SWIFT INSTITUTE

WORKING PAPER

FAIRNESS AI – FINFAIR AI INDEX FRAMEWORK

BADR BOUSSABAT

PUBLICATION DATE: JUNE, 2024

I would like to express my sincere gratitude to Dr. Jassim Haji and Dr. Patrice Latinne for their invaluable guidance and advice in the preparation of this paper.

Table of content:

Abstract

In the universe of financial analytics, particularly business-to-business (B2B) models that employ AI, fairness is of utmost importance. This study investigates the aspects of fairness within B2B financial transactions and addresses a significant gap in the existing literature. It presents a comprehensive framework for assessing fairness in B2B financial models, emphasizing the authenticity and reliability of the data inputs. This study introduced an innovative approach and proposed a pioneering method for fairness scoring with customizable parameters. This study offers a dual contribution to the field of AI fairness, enriching theoretical understanding and providing practical insights for financial institutions. This study underscores the need for a broader perspective on AI fairness, extending beyond regulatory compliance to embrace holistic fairness in finance. This research direction will guide future inquiries into AI practices in the B2B financial domain that are ethically grounded and attuned to the nuances of fairness.

1. Introduction

This study is situated within the evolving landscape of financial systems, specifically focusing on the transition from mechanical to organic solidarity in finance, as framed by the structural functionalist paradigm. This shift is particularly relevant in the context of AI-driven B2B finance, where traditional practices are reshaped by technological advancements and ethical considerations. The core of this study revolves around the development of a fairness-scoring model designed to navigate and enhance fairness in financial decision-making.

This study unveils the *FinFair AI Index* framework, a comprehensive tool that encapsulates critical elements of fairness in AI applications within the financial sector. This framework integrates the variables pertaining to a fair process and model/data bias. The introduction of this framework represents a significant shift in financial models, moving from a uniform, one-size-fits-all approach (mechanical solidarity) to a more nuanced and adaptable strategy (organic solidarity). This transition is not only a response to technological evolution but also a proactive effort to embed ethical considerations into AI practices in finance.

This study emphasizes the importance of collaboration between public authorities and financial players in this endeavor. By fostering a collaborative environment, the study seeks to ensure that the financial sector can effectively adapt to the challenges and opportunities presented by AI and machine learning, thereby contributing to AI fairness in the B2B financial system.

The Challenge of Integrating Fairness in Predictive Models

The integration of fairness into predictive models in the B2B financial sector presents a complex and multi-faceted challenge. This complexity stems not only from the technical aspects of AI, but also from the nature of decision-making processes. In this section, we explore the hurdles and considerations that complicate the pursuit of fairness in predictive modeling in the B2B financial context.

One primary challenge lies in defining fairness. While conceptually universal, fairness is a construct that varies significantly in interpretation and implementation across different domains and applications. In predictive models, fairness must balance accuracy with ethical considerations to ensure that decisions are made without undue bias or discrimination against groups or entities. However, what constitutes bias in a B2B financial model can be substantially different from consumer-facing models, which often require a deeper analysis of industry-specific dynamics and stakeholder relationships.

Another significant hurdle is the inherent limitations and biases present in the data fed into these models. Predictive models are only unbiased as they process data. In many cases, historical data in the financial sector may reflect existing prejudices or skewed representations, leading to models that perpetuate these biases inadvertently. Identifying,

quantifying, and correcting such biases pose substantial challenges, especially in complex financial environments where data sources are diverse and multi-layered.

Moreover, the technical design and implementation of predictive models introduces additional layers of complexity. The choice of algorithms, weighting of inputs, and interpretation of outputs influence the fairness of the model. Balancing the technical demands of accuracy and reliability with the ethical imperative of fairness requires not only advanced analytical capabilities, but also a profound understanding of the ethical implications of AI in finance.

Purpose and Scope of the Research

Purpose

The main objective of this study is to create an extensive and progressive scoring framework for assessing the fairness of AI systems in the business-to-business (B2B) financial sector. This framework, based on a multifaceted approach to fairness, aims to instigate a transformative change in the financial industry towards the realization of fair outcomes. The intention was to enhance the present structure of the financial sector and promote fairer applications of AI.

Scope

This study encompassed several key dimensions.

- **Definition of fairness**: The definition of fairness will be the starting foundation of the development of our framework
- **Presentation of the new framework**: We present a new paradigm based on structural functionalism to develop our *FinFair AI Index* for AI models.
- ***FinFair AI Index* Development**: This study introduces a mathematical equation. This serves as the foundation of the scoring model and quantifies the fairness of the AI models utilized by the financial sector.
- **Debate on the framework**: Implementation of the framework in the financial sector presents several challenges that must be addressed.

2. Literature review

In this comprehensive literature review, we explore the development of fairness as a concept in AI with a specific focus on its implications within the B2B financial sector. This inquiry is organized to systematically analyze seminal works in the field, thereby elucidating the evolution of the understanding of fairness in financial AI. Our objective is to present a cohesive narrative, tracing the contributions of each research study to a broader comprehension of fairness, its theoretical foundations, and its practical applications in AI systems within finance. Our aim is to synthesize these works into a coherent overview, emphasizing significant developments and shifts in thought within this critical domain of study. By scrutinizing the trajectory and intricacies of these scholarly contributions, this study aims to establish a comprehensive and well-informed understanding of fairness in financial AI, which is crucial for our approach in this research.

This review provides a starting point for our exploration of fairness in the financial sector, particularly in B2B settings. Building upon and expanding upon the foundational works discussed herein, we seek to offer new perspectives and innovative approaches to financial AI fairness. Our ultimate goal is to contribute significantly to the ongoing discourse on fairness in the financial industry, ensuring that our research has a meaningful impact on both academic and practical realms.

Fairness through science

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226. https://doi.org/10.1145/2090236.2090255

- **Approach to Fairness**: Dwork et al. present a pioneering framework in the realm of algorithmic fairness, with a focus on the concept of "individual fairness." This approach is groundbreaking in its emphasis on treating similar individuals, a principle that can be particularly transformative in the business-to-business financial sector.
- **Definition of Fairness**: Fairness is defined in this study as a metric-based approach, wherein the level of fairness is assessed by measuring the distance between individuals in a metric space. In the context of the B2B financial sector, companies that are close in terms of financial well-being, risk profiles, and other pertinent metrics ought to be treated equitably by AI algorithms.
- **Proposed Solutions**: The authors present a theoretical framework in which the fairness of an algorithm is determined by its ability to produce comparable results for comparable individuals. In the context of business-to-business finance, this entails the development of artificial intelligence models that consistently and equitably evaluate businesses based on their financial data. This paper also emphasizes the significance of carefully defining the metric space in which similarity is assessed, as

this is essential for ensuring that the AI's assessment of similarity aligns with the principles of fairness.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732. Retrieved from http://www.californialawreview.org/wp-content/uploads/2016/06/5-Big-Datas-Disparate-Impact.pdf

- **Approach to Fairness**: Barocas and Selbst's seminal work provides a deep exploration of the intricacies associated with big data and its ramifications for fairness, with a particular emphasis on artificial intelligence systems. Their investigation centers on the idea of disparate impact, a form of unintentional discrimination that can emerge even in the absence of malevolent intent, which holds relevance within the financial services sector.
- **Definition of Fairness**: This study delineates disparate impact as a situation wherein algorithms, despite ostensibly being impartial, yield outcomes that unfairly impair specific demographics. In the context of B2B financial services, this could occur when AI systems unintentionally demonstrate preferences or impose penalties on well-targeted businesses based on their size, location, or market sector as a result of biased data or inadequate model assumptions.
- **Proposed Solutions**: Barocas and Selbst underscored the significance of vigilance and proactive steps in mitigating disparate impacts. They advocate extensive testing and auditing of AI systems to detect and rectify any potential biases. Moreover, the paper recommends enhancing transparency in AI decision-making processes and implementing fairness-conscious modeling techniques. These approaches are indispensable for financial institutions leveraging AI, as they guarantee that the systems are not only productive but also just and devoid of unintentional prejudice.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315-3323. Retrieved from https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf

- **Approach to Fairness:** Hardt et al. propose a novel approach to fairness in AI, centered around the concept of "equality of opportunity." This approach is particularly relevant in supervised learning contexts, in which predictive models play a crucial role in decision-making processes, such as in the B2B financial sector.
- **Definition of Fairness:** The present document delineates the concept of fairness as a state in which individuals who exhibit similarities with respect to a specific task are awarded comparable predictions. In the context of business-to-business financial dealings, this notion guarantees that corporations with similar financial health and risk profiles are impartially evaluated by AI systems.
- **Proposed Solutions:** To attain fairness, Hardt et al. incorporated a fairness constraint into the model training process. The central concept entails equalizing the true positive rates among diverse groups, thereby guaranteeing AI's impartial and consistent performance across various segments of the financial market. This

approach represents a pivotal advancement by proposing a tangible and quantifiable method for integrating fairness into AI algorithms, particularly for developing fair credit scoring and risk assessment models in finance.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. Proceedings of the 26th International Conference on World Wide Web, 1171-1180. https://doi.org/10.1145/3038912.3052660

- **Approach to Fairness:** Zafar et al. delved into the subject of fairness in artificial intelligence by focusing on the issue of disparate mistreatment. Their work was set against the backdrop of classification problems, which are particularly relevant in financial decision-making systems.
- **Definition of Fairness:** The study's primary focus lies in the concept of fairness, which entails the prevention of disparate mistreatment arising from differential accuracy rates among various groups. In the context of B2B financial transactions, this principle demands that AI models achieve consistent predictive accuracy across a wide range of businesses regardless of their size, sector, or other distinguishing characteristics.
- **Proposed Solutions:** Zafar et al. presented a methodology that aims to alleviate disparate mistreatment by integrating fairness constraints into the classifier optimization problem. They proposed a convex proxy for fairness that enabled the provision of efficient and scalable solutions. This approach is essential for financial institutions that strive to develop artificial intelligence systems that are both highly accurate and fair for treating various business clients.

Kleinberg, J., Mullainathan, S., Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

- **Approach to Fairness**: The study conducted by Kleinberg et al. focuses on a crucial issue in the realm of AI in finance: the presence of inherent trade-offs when striving for fairness in risk score determination. Their research is of paramount importance in comprehending the intricate dynamics involved in the development of fair AI systems for financial decision making, particularly in the B2B sector.
- **Definition of Fairness**: The present study delves into the issue of fairness in the realm of risk assessments, where the aim is to guarantee that artificial intelligence systems do not perpetuate or intensify existing prejudices against particular demographics. The authors examined the notion that diverse notions of fairness may come into conflict with one another, which presents a challenge in developing an AI system that can meet all fairness criteria concurrently.
- **Proposed Solutions**: Kleinberg and his associates have examined the mathematical foundations of these trade-offs, which are achieved through the application of theoretical models. They argue that the simultaneous attainment of all types of fairness may be an unrealistic goal, and thus, a crucial understanding of these trade-

offs is necessary in making informed decisions regarding which fairness criteria to prioritize. This insight is particularly pertinent for financial institutions operating in the B2B sector, where risk assessment models must balance accuracy with fairness towards diverse client groups.

