Deep Science Publishing
https://doi.org/10.70593/978-93-49910-41-6_9

Chapter 9: Data quality, bias, and ethics: Challenges in algorithmic credit decisions

9.1. Introduction

Artificial intelligence and machine learning are enabling financial institutions to expand access to credit. It has allowed for a democratization process that makes credit available to millions of 'real people with real families with real dreams.' Beyond the philosophical issues brought by the use of AI, a critical aspect of any credit decision is trust: that we can trust the bank to identify who can or cannot access credit; and that credit decisions are not serving as vehicles to sift.

Unreliable data may drive undesirable outcomes. Among the most severe negative externalities is that financial institutions may inadvertently sort out plaintiffs for premiums that make access to credit expensive. Moreover, there are financial markets that are underbanked and for which AI and alternative data sources have the potential to open doors for thousands of individuals. However, as the risks are barely non-identical commingling that remains unresolved, developing counter-majoritarian machines with broader training can deliver unjustified or over-stochastic outcomes. The gender, ethnicity, and nativity biases are not an example; they are just the tip of the iceberg. Underbanking and affordability of finance are major issues faced by many individuals and families around the world, and the advancement of more considerate AI algorithms aiming at minimizing information asymmetries within the financial sector can contribute to tackling these problems (Barocas et al., 2019; Binns, 2018; Cowgill et al., 2021).

9.1.1. Overview of the Topic

This short symposium piece discusses the implications of machine learning and artificial intelligence for consumer credit markets. It starts from a brief overview of underwriting,

that is, the job of deciding who should be given credit. The discussion then focuses on a key question that has come to prominence more recently: when are algorithmic credit evaluations likely to be legal and when should we worry that they are toxic? The piece highlights three potential reasons why fully automated underwriting might perform differently from a human decision maker: standard biases may be magnified; human drivers of credit risk may be missed; and algorithmic response to data quality may be inconsistent with non-discrimination norms. The consequences of these differences for economic justice and consumer welfare are considered, and the tentative conclusions of the symposium are sketched.

