Tshilidzi Marwala

The Balancing Problem in the Governance of Artificial Intelligence

Springer

The Balancing Problem in the Governance of Artificial Intelligence

Tshilidzi Marwala

The Balancing Problem in the Governance of Artificial Intelligence

Springer

Tshilidzi Marwala
United Nations University
Shibuya, Tokyo, Japan

ISBN 978-981-97-9250-4 ISBN 978-981-97-9251-1 (eBook)
https://doi.org/10.1007/978-981-97-9251-1

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Within the dynamic realm of Artificial Intelligence (AI), we encounter many intricate and interconnected obstacles that need deliberate and effective regulation. This governance challenge involves finding a careful balance between innovation and caution.

The Balancing Problem in the Governance of Artificial Intelligence explores the tradeoffs that shape AI governance in the 21st century. The dichotomy between memorization and thinking is seen in AI's ability to store extensive information while showcasing its capacity for innovative problem-solving. It involves studying the contrast between quick, intuitive thinking and deliberate, intricate assessments that influence our communities.

The book delves into the effect of divergences, such as AI opportunity seeking against risk aversion, ethical challenges of transparency versus truth, and the complex relationship between synthetic and real data on the creation and implementation of AI technology. This book explores the complex problem of algorithmic discrimination, specifically focusing on what aspects may be prevented and which are unavoidable. It also discusses the several types of computational paradigms that fuel AI, such as the central processing unit (CPU) against the graphics processing unit (GPU), quantum computing versus digital computing, and the decentralized architectures of embedded, edge, and cloud computing.

Furthermore, the book examines the regulatory frameworks that regulate AI, including standards and laws as opposed to policies and regulations, and the ongoing discussion about whether self-regulation or government monitoring is more appropriate. The book explores the international aspects of AI governance, addressing the tradeoff between globalization and localization in formulating AI laws that influence society globally.

As we traverse this terrain, it becomes evident that successful AI governance involves technological proficiency and a deep comprehension of cultural values, ethical deliberations, and power dynamics. This book aims to stimulate reflection, encourage conversation, and give a direction for responsible AI governance.

Tokyo, Japan Tshilidzi Marwala
June 2024

Acknowledgments

Authoring *The Balancing Problem in the Governance of Artificial Intelligence* has been an extraordinary journey, and I appreciate those who have supported and contributed to this book. My family, Jabulile, Khathutshelo, Thendo and Denga, have been the bedrock of my journey. Their steadfast support and encouragement, unique to each of them, have provided a solid base, enabling me to wholeheartedly follow my passion for delving into the intricacies of artificial intelligence (AI). I also thank my mother, Mrs. Regina Khathutshelo Marwala, for her guidance.

I am grateful to my research assistants, who showed unwavering dedication in collecting data, examining patterns, and offering essential assistance. Your diligent efforts and tireless commitment have been crucial in completing this text. I want to thank the several industry experts and practitioners who graciously offered their experiences and viewpoints. Your practical insights, which were instrumental in connecting theoretical concepts with real-world applications, have made this book not just informative, but also relevant and valuable for anyone at the forefront of AI research and deployment.

Lastly, I express my gratitude to the broader community of AI researchers, ethicists, and activists who strive to tackle the complexities and possibilities AI brings. Your combined efforts have been a great source of inspiration for me and profoundly impacted the ideas and solutions covered in this book. Your work has been worthwhile, and I truly appreciate your contributions.

I prepared this book using Google Gemini to search for information. I thank Ms. Sunita Menon for proofreading this book.

August 2024 Tshilidzi Marwala

Contents

About the Author

Tshilidzi Marwala is an artificial intelligence (AI) engineer, currently serving as the Rector of the United Nations University. He has contributed significantly to AI applications in engineering, economics, and education. Previously the President of the University of Johannesburg, he championed the integration of advanced technologies into academia, emphasizing the role of AI in addressing global challenges. With a PhD in AI from the University of Cambridge and numerous publications, Marwala is a prominent advocate for responsible AI development. He is also a Fellow of the American Academy of Arts and Sciences and the Chinese Academy of Sciences.