Hildebrandt, M. (2018). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. Theoretical Inquiries in Law, 20(1), 83-121. https://doi.org/10.1515/til-2019-0004

- **Approach to Fairness**: Hildebrandt's work offers a distinctive perspective on fairness in AI, particularly in the finance sector, by concentrating on the nexus of privacy, personal data, and machine learning. This approach is essential for comprehending the far-reaching consequences of AI in B2B financial settings where personal and business data are frequently intricately entwined.
- **Definition of Fairness**: Hildebrandt's analysis of fairness is informed by her consideration of privacy and the incalculable elements of individuals and organizations. She maintained that achieving genuine fairness in AI systems, especially in critical domains such as finance, necessitates considering the intricate and multifaceted nature of personal and commercial identities that cannot be fully captured through quantification and computational processes.
- **Proposed Solutions**: The paper proposes a transition from "agnostic" to "agonistic" machine learning, emphasizing the importance of respecting the incomputable aspects of identity and avoiding oversimplification or misinterpretation of these complexities. In the realm of B2B finance, this entails creating AI models that are cognizant of the subtleties of business operations and contexts and that safeguard the privacy and intricacy of business data. Hildebrandt underlines the necessity of AI systems that are transparent, accountable, and developed with an understanding of the ethical ramifications of their incorporation in decision-making procedures.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943. Retrieved from https://arxiv.org/abs/1810.01943

- **Approach to Fairness**: Bellamy et al. introduced the AI Fairness 360 toolkit, which is a comprehensive suite of metrics and algorithms designed to detect, understand, and mitigate unwanted biases in AI systems. This toolkit represents a significant advancement in the field of AI fairness, particularly for B2B financial services, in which decisions made by AI can have substantial implications.
- **Definition of Fairness**: The AI Fairness 360 toolkit is constructed on the foundation that fairness in artificial intelligence is multifaceted and demands a diverse array of metrics and methods to evaluate and guarantee. This toolkit presents a comprehensive collection of fairness measurements and algorithms, acknowledging

that fairness is not a uniform concept, particularly in intricate and diverse B2B financial settings.

- **Proposed Solutions**: This toolkit presents a comprehensive selection of algorithms designed to reduce bias in datasets and models, making it an indispensable asset for financial institutions committed to guaranteeing the fairness of their AI systems. It empowers users to select from a diverse array of fairness definitions and applies the most relevant algorithms to their unique situations. Adaptability is essential in the B2B financial industry, where AI applications encompass credit scoring, risk management, and personalized financial advice.

Mehrabi, N. et al. (2019) - "A Survey of Bias and Fairness in Machine Learning"

- **Approach to Fairness:** Mehrabi et al. offer a comprehensive examination of fairness in machine learning, emphasizing its multifaceted character. They delineate the diverse biases and fairness definitions that have arisen in AI research, rendering their study an indispensable reference for comprehending the intricacies of fairness**.**
- **Definition of Fairness:** The study delineates various facets of fairness, including, but not limited to, demographic parity, equality of opportunity, and individual fairness. The former pertains to the achievement of fair results across divergent demographics, whereas the latter seeks to eliminate disparities in error rates among these demographics. In contrast, individual fairness champions the principle of treating similar individuals consistently.
- **Proposed Solutions:** This document elucidates diverse aspects of fairness, such as demographic parity, equality of opportunity, and individual fairness. The former entails the attainment of fair outcomes across diverse demographics, whereas the latter aims to eliminate disparities in error rates among these entities. Conversely, individual fairness advocates the principle of treating comparable individuals consistently.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635. Retrieved from https://arxiv.org/abs/1908.09635

- **Approach to Fairness:** This study examines the practical difficulties of incorporating fairness into financial machine-learning models while simultaneously grappling with the complex interplay between various fairness concepts.
- **Definition of Fairness**: This study examines various fairness measures, including statistical parity, which advocates for equal treatment of all groups, and individual fairness, which aims to provide consistent outcomes for individuals with similar characteristics. Additionally, this study explores conditional statistical parity, which considers legitimate factors that may warrant different treatment between groups.
- **Proposed Solutions:** The study's authors advocate for a practical methodology wherein financial institutions must evaluate and weigh these competing fairness criteria according to their scenarios. This could entail the development of tailored fairness metrics that reconcile both the business objectives and regulatory parameters of the financial sector.

Zhang, Y., & Zhou, L. (2019). Fairness Assessment for artificial intelligence in the Financial Industry. arXiv preprint arXiv:1912.07211.

- **Approach to Fairness:** This study presents a framework for evaluating the fairness of AI within the financial sector, considering the distinctive data characteristics and regulatory requirements of the industry.
- **Definition of Fairness:** The framework incorporates multiple aspects of fairness, such as procedural fairness, which pertains to the methodology of decision-making, and outcome fairness, which focuses on the consequences of decisions. It recognizes the necessity for fairness evaluations attuned to the intricacies of financial data and regulatory requirements.
- **Proposed Solutions:** The authors advocate for a tailored fairness assessment approach that encompasses data collection, model development, and post-deployment evaluation for financial institutions that utilize AI systems. This comprehensive approach is essential for guaranteeing fairness, transparency, and adherence to industry standards in the financial sector.

United Nations Educational, Scientific and Cultural Organization. (2021). Recommendation on the Ethics of Artificial Intelligence. [PDF file]. UNESCO. https://unesco.org/EthicsAI

- **Approach to Fairness:** This UNESCO document advocates for an inclusive, human-rights-centered approach to AI ethics, emphasizing fairness as a fundamental principle. It highlights the necessity of developing AI systems that respect diversity, promote non-discrimination, and ensure fair access to AI's benefits for all.
- **Definition of Fairness:** Fairness in this context is defined as the fair treatment of all individuals and groups, ensuring that AI technologies do not replicate or exacerbate biases and inequalities. It involves respecting human rights and fundamental freedoms, with a focus on promoting social justice and inclusivity.
- **Proposed Solutions:** The UNESCO recommendation proposes several solutions to achieve fairness in AI. These include developing ethical governance frameworks, ensuring transparency and explainability in AI systems, and promoting human oversight and accountability. Additionally, the document emphasizes the need for multi-stakeholder collaboration, adequate privacy and data protection measures, and the assessment of AI's impact on various aspects of society, such as health, education, and the environment.

Literature suggests a developmental trajectory in which fairness in AI has progressed from an abstract concept to a practical challenge, particularly in the financial services sector. This highlights the urgent need for a comprehensive interdisciplinary approach that combines technical expertise with ethical foresight to address the complexity of fairness in AI. The

adoption of this multifaceted perspective is not only academic in nature but also crucial for the ethical and responsible deployment of AI in the influential and far-reaching domains of finance.

3. Fairness: definition

Sensu lato definition

In the world of AI, particularly within the B2B financial sector, the principle of fairness is a fundamental aspect of responsible AI. This requires a sophisticated, multidimensional approach rooted in both introspective analysis of AI systems and a comprehensive understanding of the external business environment. Fairness in this context is more than mere adherence to ethical standards; it also represents a dynamic and continual process of learning and adaptation, which is essential for the interaction and decision-making processes of AI systems in diverse business scenarios.

Central to this discourse is the fundamental principle of fair treatment of businesses that are alike in nature. In the intricate realm of financial artificial intelligence, this principle mandates that algorithms assess businesses in an impartial and systematic manner. To illustrate, businesses with comparable financial soundness or market hazards should be evaluated uniformly using AI algorithms. This calls for AI systems to possess a profound and discerning comprehension of the business environment, allowing them to precisely discern and respond to delicate variations among company profiles.

In addition, the implementation of fairness in artificial intelligence presents significant obstacles, precisely guaranteeing that AI models do not inadvertently produce outcomes that disproportionately harm certain business groups. This concern demands extensive observation and continuous monitoring of AI systems to detect and mitigate potential bias. Consequently, the principle of fairness extends beyond the design of algorithms to encompass the entire lifecycle of AI systems, including data collection, model development, deployment, and continuous refinement based on feedback.

Ensuring fairness in financial artificial intelligence necessitates a balance between predictive accuracy and ethical considerations, particularly in financial decision-making, where the implications of outcomes are significant. It is crucial to maintain a balance between high algorithmic accuracy and comprehensive fairness protocols, a balance that is not constant but evolves in response to modifications in the financial market and regulatory environment.

Sensu stricto definition

As we endeavor to examine fairness within AI, particularly in the B2B context, our investigation now turns towards the critical topic of bias. This concept has been repeatedly emphasized in scholarly literature on AI fairness. Bias in AI is essentially characterized by two main forms: model bias and data bias. These two aspects of bias have a significant influence on the fairness of AI systems, and understanding them is of great importance for the creation of fair AI models in the financial sector.

The term *model bias* refers to the inherent inclinations within artificial intelligence algorithms that can result in biased outcomes. This form of bias frequently stems from the design of the algorithm itself, where decision-making rules or standards may unintentionally favor specific groups over others. Model bias presents a daunting challenge because it can

be discreetly integrated into the workings of AI systems, making it challenging to identify and rectify.

On the other hand, *data bias* originates from the data scientists utilized to train artificial intelligence models. It materializes when these datasets are not reflective of genuine world circumstances, or when they comprise historical biases. In the B2B context, data bias can considerably influence AI's decision-making, resulting in unfair outcomes for specific business entities, specifically those that are underrepresented or misrepresented in the training data.

Fair process

Fairness, as a concept within the world of artificial intelligence in the B2B financial sector, inherently possesses a dynamic nature that evolves in tandem with the market. This characteristic necessitates that fairness be perceived not as a static target but as an asymptote, an ever-approachable ideal that continually adapts to changing market conditions. Consequently, what makes a fair transaction in this context is subject to the temporal variability. A decision deemed fair today may no longer hold the same status tomorrow, even when assessed using the same AI model with identical data. This fluidity highlights the importance of continuously monitoring and updating AI systems to align with evolving market dynamics. As market conditions, business practices, and regulatory landscapes shift, so do the parameters and assessments of fairness within AI systems, ensuring that their decisions remain in accordance with the current state of the market. This approach underscores the obligation of AI models used by the financial sector to embody a flexible and responsive framework that can adjust to the ever-changing definitions and standards of fairness. Thus, the market is an ideal benchmark.

The pursuit of fairness in AI is an iterative and asymptotic endeavor characterized by a journey rather than a final destination. This implies that AI systems are perpetually immersed in the process of learning and adaptation, whereby they continually assimilate and apply fresh information regarding the multifarious financial entities they engage with. This ongoing refinement process mirrors the intricate and ever-transforming nature of the financial market, thereby demanding AI systems capable of simultaneously evolving alongside it.