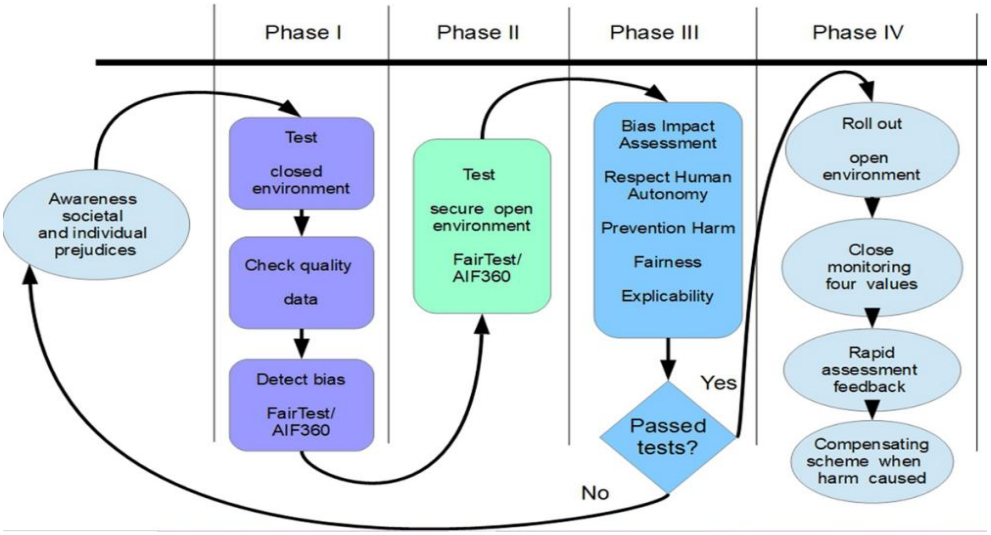

Fig 9 . 1 : AI bias

9.2. Understanding Algorithmic Credit Decisions

Algorithmic credit decisions are increasingly being used by governments, businesses, and nonprofit organizations for a variety of purposes. At its core, the algorithms inspect an individual's personal attributes, such as their credit history, debt-to-income ratio, zip code, etc., and then decide whether to grant that person credit and under what terms. The goal of these decisions is to maximize revenue or minimize cost for the organizations. The algorithms predict the risk of people defaulting on their loans through a statistical model trained on past and present demographics and credit histories in this situation. Lower predicted risks correlate with better loan terms and provide the incentive to pay off their loans in full.

In many countries, questionnaires, interviews, behavioral science, statistical and quantum methodologies, and data provided by organizations were utilized as the tools to

explore the accuracy, limits, and possible biases of the credit risk algorithms currently used. Three general examples included: understanding how local demographics and life events affect credit scores on average for different groups of people; learning how these models, trained on consumers' and small businesses' credit data, decide whether a new application should be approved or declined; and exploring the existing systems that attempt to establish businesses' and large entities' creditworthiness through a public image provided by partnerships and mergers. In each case, associations between the system's exposure to various customer events and the decision outcomes were investigated (Eubanks, 2018; Hurley & Adebayo, 2017).

9.3. The Role of Data Quality

Our theoretical model connects data quality to reduced-form prediction error and to the potential for harm produced by an algorithm's policy. That sharp theoretical illustration becomes realistic in our applied context. We link the issue of biased integrated imbalance loss prediction with the commonly discussed concepts of completeness, accuracy, consistency, and timeliness, thereby providing a straight-line ratio control interpretation while following the path in previous studies. Then, we provide a case study of building mortgage credit scoring models based on data from a dynamic financial sector. The results indicate that timeliness is the key quality factor in the credit scoring environment. The relationship between data quality and the cross-sectional accuracy of credit scoring models becomes non-linear, and a heuristic way for simulating a decomposed performance for imbalanced classification problems is suggested. Since this study powered the context in which miscounting is prohibitive, it might be possible to motivate the cost of providing high-quality imbalanced data versus the punishment of using low-quality imbalanced training data. Such implications indicate that a miscounting penalty should generally provide an intrinsic incentive to introduce sufficient quality control measures throughout the data production cycle of statistical processes.

9.3.1. Defining Data Quality

The term "data quality" (DQ) is often used within descriptive frameworks, but it has only been explicitly defined by a few authors. A comprehensive overview of these definitions outlines shared properties and explicates that data quality refers to the specific context of data use and the relevance of data to fulfill a purpose, along with which data characteristics and their accuracy, reliability, timeliness, and completeness need to be understood. "Due to the specific character of data, which are generated by processes in systems and their use in other processes, the notion of data quality, for instance, goes beyond conventional notions of quality," or "the fit of data playing a role in a task to be

performed by the user and the situational context." A two-level, eight-dimensions framework is still being introduced at conferences and in international reports and ensuing analyses of existing DQ frameworks. Our own prior use of this framework in a major legal data trial showed that the proposed DQ dimensions are, in fact, highly relevant to the quality analysis of data used in algorithmic credit decisions.

9.3.2. Impact of Poor Data Quality

An extension of data quality is data discrimination. In the financial industry, individuals have no uniform right to a good credit rating or to receive credit on decent terms. Therefore, trying to systematically separate the useful data from the noise can introduce problems with the assumptions we can make about the resulting filtered data. Moreover, the tools lenders use to collect day-to-day information about customers can sometimes help them, impair them, or break their trust. Especially since people often prefer not to talk about their financial situation, a variety of company services, points systems, and associated marketing must be noticed. These systems act as powerful tools to collect the data lenders demand. In addition, elements that make people even more reluctant to discuss their problems with a financial institution, such as delicate personal or social contexts or some frankly stigmatizing status labels in lending companies' terminology, fade in conventional creditworthiness assessments.