Acronyms

$\boldsymbol{A}$	Avoidable discrimination or Responsibility function
AI	Artificial Intelligence
ALU	Arithmetic/Logic Unit
APUs	Accelerated Processing Units
ARM	Advanced RISC Machine
ATMs	Automated Teller Machines
AWS	Amazon Web Services
BERT	Bidirectional Encoder Representations from Transformers
$\boldsymbol{C}$	Cost or expenses function or Cloud computing or Adherence
CCPA	California's Consumer Privacy Act
$\boldsymbol{C_{cpu}}$	Cost of CPU
CDNs	Content Delivery Networks
$\boldsymbol{C_{gpu}}$	Cost of GPU
CPUs	Central Processing Units
CRM	Customer Relationship Management
CSR	Corporate Social Responsibility
CTPA	Computed Tomography Pulmonary Angiography
CU	Control Unit
$\boldsymbol{D}$	Negative consequences function or Digital Computing
$\boldsymbol{Dc}$	Deception function
$\boldsymbol{D_m}$	Measured data
$\boldsymbol{D_s}$	Synthetic data.
$\boldsymbol{E}$	Embedded computing
$\boldsymbol{E}$	Expected function
$\boldsymbol{E_{cpu}}$	Energy Consumption of CPUs
EEG	Electroencephalogram
$\boldsymbol{E_f}$	Effectiveness function
$\boldsymbol{E_{gpu}}$	GPUs Energy Consumption
EHRs	Electronic Health Records
EPA	Environmental Protection Agency
EU	European Union

$\boldsymbol{F}$	Benefits of thinking fast or Adaptability Function
FTC	Federal Trade Commission
$\boldsymbol{G}$	Edge computing or Global governance
GANs	Generative Adversarial Networks
GDDR6	Graphics Double Data Rate 6 (GDDR6)
GDPR	General Data Protection Regulation
GHz	Gigahertz
GPT	Generative Pre-trained Transformer
GPUs	General processing units
HBM2	High Bandwidth Memory 2
HFT	High Frequency Trading
HIPAA	Health Insurance Portability and Accountability Act
HPC	High-performance computing
$\boldsymbol{I}$	Integration Function or flexibility of self-regulation
IaaS	Infrastructure as a Service
ICs	Integrated Circuits
IEEE	Institute of Electrical and Electronics Engineers
IHL	International Humanitarian Law
IoT	Internet-of-Things
ISA	Instruction Set Architecture
ISO	International Organization for Standardization
ITU	International Telecommunication Union
JEDEC	Joint Electron Device Engineering Council
KPIs	Key Performance Indicators
$\boldsymbol{L}$	Dependability or Local Governance
L1	Lasso
L2	Ridge
LLMs	Large Language Models
$\boldsymbol{M}$	Advantages of memorizing or Money saved
MCDA	Multi-Criteria Decision Analysis
ML	Machine Learning
MLE	Maximum Likelihood Estimation
$\boldsymbol{N_A}$	Unavoidable bias function
NLP	Natural Language Processing
PaaS	Platform as a Service
$\boldsymbol{P_{cpu}}$	CPU Performance
PE	Pulmonary Embolism
$\boldsymbol{P_{gpu}}$	GPU Performance
PoS	Point of Sale
$\boldsymbol{Q}$	Quantum
$\boldsymbol{R}$	Performance Enhancement or Accomplishment function
RFID	Radio-Frequency Identification
$\boldsymbol{R_i}$	Risk function
$\boldsymbol{S}$	Advantages of thinking slow or Security or Self-regulation
SaaS	Software as a Service