An essential aspect of this endeavor is the AI system's extensive grasp of a variety of business models and market trends. This necessitates the merging of advanced AI algorithms with a thorough understanding of diverse business types, their financial performance indicators, and the challenges associated with each market participant. Establishing such a refined understanding is crucial to guarantee that AI-guided decisions are fair and customized to the specific attributes of each financial entity.

Furthermore, the concept of fairness in AI has been extended to include proactive bias management. This entails a vigilant and adaptive approach for identifying and mitigating biases, both within the AI models and the underlying data. Recognizing that biases are dynamic, AI systems must be equipped to respond in real time and continuously update their processes and datasets to maintain fairness.

Lastly, a growing academic discourse[1] suggests a potential correlation between fairness in AI and improved business performance. This theory posits that a deeper understanding of business counterparties, facilitated by fair AI systems, can lead to more effective and informed decision making. Such decision-making could in turn enhance the performance and profitability of AI-driven financial services, underscoring the multifaceted benefits of fairness in AI.

Data bias

Data bias in AI, particularly in the context of the B2B financial sector, is a critical issue that arises when the datasets used to train and inform AI systems are not fully representative of diverse market actors or are not updated in real time to reflect current market conditions. This type of bias can significantly skew the AI-driven decision-making processes, leading to unfair or inaccurate outcomes.

Origin of data bias

Data bias arises when the data utilized in the training of an AI model do not represent the diversity of the market or are outdated, which can lead to inaccuracies in the AI learning process. In the B2B financial sector, this may manifest as the absence of sufficient and up-to-date data from certain types of businesses, such as niche economic agents or emerging industries, in the training of a credit-scoring model. Consequently, the AI system may be biased towards the characteristics of more well-represented entities, such as systemic banks.

Manifestation of data bias

Data bias can result in the discriminatory treatment of certain market participants. An illustration of this can be seen in loan approval AI that has primarily been trained on data from systemic banks. In such a scenario, AI may underestimate or misconstrue the financial well-being of less significant banks, consequently leading to unfair credit determinations.

Impact of data bias

The ramifications of the data bias are substantial. This can result in systemic disparities where specific businesses, such as less significant banks, are constantly at a disadvantage. This issue extends beyond individual financial entities, and has the potential to skew market competition and compromise financial stability.

Model Bias

Model bias in AI refers to inherent predispositions within AI algorithms that lead to skewed or prejudiced outcomes. This bias arises from the underlying mechanisms and structures of AI models, which may inadvertently favor certain patterns, characteristics, or groups over

[1] Frontiers Editorial Office. (2020). The impact of artificial intelligence on firm performance: An application of the resource-based view to e-commerce firms. *Frontiers in Psychology*, 11, 580406. https://doi.org/10.3389/fpsyg.2020.580406

others, leading to discriminatory decisions. Understanding model bias is crucial in the context of AI-driven financial services, as such biases can have significant and far-reaching implications.

Origin of model bias

Model bias often originates from the manner in which AI algorithms are designed and trained. These biases can be introduced through various factors, including the subjective decisions of the model developers, algorithmic structure, or choice of features and parameters used in the model. For instance, if an AI system for credit scoring is predominantly trained on data from large corporations, it may be less accurate or fair when assessing smaller businesses because of its overfitting to the characteristics of larger entities.

Manifestation of model bias

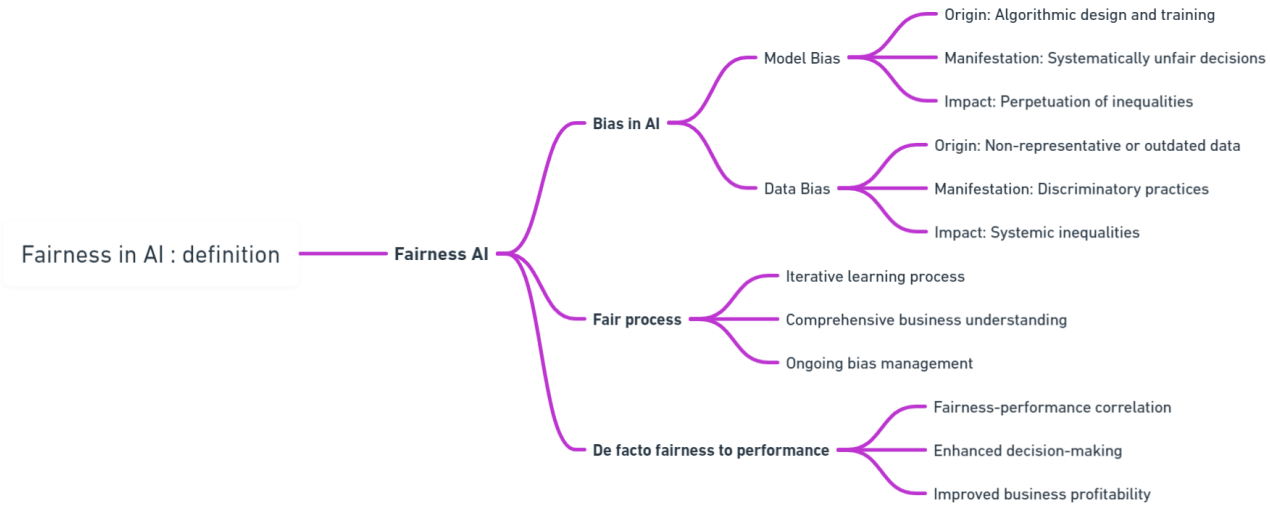
In practical applications, model bias is revealed through a systematic inclination of AI systems to make decisions that apply unfairness towards specific groups. This may manifest as a persistent disadvantage experienced by specific types of business. Bias can also stem from the learning algorithms themselves. For instance, if an algorithm is designed to prioritize certain financial indicators that are more common in larger banks, it inadvertently discriminates against smaller banks.

Impact of model bias

The impact of model bias can lead to a perpetuation of existing unfairness[2], where some entities are unduly favored or penalized, not based on their actual performance or risk profile; but due to the biased nature of AI models. This not only affects individual businesses, but also has broader implications for market dynamics and industry competition.

Hence, the goal of achieving *de facto fairness* in the B2B context is achieved by integrating a fair process with a controlled model and data biases, as previously presented. Controlling model bias involves calibrating the AI to prevent any inherent bias in decision-making, whereas managing data bias necessitates the AI to be trained on diverse and current datasets. Collectively, these components culminate in *de facto fairness*, in which the functionality of AI is inherently fair in practical applications in the market, reflecting an optimal equilibrium between ethical standards and functional feasibility in the financial services industry.

[2] Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., ... & Zafar, M. B. (2021). Fairness measures for machine learning in finance. *The Journal of Financial Data Science*.



4. *FinFair AI Index* framework: methodology

Description

Artificial Intelligence has ushered in a significant transformation in the financial sector, not only in terms of technological progress, but also in the socio-economic principles that underpin this domain. This study endeavors to explore this transformative phase, drawing on Durkheim's theoretical perspective and, specifically, the manifestation of *mechanical* and *organic solidarity* in the context of the financial sector's adoption of AI, with a focus on B2B fairness scoring models.

Traditional financial models, which have historically served as the foundational framework for risk assessment and decision-making processes in the sector, have long been characterized by a focus on *mechanical solidarity*. These models often reflect the homogeneity and shared norms that typify this approach, and may not adequately address the diverse needs of various business entities. However, the integration of AI, particularly fairness-oriented scoring models, has ushered in a compelling shift towards a fairer approach. This transition is not merely a technological upgrade, but represents a significant change in the *organic solidarity* of the financial sector.

The application of AI fairness models in a B2B context challenges conventional notions of mechanical solidarity and necessitates a reevaluation of established norms and practices. This shift towards organic solidarity is characterized by a more inclusive approach. This research contends that the development and implementation of AI models are not merely a transition from mechanical solidarity to organic solidarity, but rather an evolution towards a system that acknowledges and corrects the disparities and biases inherent in traditional financial models. This movement aligns with an ethically driven practice that prioritizes fairness.

The AI fairness scoring model proposed in this study is not an individual tool but rather form part of a comprehensive collaborative framework that involves public authorities and businesses. This cooperation is essential for aligning models with the ever-changing landscape of businesses and their dedication to fairness. By adopting this approach, we can ensure a scoring system that is better suited to the unique context of individual businesses and reflects a shift in how financial risks and opportunities are assessed and managed, transitioning from mechanical to organic solidarity.

In essence, this study investigates the transformation of the financial sector through AI-driven fairness models, using Durkheim's concept of *mechanical* and *organic solidarity*[3]. It emphasizes the need for a holistic transformation that extends beyond technical advancements to encompass a reform of cohesion and institutional practices, paving the way for a fairer financial ecosystem by promoting the adoption of fair AI models within the B2B sphere.

[3] Durkheim, E. (2010). From mechanical to organic solidarity. *Sociology: Introductory Readings*, *2*(1), 25-29.

Organic Solidarity vs Mechanic Solidarity

The concepts of mechanical and organic solidarity, as elucidated by Émile Durkheim, provide a refined perspective on fairness and decision-making processes in the business-to-business financial sector when applied to artificial intelligence. . Mechanical solidarity, characterized by homogeneity and a set of shared norms, typically underpins the conventional reporting frameworks between banks and regulators, aimed at promoting a stable financial system. This framework comprises the ensemble of norms that bind the regulator to the bank, focusing primarily on the relationship between these two entities without considering the broader interactions among banks in the market. The existing financial risk assessment models, rooted in the principles of mechanical solidarity, often overlook the diversity and distinctiveness of business entities within the financial market

Consequently, these models tend to rely on a standardized approach to evaluations, perpetuating a one-size-fits-all mentality that does not adequately address the nuances of individual entities. This approach, while providing a degree of consistency and predictability in regulatory reporting, often fails to capture the complexities and specificities inherent in inter-bank dynamics.

A fairness-centric scoring models heralds a profound evolution towards organic solidarity within the financial domain. This evolution marks a departure from uniform methodologies, ushering in an era that acknowledges and aligns with the intricate interdependencies of financial institutions. Such a tailored approach gains paramount importance in the B2B sphere, where the interplay and transactions between entities, like banks, are critical. This transition encapsulates the quintessence of organic solidarity in the financial ecosystem, reflecting a more nuanced and interconnected approach to financial interactions and assessments.

In a more nuanced perspective, organic solidarity within the financial sector is underpinned by the intricate interplay among banking institutions. A pivotal element in this dynamic is the mitigation of endogenous factors that frequently trigger systemic anomalies. This goal transcends the scope of standard regulatory reporting, characteristic of orthodox financial regulations, and delves into the enhancement of symbiotic inter-bank relationships. Central to this pursuit is the necessity to refine the fairness of AI systems, which are integral to the foundational interactions among market participants in the financial arena.

