Using structural causal models, the authors describe interventions, control for confounders, and address counterfactual questions such as "Would the applicant have been rejected had she been of a different race?" These tools are used to analyze causal and counterfactual questions about discrimination and to introduce counterfactual fairness criteria. The chapter concludes with an assessment of causal modeling's validity, noting its advantages and limitations.

Chapter 6 reviews the main anti-discrimination laws and doctrines, such as disparate treatment and disparate impact, outlining their history, uses, and limitations. It shows how these frameworks can inform the regulation of machine learning but argues that they are insufficient, suggesting that privacy and consumer protection law should also play a role.

The Reasoner 20(1), 2026. https://doi.org/10.54103/1757-0522/29851

THE REASONER 20(1), 2026. https://doi.org/10.54103/1757-0522/29851

Chapter 7 surveys traditional methods for testing discrimination, from audit studies to regression analyses, and shows their application to machine learning in different domains: natural language processing, computer vision, and online platforms. These examples illustrate that unfairness arises in subtle, context-dependent ways, reinforcing the idea that there is no single fairness test and its choice must follow moral and domain-specific considerations.

Chapter 8 broadens the scope by analyzing discrimination at three levels——structural, organizational, and interpersonal. It highlights how laws, policies, and education shape structural inequalities, which can be reinforced by machine learning systems. The authors propose interventions at both structural and organizational levels, stressing the need to move beyond narrow views of fairness that focus only on minimizing numerical disparities.

Chapter 9 turns to datasets, describing some widely used examples and the multiple roles they play in machine learning. The authors examine the "benchmark approach" in building and using datasets, noting its benefits but also its risks, and emphasize that problems stem not only from data but also from the cultural and normative practices surrounding them. The chapter closes with proposals for more responsible dataset practices.

After reviewing the main contents of the book, several strengths stand out. First, it offers a broad and comprehensive account of fairness in machine learning. The authors not only define fairness and explain its technical implementation, but also raise deeper questions: Why is discrimination wrong? What effects does it have? What moral justifications support competing definitions of fairness, and how should we choose among them?

Second, the authors emphasize the importance of context in fairness interventions. A fairness criterion applied through a technical adjustment or validated by a test does not guarantee that a system is fair. Discrimination is deeply rooted in culture and society, and fairness interventions can generate feedback effects——for instance, adjusting thresholds in credit scoring may reshape lending patterns, resulting in new data used for future models. Fairness is thus not a static property of a model but requires continuous monitoring and long-term evaluation. Meaningful interventions must be context-sensitive, guided by moral reasoning and domain-specific considerations.

No single book can exhaust the topic, and as noted by Cohen, J. and Liu, L.(2025:

The Reach of Fairness, *ACM J. Responsib. Comput.* 10.1145/3718989), the discussion could be extended. Barocas, Hardt, and Narayanan focus mainly on organizational decision-making, devoting less attention to structural interventions. Cohen and Liu also argue that fairness should not be reduced to equality of opportunity, the notion on which the book primarily focuses, but should be understood more broadly.

Beyond these limitations, two further strengths deserve emphasis. First, the book is multidisciplinary, drawing on ethics, law, the social sciences, and computer science. This breadth makes it appealing to a wide audience and situates the technical material in its normative and institutional context. Second, it is open access, ensuring that its insights are freely available to researchers, practitioners, and policymakers alike.

Overall, the book is accessible without oversimplifying. The authors guide the reader step by step, though some chapters—particularly 3 and 5—are more technically demanding, and non-experts may need to approach them slowly. Yet the combination of an introductory style, interdisciplinary reach, and open-access availability makes it a valuable resource for a wide audience and strongly recommended to anyone interested in the ethical challenges of machine learning.

Acknowledgments This work was supported by the project "Ethical Design for AI", part of the Spoke 6 of the NRRP project FAIR (PE00000013, CUP H97G22000210007), funded by the European Union-NextGenerationEU and the Italian Ministry of University and Research (MUR).

Leonardo De Pin
https://orcid.org/0009-0002-0573-1163
Università del Salento