$\boldsymbol{S_{cpu}}$	CPU scalability
$\boldsymbol{S_{gpu}}$	GPU scalability
SIEM	Security Information and Event Management
SMs	Streaming Multiprocessors
St	Standardization
STEM	Science, Technology, Engineering and Mathematics
T	Transparency function
Th	Thinking function
TPUs	Tensor Processing Units
Tr	Truth function
U	Payoff/Utility/Reward function
$\boldsymbol{W_{cpu}}$	CPU Water Consumption
$\boldsymbol{W_{gpu}}$	GPU Water Consumption
$\boldsymbol{x}$	Choice configuration
XAI	Explainable AI
$\boldsymbol{X_u}$	Improved user experience
$\alpha, \beta, \gamma, \lambda, \omega$	Coefficients (weights) values

Chapter 1
Introduction to the Artificial Intelligence Balancing Problem

Abstract This chapter introduces the Artificial Intelligence (AI) Governance Balancing Problem. While much of the AI governance models govern aspects of AI, such as efficiency, accountability, transparency, security, fairness, or privacy, there is a problem of balancing (trade-off) conflicting aspects of AI, e.g., transparency versus security. The Pareto Optimal Frontier, from economics and decision theory, helps find good trade-offs between competing goals. This paradigm should guide solving the AI balancing problem in designing and deploying AI in applications such as healthcare, conflict resolution, and the financial sector. The AI balancing problem can help solve ethical and practical difficulties for AI development and promote the creation of intelligent, ethically, and socially beneficial AI systems.

Keywords Artificial intelligence · AI governance balancing model · AI governance hierarchy model · Pareto frontier

1.1 Introduction

The AI balancing problem involves the complex task of harmonizing the extensive capabilities of AI with a wide range of human values, ethical issues, and societal requirements (Marwala 2001, 2010, 2014, 2023, 2024a, b; Marwala and Mpedi 2024; Marwala et al. 2023a, b). As AI systems grow more integrated into many areas, such as healthcare, education, security, and entertainment, they are responsible for making judgments that can significantly affect individuals and communities (Mbuvha and Marwala 2020; Russell et al. 2008). This integration poses crucial inquiries about how these systems can be formulated to accurately represent the intricate preferences of the varied populations they cater to, encompassing equity, confidentiality, effectiveness, and security.

The AI balancing problem arises from the understanding that AI systems are built to optimize specific goals. Nevertheless, numerous goals must be considered concurrently, which may occasionally be contradictory. For instance, in healthcare, an AI system may have to balance safeguarding patient confidentiality and the requirement

T. Marwala, *The Balancing Problem in the Governance of Artificial Intelligence*,
https://doi.org/10.1007/978-981-97-9251-1_1

for a thorough collection of data to enhance the accuracy of diagnoses. Similarly, in content recommendation systems, the objective of tailoring content to suit individual user preferences must be carefully governed to prevent the establishment of isolated communities with limited perspectives or the dissemination of false or misleading information.

Stakeholder diversity refers to individuals or groups interested in or concerned about a particular issue or project. Competing interests, on the other hand, are conflicting or opposing desires or goals among these stakeholders. The multitude of stakeholders engaged in or impacted by AI systems further exacerbates the issue. Developers, users, regulatory authorities, and individuals affected by AI judgments may possess significantly divergent agendas and levels of risk tolerance. For example, a developer may give the highest importance to the precision and effectiveness of an AI system. In contrast, a regulatory body may concentrate on ensuring safety and impartiality, and users may be most concerned about privacy and customization. A balance between these conflicting objectives necessitates a meticulous and intentional strategy considering AI technologies' ethical ramifications.

The ethical aspect of the AI balancing problem is of utmost importance as the societal ramifications of decisions made by AI systems, such as in judicial sentencing, loan approvals, or autonomous vehicles, are substantial. The difficulty lies in incorporating ethical ideas into AI systems that uphold cultural diversity and individual liberties while advancing collective welfare. This entails addressing prejudice, unfair treatment, responsibility, and possible unforeseen outcomes from AI actions.