A *FinFair AI Index* in the B2B context fosters a dual responsibility framework: one that encompasses both regulatory bodies and inter-bank dynamics. This paradigm shift leads to a more robust operational process, transitioning from a one-to-many (regulator to individual bank) approach to a many-to-many framework[4]. This holistic model not only promotes accountability but also enhances the resilience of the financial system in terms of fairness promotion.

Moreover, this approach proves to be cost-effective, as it inherently reduces the likelihood of a general financial crisis. It is essential to recall that global financial crises are often the result of inefficiencies within the system – a manifestation of endogenous malfunctions. The

[4] bank-to-bank and bank-to-regulator

interaction between banks, especially when utilizing AI in B2B transactions, plays a crucial role in mitigating these endogenous constraints. It allows for a more accurate assessment of the fairness of AI systems employed by banks, thereby contributing to a more efficient and fair financial ecosystem.

In essence, by fostering a network of interactions underpinned by fair AI systems, the financial sector can move towards a more interconnected and resilient framework. This not only enhances the overall efficiency of the system but also aligns with the principles of organic solidarity, ensuring that each entity's actions contribute positively to the stability and sustainability of the broader financial market.

In the business-to-business financial sector, organic solidarity is of immense importance. It acknowledges the network of connections among the various financial institutions. For example, in a transaction involving artificial in predicting fraud, it is essential that Bank A does not unfairly discriminate against Bank B, and vice versa. This interconnectedness necessitates a refined comprehension and utilization of AI models that can discern these business relationships.

The concept of organic solidarity enables a specific and balanced methodology for the financial industry. It fosters the creation of models that not only exhibit uniform fairness but also address individual fairness. This will be elucidated in future research through delineation of the *FinFair AI Index* equation, which seeks to quantify and embody this concept of fairness in an operational manner.

Building upon our defined concept of fairness in AI within the B2B financial sector, we now delineate a methodology for developing a fairness scoring system in this context, drawing inspiration from the structural functionalism paradigm. This theory, contributed to by Spencer, Merton and even Durkheim, views interactions as part of a complex system where various parties collaborate to maintain stability, enhance accuracy in the AI predictions and promote fairness. This theoretical lens provides a valuable framework to understand and enhance AI fairness in the financial sector.

Structural functionalist paradigm

This theoretical perspective conceives of the financial sector as a collective of interconnected entities comprising businesses, regulatory bodies, financial institutions, and AI technologies. Each entity assumes a unique role: collectively, it contributes to the sector's operational efficiency. By adopting this approach, one can examine the intricate web of relationships within a sector and how these relationships shape the pursuit of fairness.

The objective is to evolve the structure of the B2B financial sector to realize factual fairness (*de facto fairness*). This evolution involved the introduction of an incentive-based fairness-scoring model. This model is designed to measure fairness and foster practices that promote fair outcomes in AI-driven decisions. This scoring system is envisioned to be fair in its process, thus incentivizing the adoption of fair AI practices across the sector.

Fairness scoring model development

Our methodology, rooted in the principles of structural functionalism, seeks to establish a comprehensive framework within the B2B financial sector that emphasizes fairness in AI.

This approach entails a fair and transparent process to implement a fairness scoring system designed to elevate fairness scores and foster a continuous cycle of improvement. By collaborating and adhering to the principles of fair AI, financial entities can actively contribute to creating an environment in which fairness is realized and reinforced. This framework evaluates current fairness levels and encourages the adoption of practices that enhance fairness, thereby promoting the financial sector's communion, the latest being the prerequisite of de facto fairness

As we continue to refine our methodology for developing a fairness scoring system for AI in the B2B financial sector, we delve deeper into the structural functionalism paradigm. This paradigm draws on Émile Durkheim's concept of mechanical and organic solidarity within a functioning system, offering a nuanced understanding of systemic cohesion and the evolution of social structures. By adapting Durkheim's framework, we explore the transformation of standard financial practices towards a fairer and inclusive AI-driven environment.

From mechanical to organic solidarity in finance

Durkheim's theory of mechanical solidarity aligns with the traditional standardized risk management models prevalent in the financial sector, such as the Internal Ratings-Based (IRB) approach used by banks. While forming the technical backbone of the sector, these models often reflect and perpetuate existing power dynamics, favoring institutions with more extensive data and resources. This scenario is akin to mechanical solidarity, in which the homogeneity of practices and perspectives dominates.

In contrast, our proposed *FinFair AI Index model* signifies a concrete shift towards organic solidarity, embodying a move away from uniformity towards a system that values diversity and interdependence. This model represents not only technical recalibration but also broader cultural and ideological evolution in the financial sector. It advocates a more nuanced and fair approach to AI evaluation, recognizing the importance of the process of achieving fairness alongside the outcome.

Collaboration between public authorities and financial players

The core of this model is cooperation between public authorities and financial institutions. This partnership is essential for the shift from mechanical solidarity, which relies on standardized models, to organic solidarity, which is characterized by *de facto* fairness. By enabling banks to modify their AI models to reflect their evolving commitment to fair practices, public authorities can contribute significantly to this transition. This collaboration ensures that fairness scoring is not only more tailored and pertinent, but also inspires banks to strive for fairness through a collaborative and calibrated strategy. Finally, the fairness scoring system was designed to be adaptable and personalized, acknowledging the unique contexts and capabilities of different banks. This approach allows for a more tailored assessment of AI systems and data, promoting fairness in the AI applications.

FinFair AI Index framework

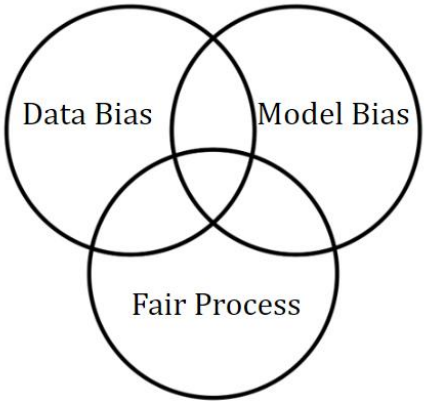
We propose a new equation to evolve the structure of the B2B financial sector towards actualized fairness: the *FinFair AI Index.* This model focuses on the development of an incentive-based and inherently fair AI fairness scoring system. Our approach seeks to

embrace a new paradigm that first thoroughly comprehends the current structure of the financial sector and then responds to it using a specially formulated equation aimed at achieving fairness objectives. We are now advancing towards the practical realization of the previously discussed theoretical approach by implementing a non-competitive and individual fairness scoring. This scoring takes into account the interaction between banks and is based on a collaborative effort with regulatory authorities.

The equation developed in this study was not intended for random or arbitrary applications. Rather, it is designed for implementation within a framework founded on public-private collaboration, as the outcome of the scoring is inherently linked to the organization and quality of this collaboration. This represents a novel approach to our research. Additionally, this framework facilitates the positive evolution of the scoring process, incentivizing banks to improve their fairness scores. The weights and values assigned to the specific variables in the equation are critical for demonstrating this aspect. These are elements in which businesses have less flexibility and must provide justifications. It also underscores the fairness of the adjustable application of the fairness model. Consequently, any random testing of public data would inherently be unfair, as it would be disconnected from the established framework. This approach allows us to move away from arbitrary standardized models or internal banking models that often do not reflect reality within conventional frameworks.

When constructing the equation, it is essential to delineate and articulate the variables that contribute to the integrity of AI systems. The blueprint for this endeavor starts with the "fair process" dimension, a systematic protocol that refines AI practices to ensure fair outcomes. Within this process, pivotal junctures are encountered, each of which presents critical variables that require attention.

Firstly, we address *model bias*, a variable that encapsulates the predispositions embedded within algorithmic decisions. This bias can manifest through historical and structural inequalities that the model may inadvertently learn and perpetuate. Understanding and mitigating model bias involves examining the design of the algorithm and the assumptions underlying its predictive judgments. Subsequently, we shifted our focus to "data bias," a variable deeply entrenched in the data that serves as the foundation for machine learning. The representativeness and quality of the data are paramount, as any oversight can lead to AI systems that reinforce existing disparities. Data bias is often a reflection of the financial sector as it is, not as it should be, and identifying this discrepancy is crucial for the advancement of fair AI practices.



Understanding the systematic unraveling of these variables and their development paves the way for their calibration. The path we embark on necessitates a diligent presentation of each variable, guaranteeing that our trajectory adheres to the overarching objective of fairness.

0. Final equation:

$$\text{Fairness Score} = \text{Base Score} \times \frac{(1+\alpha \times F_a) \times (1+\beta \times F_d) \times DR \times PP}{(1+k \times \text{SPD}_{\text{normalized}}) \times (1+\text{Shannon Index}_{\text{normalized}}) \times (1+\text{FNR}_{\text{normalized}}+\text{FPR}_{\text{normalized}})}$$

Fairness scoring ranged from 0 to 10. A score closer to 10 indicated that the AI used in B2B was fair. Conversely, a score closer to 0 suggests that the AI is less fair:

$$f: A \rightarrow [0,10],$$

where A denotes the set of AI systems under consideration. Function f assigns to each AI system $a \in A$ a real number s, which represents its fairness score. A higher value (closer to 10) implies a higher fairness of the AI model, whereas a lower value (closer to 0) implies a lower fairness of the AI system.

For example, if a is an AI system with perfect fairness, *then* $f(a)=10$. Conversely, if a is the AI system with the lowest possible fairness, $f(a)=0$.

1. Fair process: variables

Base score

The calibration and engagement aspects within our fairness scoring framework are crucial for its effective implementation in AI systems used by financial institutions, particularly in the B2B sector. This facet of the model hinges on the dynamic relationship between the *base score* and concrete commitments of banks to enhance fairness, specifically in areas such as *Fa* (awareness factor) and *Fd* (documentation factor). The *base score*, initially set to 1, serves as a benchmark for assessing the initial level of fairness in AI models used by banks or financial institutions.

As financial institutions progress in their fairness initiatives, as reflected in improvements in factors such as *Fa* and *Fd*, the base score is designed to respond proportionally. Public authorities and regulatory bodies play a pivotal role in this process, with the discretion to adjust the base score in subsequent evaluations. For instance, one year after the initial

assessment (at time *t*+1), if a bank demonstrates substantial improvements in integrating fairness into its AI systems, the base score can be increased accordingly. This increase is meant to be a direct reflection of banks' engagement and advancements in fairness. Conversely, if there is no significant progress or decrease in engagement, the base score remains at its initial value, ensuring that institutions are neither unfairly penalized nor rewarded without merit.
An important characteristic of this model is its individualized and non-comparative nature. The *FinFair AI Index*, which measures AI fairness in the B2B financial sector, is a standardized and individualized measure. It provides a consistent framework for fairness assessment, while acknowledging the unique circumstances and efforts of each bank. Consequently, this precludes a direct comparison of scores or base scores between banks. The focus is on individual improvement and progress in fairness rather than competitive benchmarking.
Moreover, this approach encapsulates the symbiotic partnership between financial institutions' dedication to *de facto* fairness, and the guiding role of public authorities. By dynamically calibrating the base score in response to each bank's fairness efforts, the scoring model remains a relevant, effective, and equitable tool to advance responsible AI practices in the financial sector.

In our research, the concept of a *base score* within the fairness scoring equation plays a pivotal role, symbolizing the calibration of scoring in alignment with banks' commitment to fairness in their operations. This approach underscores a key aspect of our methodology: incentivizing and recognizing the efforts of financial institutions to achieve fairness in their AI systems.
Let *B* denote the *base score*. Mathematically, *B* is defined as a function of various fairness-related variables and their respective improvements over time. Specifically, the *base score* can be formulated as

$$B(t+1) = B(t) + \sum_{i-1}^{n} \Delta_i(t)$$

Where:

- $B(t)$ is the *base score* at time *t*.
- $B(t+1)$ is the *base score* at time *t*+1 (e.g., after a year).
- $\Delta i\ (t)$ represents the incremental change in the base score at time t owing to improvements in the i-th fairness-related variable.
- *n* is the total number of fairness-related variables considered in the model (such as *Fa* and Fd).

The incremental change $\Delta i\ (t)$ for each variable was determined based on the specific criteria established for that variable's contribution to overall fairness. For instance, improvements in *Fa* or *Fd* would lead to a positive increase in $\Delta i\ (t)$, enhancing the base score. This formulation allows for dynamic adjustment of the *base score* in response to ongoing improvements in AI fairness by financial institutions.

2. Data bias: variables

Awareness factor: Fa

To delve deeper into the role of *Fa* (awareness factor) in assessing AI fairness in the B2B financial sector, let us consider its significance and technical application in the context of this research.

$$\text{Adjusted Fa} = (1 + \alpha \times \text{Fa})$$

- *Fa* represents the degree of awareness and proactive measures an organization has regarding biases in AI systems (webinars, trainings, etc.)
- Mathematically, this is represented as *(1+α×Fa)* in the fairness equation. Here, *α* is a coefficient that amplifies the effect of *Fa* on the overall fairness score. This formulation suggests that a higher awareness correlates with a higher fairness score, acknowledging efforts to identify and mitigate biases.
- α is a coefficient that amplifies the effect of *Fa* on the overall fairness score. This coefficient signifies the extent to which the organization's awareness and efforts to address biases contribute to the fairness of the AI system.
- This approach also allows human involvement in the fairness process, as it is essential that humans are engaged in preventing the *FinFair AI Index* from becoming an exclusively technical matter. In addition, the OECD principles[5] on AI emphasize the integration of human oversight to ensure fairness, aligning AI systems with human implications. This included the implementation of human intervention and oversight mechanisms. In the B2B financial sector, adhering to these standards reassures fairness and aligns with auditing requirements and compliance.

Importance in the financial sector

- In B2B finance, decisions made by AI systems, such as loan approvals or risk assessments, have significant economic and social impacts. An elevated Fa indicates a financial institution's commitment to understanding and addressing potential biases in AI-driven decisions.
- This is crucial because biases in AI can lead to unfair practices, affecting businesses and economies. A high Fa score implies that the institution not only acknowledges these risks, but also actively engages in practices to ensure that their AI systems are as unbiased as possible.

Fairness approach: The awareness factor (*Fa*) plays a critical role in mitigating the risks associated with overreliance on mathematical models. This factor quantifies the extent of organizational awareness and proactive measures against biases in AI

[5] OECD. (n.d.). *Human-centred values and fairness (Principle 1.2)*. Human-centred values and fairness (OECD AI Principle) - OECD.AI. https://oecd.ai/en/dashboards/ai-principles/P6

systems, encompassing regular employee training and internal surveys. The inclusion of Fa in the fairness equation, often represented as (1+α×Fa), highlights the balance between mathematical quantification and human expertise. This approach counters the trend of overmathematisation, which historically emphasized reliance solely on models without human oversight, a stance epitomized by eminent figures such as Milton Friedman. The incorporation of Fa aligns with the principles of responsible AI, advocating a blend of human expertise and quantitative analysis to achieve fairness in AI applications.

In addition, the awareness factor (*Fa*) was assigned a weight to underscore the integral role of human discretion in the fairness evaluation process. This approach recognizes that fairness is not solely a product of models and data but encompasses the entire process, including human oversight. Reimagining fairness involves the synthesis of algorithmic logic, data, and public–private partnerships. Mandating corporate self-assessment in terms of fairness can drive improvement, as models inherently cannot fully capture fairness. The weight increase for *Fa*, particularly for efforts in biased documentation and awareness, signifies the long-term investment in education for the success of fair AI in B2B contexts. This evolution of fairness, transcending technological confines, advocates collaborative approaches to holistic and responsible AI development.

Documentation factor: Fd

In examining the role of *the documentation factor (Fd*) within the context of AI fairness in B2B financial services, it is pertinent to recognize its function as a quantifiable metric of an organization's commitment to recording and understanding biases that emerge during AI utilization.

$$\text{Adjusted Fd} = (1 + \beta \times \text{Fd})$$

Fd encapsulates the degree to which a financial institution document encounters bias in its AI system. Mathematically, this is denoted as *(1+β×Fd)* within the fairness equation, where β serves as the scaling coefficient that influences the impact of documentation efforts on the overall fairness score. Thus, the inclusion of *Fd* directly correlates an institution's dedication to transparency and accountability with its enhanced fairness score.

Importance in the financial sector: In the B2B finance realm, AI systems are pivotal in critical decision-making processes. Documentation of biases is not just a reactionary measure but also a proactive strategy. A robust *Fd* score indicates that a financial institution is not only aware of potential biases, but is also actively documenting these issues, which is vital in the iterative process of AI improvement. This commitment to documentation assists in navigating through the complexities of biases, ensuring that AI systems evolve to become fairer over time.

Fairness Approach: The index's novel incorporation of *Fd* reflects a comprehensive approach to fairness that surpasses that of traditional methods. It embodies a partnership between public regulatory bodies and private financial entities, advocating for calibrated measures of fairness that are contingent upon the size and technological prowess of the institution in question. The introduction of a weight, for *Fd*, could represent this nuanced calibration, aligning with a more sophisticated, multidimensional view of fairness that integrates human-driven documentation as a counterbalance to the potential 'overmathematization' of fairness.

Shannon Index

Incorporating the normalized Shannon index into the fairness equation is a significant advancement in assessing data diversity in AI systems, especially in the B2B financial context. This integration is a synthesis of theoretical understanding and practical applications, highlighting the crucial role of diverse data in facilitating fair AI decision-making processes.

$$H = -\sum_{i-1}^{n} p_i \log(p_i)$$

Where:

- H is the Shannon index.
- *where n denotes* the number of categories (or classes) in the dataset.
- pi is the proportion of the dataset belonging to the *ith* category.
- The summation ($\sum$) was performed for all categories in the dataset.
- $\log(pi)$ is the natural logarithm of proportion pi.

To normalize the Shannon index, we adjusted it to fall within a specific range, typically between 0 and 1. This is done by dividing the Shannon Index by the logarithm of the number of categories ($\log(n)$), which is the maximum possible value under the assumption of equal distribution across categories:

$$\text{Shannon Index}_{\text{normalized}} = \frac{H}{\log(n)}$$

The Shannon index, especially when normalized, quantifies the diversity and richness of a dataset. It calculates the proportional representation of each feature in the dataset. In our context, the features of this scoring system include the proportional representation of various financial actors[4] within the market. For instance, if a bank's transactional data with these financial actors are proportionally reflected in the datasets used to train the AI, it could potentially lead to a higher fairness score. Integrating this measure into the fairness equation is theoretically justified by the premise that greater data diversity leads to fairer AI systems. This diversity ensures exposure to various scenarios and conditions, thereby mitigating the biases inherent in more homogeneous datasets.

The Shannon index was computed as follows:

$H' = -\log(n)\sum(pi \cdot \log(pi))$

$$H' = -\frac{\sum(p_i \cdot \log(p_i))}{\log(n)}$$

where *pi* represents the proportion of each category, and *n* is the total number of categories.

Range of the Shannon index: The normalized Shannon Index ranges from 0 (no diversity) to 1 (maximum diversity). A higher value of H' indicated greater diversity within the dataset.

In summary, the inclusion of the normalized Shannon Index in the fairness equation is a methodological innovation that emphasizes the essential role of data diversity in AI fairness assessments. This provides a quantifiable measure to address the issue of data bias.

3. Model bias: variables

Statistical Parity Difference (SPD)

The Statistical Parity Difference (SPD) is an important metric in the fairness equation that gauges the balance of positive outcomes between different groups. It is particularly vital in B2B financial settings, where the AI's decision-making process may inadvertently favor certain types of businesses over others due to historical data trends or inherent algorithmic biases.

$$\text{SPD}_{\text{normalized}} = \frac{|P(Y=1|G_1) - P(Y=1|G_2)|}{1 - \min(P(Y=1|G_1), P(Y=1|G_2))}$$

Normalized SPD is calculated by taking the absolute difference in positive outcome rates between two distinct groups, divided by the maximum possible disparity. Mathematically, it can be represented as:

Where:

- *P(Y=1|G1)* is the probability of a positive outcome for the privileged group *G1.*
- *P(Y=1|G2)* is the probability of a positive outcome for the disadvantaged group *G2.*
- The absolute difference in these probabilities is divided by *1−min(P(Y=1|G1),P(Y=1|G2))*, which normalizes the SPD.

Range of SPDnormalized: The range of normalized SPD is between 0 and 1, where 0 indicates perfect parity between groups and 1 indicates the maximum possible disparity. A low value indicates little to no disparity in positive outcomes between groups, whereas a higher value indicates potential bias in the AI system that needs to be addressed.

Importance in the financial sector: For financial institutions, a fair AI system is one that does not skew its decision-making process based on the size of the bank, industry sector, or geographical location. Maintaining a low normalized SPD is vital to ensure that all businesses have equal treatment, regardless of the group.

Fairness approach: The inclusion of normalized SPD in the fairness score calculation reflects an advanced approach to ensure that AI models in the B2B context do not perpetuate or exacerbate existing inequalities. By striving to minimize normalized SPD, financial institutions can demonstrate their commitment to fair treatment and decision-making towards their counterparties, fostering a more trusted and fairer financial ecosystem.