Exploring AI's ability to manage and prioritize various objectives requires a multidisciplinary strategy that combines knowledge from computer science, ethics, law, social sciences, and other related topics. It entails creating technologically sophisticated AI systems with social awareness and ethical alignment. This involves developing frameworks to ensure ethical AI, establishing robust systems for transparency and accountability, and promoting an open discourse among all stakeholders to ensure that AI technologies align with a wide range of human values and goals.

AI involves multiple objectives. As AI advances and becomes more integrated into everyday life, finding ways to manage this balance is essential to ensure that AI technologies improve human welfare, foster positive societal outcomes, and uphold the diverse human values and ethical principles that shape our shared existence.

1.2 Pareto Frontier

One way of balancing conflicting objectives in using AI is through the Pareto Efficient Frontier, derived from economics and operations research disciplines, which is fundamental for comprehending resource allocation in a specific system (Marwala 2024b; Censor 1977). The Pareto Efficient Frontier is named after Vilfredo Pareto, an Italian economist who created the notion of Pareto efficiency during the late nineteenth and early twentieth centuries (Cai et al. 2024). Pareto efficiency, also known as Pareto optimality, refers to a situation where the distribution of resources cannot be

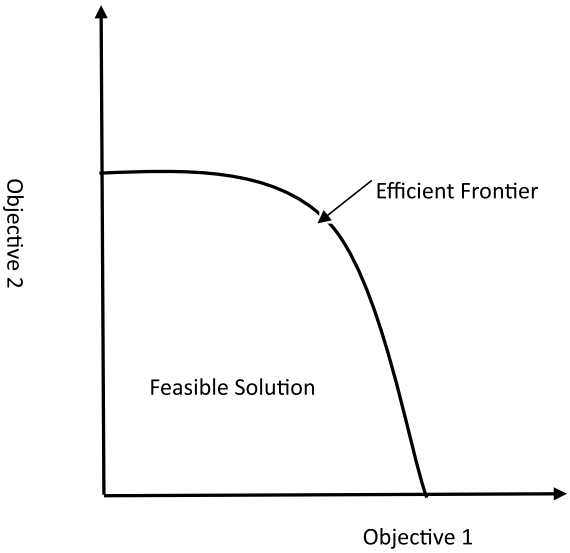

Fig. 1.1 Pareto Efficient Frontier

improved for one person without harming at least one other person. The Pareto Efficient Frontier, graphically depicted, is the arc or boundary that covers all conceivable locations of Pareto efficiency within a particular system, as shown in Fig. 1.1.

The Pareto Frontier depicts the trade-offs between different objectives in a two-dimensional space, where each axis represents a particular party's benefit level. Each point on the border symbolizes a stage when enhancing the situation of one partner requires compromising the other, exemplifying the fundamental concept of optimal allocation within the system's limitations. The Pareto Efficient Frontier is crucial in analyzing market efficiencies and resource distribution in economics. It facilitates comprehension of how various distributions might influence the well-being of individuals in an economy, guiding policymakers in making decisions that strive to enhance social welfare without harming others.

In finance, the Pareto frontier is utilized in portfolio optimization (Marwala 2013; Marwala and Hurwitz 2017; Maumela et al. 2022; Hurwitz and Marwala 2012; Hurwitz 2014). In this context, the frontier refers to the portfolio collection that provides the greatest anticipated return for a specific level or the least amount of risk for a particular level of expected return. Investors utilize this framework to make well-informed decisions that align with their risk tolerance and investment goals.

Operations research uses mathematical models and analytical techniques to make informed decisions (Bajaj et al. 2018). It involves analyzing complex problems and finding optimal solutions. The Pareto Frontier is employed in operations research to address multi-objective optimization problems, including simultaneous optimization of multiple conflicting objectives (Petchrompo et al. 2022). It allows one to understand the best outcomes in which you cannot improve one objective without

worsening another. It thus offers an analytical instrument for recognizing the compromises between various goals, assisting decision-makers in choosing the most suitable approach according to their priorities.