Parameter K

Parameter *k* in the fairness equation encapsulates the equilibrium state of the fairness assessment, serving as a statistical baseline against which the impact of SPD normalization is measured. This represents the average or expected level of SPD normalized across the dataset, ensuring that the fairness score is grounded in the historical context of the AI system's decisions. Consequently, the selection of *k* is a strategic choice that aligns with the principle of maintaining a balanced and objective framework for evaluating fairness, ensuring that the model's assessment begins from a neutral standpoint before accounting for the variability introduced by the SPD. This neutrality is pivotal because it avoids skewing the fairness score towards either extreme of the dataset and allows for a fair comparison across different AI models and applications in the financial sector.

In the context of the fairness equation, parameter *k* is adapted as the *FinFair AI Index* evolves. The selection of *k* was based on several principles to ensure the robustness and relevance of the fairness score.

- **Avoiding division by zero**: *k* must be positive and sufficiently large to prevent division by a value close to zero when the normalized SPD is very low.
- **Preserving the score scale**: The magnitude of *k* is critical; it must be calibrated such that the fairness score remains within the range of 0 to 10. If *k* is too small, variations in SPD will disproportionately influence the score; if k is too large, the impact of SPD becomes negligible.
- **Base neutrality**: *k* may be set at a value that reflects a neutral starting point for fairness before considering the SPD.
- **Statistical analysis**: *k* can be determined based on historical fairness scores, calibrated such that the distribution of scores is centered around a value representing fairness.

Essentially, *k* should be congruent with a bank's historical data to prevent biasing the equation, allowing for a personalized result that is truly fair. For a rational choice of *k*, one might consider the mean or median of the historical or projected values for *SPD*normalized, or even half the range of SPD normalized if it varies from 0 to 1. This establishes *k* as a point of equilibrium: below it, *SPD*normalized enhances the fairness score relatively; above it, the effect is reductive. Using the mean or median ensures *that k* is representative of what is typical for the dataset, and hence, neutral.

Disparity Ratio (DR)

Disparity Ratio (*DR*) is a critical metric for evaluating fairness in AI systems, particularly in the financial industry. As a measure, it examines the essential question of whether the AI system treats different groups fairly when making decisions.

DR's significance stems from its ability to quantify disparities in positive outcomes between groups. AI systems might assess loan applications or detect fraudulent activities, ensuring that all types of banks are treated fairly. A disproportionate positive outcome rate, indicated by a *DR* significantly different from 1, could reveal inherent bias in the AI system.

The significance of *DR* in fairness scoring is two-fold. First, it aids in detecting potential biases present in the system. Second, it serves as a measure to guide rectifications in the AI models. By aiming for a *DR* value close to 1, financial institutions can improve the fairness of their AI systems, ensuring that decisions are made impartially across diverse groups.

DR is mathematically articulated as the ratio of positive outcomes (such as accurate fraud detection) between privileged (Group A) and unprivileged groups (Group B). *DR* is calculated as:

$$DPR = \frac{P(\text{Positive Outcome}|\text{Group A})}{P(\text{Positive Outcome}|\text{Group B})}$$

Where:

- *P(Y=1*|Group A) represents the probability of a positive outcome (such as correct fraud detection) for Group A.
- *P*(*Y*=1|Group B) represents the same for Group B.

The term "positive outcomes" refers to AI's accurate detection of fraudulent activity. *DR* is used to measure the relative effectiveness of the AI in different groups, such as those defined by business type or geographical location.

Range of DR: The DR value ranges from 0 to infinity. A value of 1 indicates no disparity, values less than 1 suggest fewer positive outcomes for the numerator group, and values greater than 1 indicate a potential bias in favor of that group.

Fairness approach: From a scientific perspective, *DR* addresses a gap often overlooked in traditional fairness metrics: the differential impact of AI decisions across diverse business entities. By quantifying the ratio of positive outcomes between the groups, DR challenges the

notion that a universally fair AI system can be achieved solely through accuracy and precision. For instance, an AI system in finance might show high accuracy in fraud detection; however, if this accuracy disproportionately favors larger enterprises over smaller ones, the system can hardly be deemed fair.

The deeper implication of *DR* in fairness scoring presents a subversive yet essential critique of the prevailing AI practices. It calls upon us to grapple with and rectify instances where AI professes objectivity and inadvertently perpetuates systemic biases. In the realm of finance, AI models are reassessed to prevent them from inherently favoring specific types of businesses, such as those with copious data availability, extended financial histories, overburgeoning entities, or new counterparties that may contribute innovatively to the market.

Predictive Parity (PP)

Predictive Parity (*PP*) in AI fairness, especially in applications such as anti-money laundering and fraud detection, is a key metric for evaluating the fair performance of algorithms. Its role in a comprehensive fairness equation is to ensure that AI predictions are accurate across all analyzed groups.

Predictive Parity is achieved when the probability of a positive prediction (such as detecting a fraudulent transaction) being correct remains consistent across all groups. For instance, if an AI system examines transactions from small banks (Group A) and systemic banks (Group B) and marks certain transactions as suspicious, *the PP* would require that the true positive rate, the percentage of correctly identified fraudulent transactions, be equivalent for both groups.

Mathematically, PP can be expressed as the ratio of true positive rates between privileged groups (Group A) and unprivileged groups (Group B):

$$PP = \frac{TPR_{G_A}}{TPR_{G_B}}$$

Where TPR_{GA} and TPR_{GB} are the true positive rates for groups A and B, respectively.

Fairness approach: Achieving Predictive Parity is inevitable for ensuring fairness in fraud detection. If one group has a significantly different true positive rate, it could indicate a bias in the AI model, favoring or disadvantaging one counterparty over another. Therefore, a balanced PP suggests that AI's predictions are reliably fair, regardless of the business type.

In the context of a fairness equation, *PP* complements *DR* by offering an alternate perspective. While *DR* assesses the balance of positive outcomes, *PP* focuses on the accuracy of positive predictions. It is crucial to consider both *DR* and *PP* in our *FinFair AI Index* to provide a holistic view of AI performance. Balancing these metrics ensures that no aspect of fairness is overlooked, thereby achieving a more comprehensive and fair AI system.

False Negative Ratio (FNR)

The False Negative Rate (*FNR*) is a metric for assessing the fairness and accuracy of AI systems, especially in the financial sector. It was calculated as the ratio of false negatives to the total number of actual positive cases. Mathematically, the *FNR* is represented as:

$$\text{FNR} = \frac{\text{False Negatives}}{\text{False Negatives} + \text{True Positives}}$$

Range of FPR: An *FNR* of 0 (min) signifies no false positive errors, whereas an *FNR* of 1 (max) indicates that all negative cases are incorrectly identified.

Fairness approach: Achieving a low False Negative Rate (*FNR*) is of paramount importance for fairness in AI systems, particularly in financial services, such as loan processing or fraud detection. A high *FNR* could indicate that the AI system erroneously rejects legitimate cases, such as denying loans to qualified customers or failing to detect genuine transactions. This can result in the unfair treatment of certain individuals or groups, potentially exacerbating existing biases. To guarantee fairness, it is imperative that the AI system minimize *the FNR*, thereby accurately identifying true positive cases. Incorporating this metric into the fairness equation enables a nuanced assessment of the model's sensitivity and capacity to refrain from unfairly penalizing legitimate cases.

False Positive Ratio (FPR)

False Positive Rate (*FPR*) is a metric for evaluating both fairness and accuracy. The *FPR* assesses how frequently an AI model incorrectly predicts a positive outcome for factually negative cases. Here, is the mathematical formulation of *FPR*:

$$\text{FPR} = \frac{\text{Number of False Positives}}{\text{Total Number of Actual Negative Cases}}$$

Range of FPR: An *FPR* of 0 (min) signifies no false negative errors, whereas an *FPR* of 1 (max) indicates that all positive cases are incorrectly identified.

Fairness approach: Similarly, maintaining a low False Positive Rate (*FPR*) is vital for fair AI decision making. In scenarios such as anti-fraud measures, a high *FPR* might result in falsely flagging legitimate activities as fraudulent, leading to unfair practices to counterparties. Such biases could disproportionately affect certain groups depending on the characteristics of the data on which the AI system was trained. Thus, the fairness approach involves rigorously fine-tuning the AI model to reduce *FPR*, ensuring that it does not unduly impact any group. Balancing *the FPR with* other fairness metrics is key to developing AI systems that are both accurate and fair.

4. Model bias: Groups

In addressing model biases, particularly after covering all variables of the *FinFair AI Index*, it becomes crucial to identify and categorize the groups upon which the framework will be applied. This entails calculating key fairness metrics such as the Statistical Parity Difference (*SPD*), Disparity Ratio (*DR*), Predictive Parity (*PP*), False Negative Rate (*FNR*), and False Positive Rate (FPR). These metrics serve as fundamental tools for assessing and comparing the performances of AI models across different groups. Comparing these groups is a critical commitment that banks must undertake to combat model bias. It is not just about recognizing disparities, but also actively engaging in their mitigation. To facilitate this process, we established a table outlining the groups and subgroups. This table serves as a structured guide for banks to systematically examine and address the biases within their AI models. By rigorously analyzing these metrics, financial institutions can ensure that their AI systems are fairer, thereby aligning with the broader goal of reducing bias in model outputs.

Category	Subcategory	Definition	Rationale
1. Transaction type Groups	**1.A Domestic vs. International**	Transactions within a single country vs. across borders.	Different fraud patterns due to varying regulatory environments and currency risks.
	1.B Interbank vs. Customer	Transactions between banks versus those between banks and their clients.	Different risk profiles and transaction scales between interbank and customer transactions.
	1.C Payment vs. Credit	Immediate payment transfers versus transactions that extend credit.	Complexity in risk assessment differs between straightforward and credit transactions.
	1.D High-Value vs. Low-Value	Grouping is based on the monetary value of transactions.	Transaction value influences risk assessment and the decision-making process.
2. Geographic Groups	**2.A Europe (EU and Non-EU)**	Transactions within EU countries vs. non-EU European countries.	Variance in fraud risk due to different regulatory standards in EU and non-EU countries.

	2.B North America	Transactions in the United States and Canada.	Unique financial regulations and fraud patterns in North America.
	2.C Asia-Pacific	Transactions in East Asia, Southeast Asia and Oceania.	Diverse economic systems and regulatory environments impact fraud risk differently in the Asia-Pacific region.
	2.D Middle East and Africa	Transactions in Middle East and African countries.	The distinct economic and political landscapes in these regions influence fraud risk.
	2.E Latin America and Caribbean	Transactions in South American countries and the Caribbean.	Various levels of economic development and financial regulations affect fraud patterns.
	2.F Emerging vs. Developed Markets	Based on the economic development status of the countries involved.	Emerging markets have different fraud risks and patterns than developed markets.
3. Sector-Specific Groups	**3.A Financial vs. Non-Financial**	Transactions involving the financial and non-financial sectors.	The financial sector has various risk profiles and regulatory requirements.
	3.B High-Risk vs. Low-Risk Industries	Based on the industry's risk level:	Industries such as construction and energy might have higher risk profiles than others.
	3.C Emerging vs. Established Industries	Transactions in emerging and established industries.	Emerging industries may present different financial behaviors and risks.
	3.D Regulated vs. Less Regulated Industries	Based on the level of regulatory oversight in the industry.	Industries with heavy regulatory oversight may exhibit different

			transaction patterns.
4. Transaction Size Groups		Group transactions by monetary value.	Large transactions might be treated differently by AI models than smaller ones.
5. Counterparty Behavior Groups	**5.A High-Risk Transactions**	Transactions with higher potential for financial loss, fraud, or illegal activities.	Involving high-risk countries or large atypical transactions.
	5.B Medium-Risk Transactions	Transactions with moderate risk level	Somewhat unusual but explainable transactions.
	5.C Low-Risk Transactions	Regular small-scale transactions align with known profiles in low-risk countries.	Transactions are considered safe or pose minimal risk.

Note: This table serves as the framework for conducting a comparative analysis of the *FinFair AI Index* calculations. It aims to precisely calculate key metrics such as the Statistical Parity Difference (SPD), Disparity Parity Ratio (DPR), and Predictive Parity (PP). By categorizing transactions into these groups, we can effectively evaluate and compare the performance of AI models across various dimensions of financial activities. This classification allows for a more detailed and accurate assessment of fairness in AI systems by providing a structured approach to identifying potential disparities in treatment and outcomes across different transaction types, geographic regions, sectors, and counterparty behaviors. Utilizing this table as a reference point, researchers can rigorously analyze the fairness of AI models, ensuring that they equitably handle the complex and diverse nature of financial transactions.

5. Data bias vs Model bias

Particularly, within the financial sector, a compelling argument emerges for prioritizing *model bia*s variables over *data bias* in the development of AI systems. This recommendation is founded on the premise that the scope of measurement in AI development is confined to the available data. Consequently, a strategic emphasis on *model bias* variables does not inherently amplify the risk of *data bias*. Rather, the incorporation of a diverse array of *model bias* variables should be considered a fundamental prerequisite for the development of fair AI systems.

This approach is underpinned by the understanding that the nature and quality of data collection are inextricably linked to the specific business context of the financial entity in question. It necessitates a dynamic and long-term strategy for accruing high-quality data. Furthermore, the focus on model bias aligns with a more *human-in-the-loop* approach. Unlike data bias, which is predominantly influenced by the financial institution's business model

and the time-dependent nature of data acquisition and refinement, *model bias* is more directly subject to the entity's control and intent.

This approach is particularly significant when considering the entry of new financial institutions into the market. A requirement for extensive, diverse, and high-quality data as a precondition for market entry could inadvertently establish prohibitive barriers for these emerging entities. By focusing on *model bias* —while not undermining the importance of *data bias* — the financial sector can foster a fairer business environment. It enables new market entrants to progressively work towards de facto fairness, evolving through sustained interactions and gradual enhancement of their data pool.

Nevertheless, the significance and weighting of these parameters must be carefully calibrated in accordance with the specific AI application—be it fraud detection, credit lending, or other financial services. This nuanced approach ensures that the established fairness metrics are aptly tailored to the unique demands and challenges of each application. The forthcoming sections of this research will delve into a comprehensive taxonomy, delineating a differentiated ratio and weighting scheme based on the AI application.

Weighting, justification and thresholds

Fairness Variables Weighting in AI Applications: Credit Lending vs. Fraud Detection

Variable	Definition	Credit Lending Weighting	Fraud Detection Weighting	Justification
FNR (False Negative Ratio)	Ratio of false negatives to actual positive cases.	High	High	Crucial for avoiding unfair loan denials in lending and for effective identification of fraudulent activities in fraud detection.
FPR (False Positive Ratio)	Ratio of false positives to actual negative cases.	Moderate	High	Less critical in lending but vital in fraud detection to avoid misclassifying legitimate activities.

SPD (Statistical Parity Difference)	Balance of positive outcomes between groups.	High	Moderate	Ensures non-discriminatory practices in lending; focus in fraud detection is more on accuracy.
DR (Disparity Ratio)	Ratio of positive outcomes between groups.	Moderate	Low to Moderate	Important for fairness in lending; less critical in fraud detection.
PP (Predictive Parity)	Equal true positive rates across groups.	Moderate	High	Ensures accurate identification of creditworthy individuals in lending; critical for equitable accuracy in fraud detection.
Shannon Index	Measures diversity of the dataset.	High	Moderate	Essential in lending for a representative dataset; important in fraud detection for a balanced approach.
Fa (Awareness Factor)	Degree of organizational awareness of AI biases.	High	Moderate	High in lending due to its impact on individuals; moderate in fraud detection to balance awareness

				with model precision.
Fd (Documentation Factor)	Degree of documentation of AI biases.	High	Moderate	Vital in lending for transparency; important in fraud detection for model refinement and compliance.

Weighting :

Low Weighting:

Numerical Range: *0 – 0.33 (on a scale of 0 to 1)*

Variables within this demarcated range are posited to exert a relatively marginal influence within the overarching structure of the model. Their role, while not negligible, is characterized by a subsidiary impact on the decision-making algorithm. This classification is integral to the hierarchical structuring of model variables, a practice that gains heightened significance in complex, multi-variate AI systems, especially those deployed in our financial contexts where numerous variables can potentially interact in a multifaceted manner.

Moderate Weighting:

Numerical Range: *0.34 – 0.66 (on a scale of 0 to 1)*

This bracket encapsulates variables that assert a substantive, albeit not dominant, influence on model outcomes. Such variables are significant in shaping the predictive outcomes of the model but are balanced against other variables to prevent any disproportionate skewing of results. This moderate classification serves to maintain an equilibrium within the model, ensuring that the predictive capacity is not overly reliant on any singular variable. This aspect is particularly pivotal in some cases, where the equitable representation of variables is crucial to maintain the integrity and accuracy of the model.

High Weighting:

Numerical Range: *0.67 - 1 (on a scale of 0 to 1)*

Variables classified within this range are deemed to be of paramount importance. They are the principal drivers of the model, critically influencing its output and decision-making process. The attribution of high weighting to these variables underscores their pivotal role in the model's functionality and the accuracy of its predictions. In the realm of finance-related AI applications, the implications of decisions driven by these variables are profound, necessitating meticulous validation and continuous scrutiny.

Justifications for Variable Weightings:

FNR (False Negative Ratio)

Credit Lending: High FNR could systematically exclude creditworthy applicants, particularly those from historically underserved backgrounds, perpetuating financial inequality and reinforcing historical biases.

Fraud Detection: High FNR can result in overlooking actual fraudulent activities, compromising the integrity of financial systems and potentially leading to significant financial losses and reputational risks.

FPR (False Positive Ratio)

Credit Lending: Moderately weighted, as false positives primarily lead to financial risks for the institution but less direct social impact. However, over time, high FPR can erode the trust in AI systems.

Fraud Detection: High FPR can lead to unnecessary investigations, strain resources, and cause inconvenience to customers, affecting the user experience and potentially leading to a loss of customer trust.

SPD (Statistical Parity Difference)

Credit Lending: Ensures fair access to financial services across different demographic groups, which is crucial for promoting financial inclusion and preventing systemic discrimination.

Fraud Detection: Moderate emphasis on SPD as the focus shifts towards the precision of fraud detection, though parity remains important to prevent systemic biases against certain groups.

DR (Disparity Ratio)

Credit Lending: Ensures that the lending decisions are not disproportionately favoring or disadvantaging any particular group, crucial for maintaining public trust and regulatory compliance.

Fraud Detection: Lower emphasis as the primary goal is the accuracy and effectiveness of fraud detection mechanisms.

PP (Predictive Parity)

Credit Lending: Ensuring consistency in the accuracy of predictions across different groups mitigates the risk of biased lending decisions and promotes fairness.

Fraud Detection: High emphasis as it ensures that the system's predictive accuracy is uniform across different groups, crucial for a fair and just financial security system.

Shannon Index (Diversity of Data)

Credit Lending: High diversity in training data is essential to capture the wide spectrum of applicant profiles, reducing the risk of embedding historical biases into AI systems.

Fraud Detection: Moderately weighted, as while diverse data is important for comprehensive fraud detection, the precision and adaptability of the model are more critical.

Fa (Awareness Factor)

Credit Lending: High awareness reflects an institution's commitment to identifying and mitigating potential biases in their AI models, crucial for ethical lending practices.

Fraud Detection: Moderately weighted as it's essential to understand potential biases, but the primary focus is on the technical robustness and adaptability of fraud detection algorithms.

Fd (Documentation Factor)

Credit Lending: Essential for maintaining transparency, accountability, and continuous improvement in AI-driven decision-making processes.

Fraud Detection: Important for regulatory compliance and for ensuring that the model's development and deployment are well-documented and can be audited for biases.

Fairness Scoring Threshold

Score Range	Category	Description
0-2	Very Unfair	Severe fairness issues, clear biases, lack of awareness or documentation (low human intervention).
3-5	Unfair	Some awareness and efforts, but notable biases and shortcomings.
6-7	Moderately Fair	Good awareness and documentation, room for improvement.
8-9	Fair	Largely fair, minor biases, strong commitment to fairness.
10	Perfectly Fair	Exemplifies highest standards of fairness, minimal to no biases.

5. Debate

Organic Solidarity towards *de facto Fairness*

In this research, we employed the Emile Durkheim's paradigm to establish our *FinFair AI Index*, aiming to transcend the conventional concept of *mechanical solidarity,* which is as the prevailing paradigm in the field of finance, particularly with regard to scoring in a general context. Concerning fairness, a notable absence of a definitive framework remains.
This paradigm shift is particularly evident when considering the existence of scoring models that are prevalent in the financial sector. These models are deployed to evaluate bank risks across diverse regulatory domains irrespective of the specific compliance focus. However, the present situation seems to discourage proactive engagement as it does not inherently foster a virtuous cycle conducive to diminishing systemic risk. To date, empirical evidence remains elusive to demonstrate that models such as the Internal Ratings-Based (IRB) approach effectively contribute to the mitigation of systemic risk.

The crux of the matter lies in the dual nature of these models, which are either overly standardized or excessively individualized, primarily owing to the latitude provided for the creation of bespoke internal models. This inherent rigidity precludes any meaningful interbank collaboration towards tangible reduction in systemic risk within the financial sector.

Furthermore, empirical research convincingly illustrates that during times of financial turmoil, the probability distribution transcends the normal curve, placing financial instruments in the extremities of the tails. Consequently, even assets with ostensibly disparate behavior dynamics under normal market conditions, such as equities and bonds, tend to exhibit correlated behavior during crisis periods.