The Pareto Efficient Frontier has significant implications for decision-making in diverse disciplines. By identifying Pareto optimum points, decision-makers can assess the effectiveness of various options, enabling them to select choices that maximize overall utility or benefit. However, it is essential to recognize that Pareto efficiency does not inherently consider the fairness of resource distribution unless fairness is inserted as one of the goals, as one of the constraints, or both.

1.3 AI and Pareto Optimality

AI and Pareto Optimality are two fundamental ideas that, when combined, provide deep insights into optimizing systems for maximum efficiency and justice (Marwala 2018, 2019; Alvares et al. 2014). This section studies the incorporation of AI with the concepts of Pareto Optimality, analyzing how this combination might create systems that more effectively harmonize conflicting interests and resources in several fields, including economics, finance, healthcare, and environmental management.

A Pareto Optimal result is a state in which it is not possible to make any additional improvements for one person without causing a disadvantage for another person. This indicates a balance where resources are used most efficiently. The capacity of AI to efficiently process and analyze large volumes of data with exceptional speed and accuracy makes it a potent tool for identifying Pareto Optimal solutions in diverse domains (Khouadjia et al. 2013). AI systems can detect patterns and uncover insights that human analysts may overlook, and this allows for better-informed decision-making that closely adheres to the principles of Pareto Optimality.

AI may assist in environmental management by achieving a balance between economic growth and sustainability objectives (Mbuvha et al. 2021). It accomplishes this by identifying Pareto Optimal solutions that optimize resource efficiency while reducing environmental damage. AI algorithms may optimize energy use in industrial processes, reducing waste and emissions while maintaining output levels. This allows for a balanced alignment of economic and environmental concerns.

AI can help allocate medical resources, ensuring patients receive optimal care without overwhelming healthcare providers. AI systems can determine Pareto-optimal methods for patient care by examining patient data, treatment outcomes, and resource availability. These strategies aim to balance meeting individual patients' requirements and the healthcare system's capacity.

By combining AI with the concepts of Pareto Optimality, we can achieve more efficient and balanced solutions in different areas. Using AI to evaluate intricate datasets and determine the most advantageous allocation procedures, we may make significant progress toward achieving results that properly balance competing interests and resources.

1.4 Utilitarianism and Pareto Optimality

Incorporating AI into several aspects of human existence has had a profound and disruptive impact, presenting unparalleled prospects for progress and substantial ethical and societal dilemmas (Leke and Marwala 2019). Considering its dual-edged influence, a framework is required to steer the development and deployment of AI to maximize social benefits and minimize adverse repercussions. This section examines how utilitarianism and Pareto optimality guide AI towards beneficial and balanced results, ensuring its implementation improves general welfare without unfairly disadvantaging any social group (Mill 2016).

Utilitarianism is a normative ethics theory that states that the optimum behaviour maximizes utility. Utility is often defined as the greatest well-being of the most significant number of individuals (Sen 1979). When applied to AI, this principle implies that AI technologies should be created and implemented in manners that advance society's overall well-being (Mitov 2021; Sommaggio and Marchiori 2020). This entails improving efficiencies, generating economic benefits, and reducing the negative aspects of AI, such as privacy infringement, prejudice and unfair treatment, job displacement, and the erosion of human independence.

The utilitarian approach to AI necessitates a meticulous evaluation of AI's impact on different stakeholders, including people and communities. Developers and politicians must prioritize measures that have the most net good effects on society, considering AI's benefits (such as better healthcare results, improved learning experiences, and environmental protection) and its possible adverse effects.

Pareto optimality enhances the utilitarian paradigm by incorporating fairness into assessing AI's societal effects (Dhillon 1998; Lagerspetz 1984; Che et al. 2024). To achieve Pareto optimality in AI applications, it is necessary to identify and manage trade-offs. This means ensuring that the use of AI technology does not result in some individuals or groups benefiting at the expense of others but rather contributes to the overall improvement of society. This could entail, for instance, creating AI systems in education that customize learning experiences to improve student achievements without worsening educational disparities or implementing AI in workforce management methods that boost human capacities without causing extensive unemployment.

The difficulty is to find a balance between the utilitarian objective of maximizing general well-being and the Pareto principle of ensuring equal improvement. This balance necessitates a comprehensive strategy. First, it requires ethical AI development that incorporates ethical questions into the AI development process to ensure that AI systems are transparent, understandable, and devoid of biases that may result in uneven outcomes. Second, it requires regulatory oversight that enforces established guidelines for the ethical implementation of AI, safeguarding the rights and privacy of persons and ensuring equitable distribution of AI's advantages throughout society. Third, it requires diverse stakeholders in the decision-making processes concerning AI, particularly those most likely to be impacted by its implementation, to ensure the inclusion of varied perspectives and values. Fourth, continuous monitoring and

evaluation processes must be implemented to examine the implications of AI regularly, enabling the adaptation of strategies and policies to address evolving issues and possibilities.

Utilitarianism and Pareto optimality principles provide essential direction for understanding and addressing AI's intricate societal effects. By making a concerted effort to promote the overall welfare and ensuring that AI's advantages are shared, we may utilize AI's capabilities to establish a more prosperous, fair, and environmentally sustainable future.

1.5 Model for AI Governance Balancing Problem

This section describes a model for handling the AI balancing problem. The algorithm followed in this regard is described below and illustrated in Fig. 1.2.

1. **Establish Clear and Specific Goals**
   Good Values (G): Specify the criteria that determine the desirable qualities of AI. These factors may include efficiency, precision, user contentment, accessibility, etc.
   Negative Consequences (B): Specify the adverse effects or detrimental outcomes of AI, such as privacy issues, bias, unemployment, etc.
2. **Define Key Performance Indicators**
   Establish quantifiable measures for both positive and negative values. For instance, accuracy can be assessed by the proportion of accurate predictions, whereas the frequency of data breaches or complaints can evaluate privacy concerns.
3. **Gather Data**
   Collect pertinent data for each metric regarding the present AI systems being utilized.
4. **Identify Pareto Improvements**
   Identify areas of improvement: Seek modifications that can enhance at least one measurement of *G* without deteriorating any measurement of *B*.
   Assess Trade-offs: When faced with unavoidable trade-offs, employ Pareto efficiency as a criterion to ensure that any enhancement in *G* does not result in an unacceptable detriment to *B*.
5. **Execute Modifications**
   Implement the modifications indicated in Step 4, beginning with the ones that provide the most enhancements to *G* while causing the least disruption to *B*.
6. **Sequential Evaluation**
   Evaluate Metrics: Upon implementing the adjustments, it is necessary to reevaluate all metrics.
   Iterative optimization: Continuously iterate steps 4 and 5, consistently seeking improvements that align with the Pareto principle.