Considering these challenges, our Durkheimian-inspired approach might hold promise in steering the financial sector towards more cohesive solidarity, a more unified and structured form of solidarity, particularly in the context of AI fairness scoring within the B2B landscape. The foundation of this promise is rooted in the *FinFair AI Index*, which might offer a standardized yet personalized framework. This framework effectively bridges the gap between regulatory or public sector initiatives and financial institutions.

Moreover, it facilitates the emergence of organic solidarity, paving the way for a more probable transition towards *de facto* fairness. This underscores the multifaceted nature of our approach, as it proactively addresses model biases and their regulation. In this context, regulatory authorities wield considerable influence on calibrating fairness scores. This calibration process hinges on banks' commitment to fostering awareness, comprehensively documenting biases, and accounting for data diversity through application of the Shannon Index. In the B2B context, fairness is primarily a collective concern before becoming a bilateral concern among banks, for instance, when utilizing AI models for fraud assessment.

Finally, the paradigm rooted in organic solidarity proves to be enticing, as the *FinFair AI Index* is meticulously structured to promote incentive-driven collaboration, eschewing competitive undertones. Significantly, the *FinFair AI Index* outcome serves as a discerning tool to assess the fairness score of the AI systems employed by individual banks. Notably, direct comparisons between such scores are infeasible, owing to the inherently tailored nature of the framework. Additionally, this paradigm exhibits an inherent incentive structure by facilitating constructive alignment between a bank's commitment to fairness, its control mechanisms addressing model bias and data bias, and overall performance. Within this academic framework, an enhanced comprehension of one's model and data translates into a more nuanced understanding of market evolution and transformation, thereby cultivating heightened awareness of counterparties. Ultimately, this enriched knowledge base contributes to the optimization of performance within commercial relationships.

This paradigm shift is not merely an academic exercise; it addresses the real costs faced by financial institutions owing to regulatory divergence. The findings reveal that regulatory divergence incurs substantial financial costs, averaging between 5% to 10% of the annual turnover for financial institutions[6]. These costs create a significant burden that reduces the possibility of achieving higher performance, adding urgency to the need for innovative approaches, such as the Durkheimian-inspired framework. It not only champions fairness, but also offers a practical and financially advantageous pathway for banks to navigate the challenges of risk assessment and regulatory compliance, thereby potentially offsetting the weight of these incurred costs.

Operational risks for the *FinFair AI Index* Framework

Operational risks for the *FinFair AI Index* Framework	**Description**
Minor challenge	Despite the framework's intention to discourage direct score comparisons, a potential challenge arises from the financial institutions' ability to assess the evolution of fairness scores over time. However, given the sector's familiarity with complex mathematical indices, any disruption to the framework's efficiency is unlikely. This issue, while minor, warrants

[6] IFAAC. (2018, February). *Regulatory divergence: Costs, risks, impacts IFAC*. https://www.ifac.org/_flysystem/azure-private/publications/files/IFAC-OECD-Regulatory-Divergence.pdf. https://www.ifac.org/_flysystem/azure-private/publications/files/IFAC-OECD-Regulatory-Divergence.pdf

	ongoing monitoring of the long-term effects on the organic solidarity paradigm.
Moderate challenge	Determining a suitable regulatory authority to oversee the framework is a moderate challenge. The choice between a regional or global regulator and the establishment of a dedicated international agency requires careful consideration. This situation may necessitate a one-to-many approach, obliging all financial entities to provide their models and data for exclusive analysis by a designated agency. Such a decision ensures the accurate calibration of the *FinFair AI Index* and the ongoing monitoring of variable developments within each bank. The regulatory oversight issue requires thoughtful deliberation to guarantee the effective and impartial administration of the framework.
Major challenge	A substantial challenge arises from disparities among banks in terms of access to AI resources and funding to address model and data biases. Additionally, the Shannon Index's reliance on data diversity raises questions when banks must adapt their datasets to accommodate evolving business models. However, it is essential to recognize that these challenges may yield benefits, particularly for banks with evolving business models, diversified relationships, and distinct criteria. Over time, they may accumulate a more diverse dataset, potentially bolstering their fairness scores in the long term, albeit not necessarily in the short term. Addressing this major concern requires addressing resource disparities and potentially re-evaluating the application of the Shannon Index in varying business contexts within the financial sector.

Fairness vs Performance

In the milieu of economic downturns, it is a well-documented phenomenon that average performance across financial markets tends to wane, a trend often ascribed to endogenous

dysfunctions within the financial system itself. These dysfunctions typically arise from flawed organizational structures or suboptimal interactions among the constituent entities. This perspective aligns with the broader understanding that crises are frequently the result of internal systemic issues rather than external shocks. Consequently, the integration of enhanced fairness into AI systems is posited as a significant countermeasure against such endogenous dysfunctions. By reducing information asymmetry and fostering an environment of heightened trust and transparency, particularly through the deployment of fair AI systems, the financial sector can mitigate the adverse impacts of these crises. Empirical research corroborates that the diminution of information asymmetry leads to enhanced business opportunities[7], largely due to the resultant bolstered trust among market participants. This diminution is pivotal in facilitating a more discerning and fair treatment of diverse groups, thereby engendering a heightened level of trust among stakeholders. Contemporary research substantiates the assertion that the amelioration of information asymmetry bears a direct corollary to an amplification of business opportunities, fundamentally attributed to the bolstered trust resonating among market participants[8].

In addressing the integration of fairness in AI systems within the B2B financial sector, it becomes evident that this endeavor does not contravene the objectives of market performance. Instead, it serves as a complementary and essential aspect of the financial ecosystem. The pursuit of enhanced fairness in AI is inexorably linked to one of the quintessential challenges in finance: the diminution of information asymmetry.

The deployment of AI systems characterized by heightened fairness is instrumental in disseminating more comprehensive and transparent information about the profiles of financial counterparties. This enhanced clarity plays a pivotal role in mitigating the traditionally pervasive issue of information asymmetry in financial markets. By yielding a more nuanced understanding of the entities engaged in financial transactions, these sophisticated AI systems catalyze informed, fair decision-making.

Consequently, an increase in trust and confidence among market participants emerges, fostering a milieu conducive to improved market stability and performance. This scenario underscores the strategic significance of fairness in AI within the financial sector, transcending beyond mere compliance imperatives.

Thus, the integration of fairness into AI systems within the financial sector should be perceived not as a divergent goal but as an integral component that enhances market performance.

[7] Gudelytė, L. (2015). On the impact of information asymmetry on evaluation and risk of cluster performance. Socialinės Technologijos, 5(01), 32-43.

[8] Hicks, J. R. (1936). Keynes' Theory of Employment, Interest and Money. The Economic Journal, 46(182), 238-253.

6. Conclusion

This study has explored the transition from mechanical to organic solidarity in the B2B financial sector. By developing the *FinFair AI Index* framework, we have introduced a comprehensive tool aiming to score fairness of AI. Our research underscores the importance of collaboration between public authorities and financial institutions in ensuring fair AI systems.

The *FinFair AI Index* framework emerges as a pivotal innovation in advancing both procedural fairness and de facto fairness in the financial sector, by implementing a non-competitive and individual fairness scoring system ranging from 0 to 10. This framework transcends traditional approaches by not only ensuring compliance with regulatory standards but also by fostering an environment of *organic solidarity*. This concept of solidarity signifies a deeper integration and cooperation among financial institutions. The *FinFair AI Index* extends beyond the confines of a simple regulatory tool as it serves as a catalyst for a paradigmatic shift, wherein the interactions between banks become crucial for the cultivation of fairness.

Our framework, which accounts for both model biases and data biases, represents a significant step toward addressing the complexities associated with AI fairness. The inclusion of parameters such as the Disparity Ratio (DR) and Predictive Parity (PP) in our analysis has allowed for a more detailed and sophisticated understanding of how fairness can be quantified and improved in financial decision-making systems.

Moreover, our study highlights the significance of continuous adaptation and improvement in AI models, reinforcing the need for ongoing AI development against biases. By documenting and actively addressing these biases, financial institutions can not only comply with standards but also enhance their individual development towards fair AI systems and performance.

In conclusion, this research contributes to the growing discourse on AI and fairness in finance, providing valuable insights and tools for financial institutions to navigate and improve the fairness of their AI systems. It is our hope that the *FinFair AI Index* framework and our findings will inspire further research and action in the pursuit of a fairer financial system in the B2B context.

7. References

- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732. Retrieved from http://www.californialawreview.org/wp-content/uploads/2016/06/5-Big-Datas-Disparate-Impact.pdf.
- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943. Retrieved from https://arxiv.org/abs/1810.01943.
- Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., ... & Zafar, M. B. (2021). Fairness measures for machine learning in finance. The Journal of Financial Data Science
- Durkheim, E. (2010). From mechanical to organic solidarity. In Sociology: Introductory Readings (Vol. 2, No. 1, pp. 25-29).
- Durkheim, E. (2023). The division of labour in society. In Social Theory Re-Wired (pp. 15-34). Routledge.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226. https://doi.org/10.1145/2090236.2090255.
- Frontiers Editorial Office. (2020). The impact of artificial intelligence on firm performance: An application of the resource-based view to e-commerce firms. Frontiers in Psychology, 11, 580406. https://doi.org/10.3389/fpsyg.2020.580406
- Gudelytė, L. (2015). On the impact of information asymmetry on evaluation and risk of cluster performance. Socialinės Technologijos, 5(01), 32-43.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315-3323. Retrieved from https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.
- Hicks, J. R. (1936). Keynes' Theory of Employment, Interest and Money. The Economic Journal, 46(182), 238-253.
- Hildebrandt, M. (2018). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. Theoretical Inquiries in Law, 20(1), 83-121. https://doi.org/10.1515/til-2019-0004.
- IFAAC. (2018, February). Regulatory divergence: Costs, risks, impacts. IFAC. Retrieved from https://www.ifac.org/_flysystem/azure-private/publications/files/IFAC-OECD-Regulatory-Divergence.pdf
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. Proceedings of the 8th Innovations in Theoretical Computer Science Conference. https://doi.org/10.4230/LIPIcs.ITCS.2017.43.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635. Retrieved from https://arxiv.org/abs/1908.09635.

- OECD. (n.d.). Human-centred values and fairness (Principle 1.2). In Human-centred values and fairness. [Online]. Retrieved from https://oecd.ai/en/dashboards/ai-principles/P6.
- United Nations Educational, Scientific and Cultural Organization. (2021). Recommendation on the Ethics of Artificial Intelligence. [PDF file]. UNESCO. https://unesco.org/EthicsAI
-
- Zafar, M. B., Valera, I., Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. Proceedings of the 26th International Conference on World Wide Web, 1171-1180. https://doi.org/10.1145/3038912.3052660.
- Zhang, Y., & Zhou, L. (2019). Fairness Assessment for artificial intelligence in the Financial Industry. arXiv preprint arXiv:1912.07211. Retrieved from https://arxiv.org/abs/1912.07211.