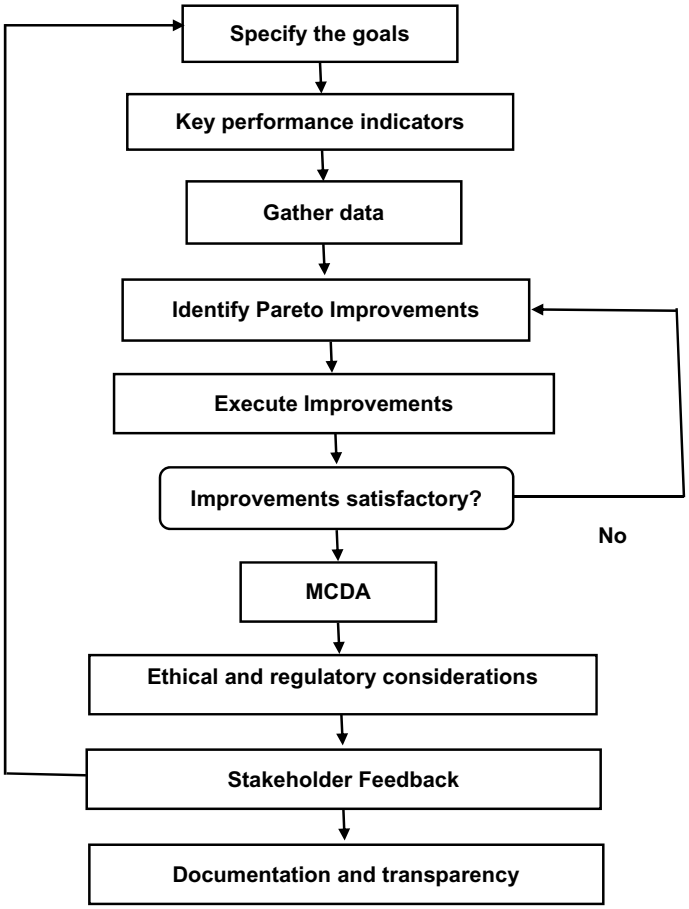

Fig. 1.2 Algorithm for applying Pareto optimality to balance the benefits versus the risks of AI

7. **Multi-Criteria Decision Analysis (MCDA)**
   This method evaluates and compares options based on many criteria (Guitouni and Martel 1998). When dealing with intricate situations with a complicated interplay of positive and negative values, it is advisable to utilize MCDA approaches to help assess the significance of various objectives and achieve a balance that maximizes overall utility based on Pareto optimality principles.
8. **Ethical and Regulatory Considerations**
   Make sure that any modifications and improvements adhere to ethical norms and regulatory obligations, considering the impact on society.
9. **Stakeholder Feedback**

Interact with stakeholders, such as users, developers, ethicists, and the public, to collect input on the perceived effects of AI and modify the objectives and metrics accordingly.

10. **Documentation and Transparency**
    Record the procedure, choices, and the reasoning behind them to ensure openness and responsibility.

The iterative nature of this algorithm necessitates ongoing review and change in response to the availability of new data and the evolution of the AI system. The objective is to establish a dynamic balance in which the advantages of AI are maximized and the disadvantages are minimized while adhering to the rules of Pareto optimality.

To illustrate this algorithm, we consider the trade-off between the transparency and accuracy of AI systems in healthcare. Here is how the algorithm discussed earlier can be utilized for this problem. AI provides unparalleled precision in diagnosis, treatment suggestions, and patient outcomes (Mbuvha and Marwala 2020; Scurrell et al. 2007). Nevertheless, this technological progress presents a notable obstacle: the need for more transparency without compromising the accuracy of AI decision-making procedures. The complex algorithms that power AI systems function as opaque entities, leaving healthcare practitioners and individuals unaware of the decision-making process. This algorithm above balances the precision of AI systems in healthcare and the crucial requirement for transparency by employing the concepts of Pareto optimality.

Pareto optimality offers a framework for maximizing the advantages of AI while reducing its disadvantages. Within this framework, the term "good values" (G) refers to the precision exhibited by AI in healthcare. In contrast, the term "bad values" (B) signifies the absence of transparency. The first step is to define the objectives explicitly using Pareto optimality. Accuracy can be tested using metrics like the percentage of correct diagnoses and the effectiveness of patient outcomes. On the other hand, transparency can be evaluated by assessing the comprehensibility of AI judgments for healthcare professionals and patient satisfaction with the explanations provided.

Once these parameters are defined, the subsequent step is to gather data on AI systems' present accuracy and transparency. This data collection is the initial reference point for recognizing and executing improvements. The essence of the Pareto optimality technique is to discover possible enhancements that increase transparency while maintaining accuracy. This could entail investigating explainable AI (XAI) methodologies that offer insights into the AI's decision-making process, clarifying the "black box" without compromising the system's diagnostic accuracy (Baniecki and Biecek 2024; Klauschen et al. 2024; Alizadehsani et al. 2024).

Implementing modifications using XAI methodologies necessitates a meticulous assessment of trade-offs, prioritizing improvements that substantially increase transparency while minimizing any decrease in accuracy. The healthcare industry should not tolerate sacrifices in patient care; therefore, any modifications must maintain or enhance the quality of care. After implementing these adjustments, a repetitive