In the face of the rapidly increasing importance of digital footprints, it seems unlikely that companies will change their emphasis on actual instead of recommendatory labeling, either voluntarily or in response to legal action. Systematic discrimination and data noise will therefore remain the lenders' problems. Even more fundamentally, there may be an approximately Gaussian distribution of people's ability to respond fluidly to demands that the banking system seeks to lay on its customers, while the normal distribution does not describe the actual use of and confidence in the system. The very banking services and products that the central nervous system is looking for may be the ones that people actually want to cover up or evade. Therefore, the current state of conventional bank systems is fundamentally in contradiction with the requirements of the financial system.

9.4. Types of Bias in Algorithms

The use of algorithmic adjudication in consumer lending potentially exacerbates long-standing biases in credit decision-making. In this section, we offer an overview of different types of potential bias due to reliance on machine learning in lending. The potential sources of bias stem largely from two characteristics of algorithms. First,

because algorithms work mainly with data, they inherit all the biases in real-world data. Second, they have the potential to introduce new bias.

Overall, the potential biases can be classified as arising from group-related protected demographic attributes, non-protected group-similarity-related bias, and group-similarity biases. Protected attribute bias refers to the possibility of illegal discrimination when decision-making groups are defined based on legally protected characteristics. Non-protected group bias refers to bias based upon classes that are similar to protected groups but not specifically protected based on a protected attribute. Group-similarity biases result from classifying applicants into 'in-groups' and 'out-groups' and rewarding or punishing based on similarity within.

These are but a few examples of deriving, detecting, and quantifying bias and fairness in machine learning. The sources of biases are many, complex, and interrelated. In practice, highlighting the unfairness observed in a credit-related predictive model alone is insufficient to derive the guidelines and interventions to manage biases. Instead, a deeper understanding of whether and to what extent data, human decisions, or algorithm outputs are responsible for observed discrimination is necessary.

9.4.1. Data Bias

The use of big data – unstructured or high-dimensional databases such as social media data, search queries, news stories, and other internet activity signals – and the ability to infer creditworthiness based on crowd predictive signals are likely to play key roles in the future of credit scoring. It is often faster and cheaper to process a loan through automated scoring technology than it would be through traditional credit adjudication that requires an application to be reviewed by a loan officer and priced based on interest rates or capital models. But one crucial concern is that the use of big data may exacerbate any underlying biases that large proprietary data sets have.

Inaccuracies, multiple identities, privacy concerns, and other data quality issues are potential reasons to be concerned about the quality of alternative lending data, but another concern is the presence of inherent biases. For example, many similar 'non-traditional' data sources – like mobile phone records, utility bills, rental payments, and digital footprints – that firms tap to gauge consumers who do not have traditional credit files are likely to contain inherent biases that would affect the effectiveness of scores estimated from the limited data that could be associated with such groups. Indeed, such data are likely to capture snapshot elements of consumer behaviors and not the broader behavior of individuals who may have had funds in traditional banking institutions but have become unbanked or underbanked. In the former case, economic decision-making capability provided to access traditional financial markets would be out of reach.

9.4.2. Algorithmic Bias

Algorithmic bias has been under scrutiny in the context of lending for quite some time. Early research asked whether the "black box" properties of learning algorithms produce discriminatory errors and, even if the weights of discriminatory factors such as race, age, or gender are minimized in the learning process, concluded that the simple fact that the output of an algorithmic model relies on those factors provides a potential source for unfairness. This conclusion is in line with theoretical work showing that there is always such a location-based price discrimination when suppliers have some pricing discretion. Researchers have also provided empirical evidence of discriminatory outcomes of algorithmic systems. The body of literature has forced governments, intergovernmental organizations, and autonomous organizations alike to produce guidelines regarding the use of these techniques.

When abuse of these models due to the categorical, biased outcomes became widely acknowledged, researchers and engineers developed methods to ideally correct them, using techniques such as post-processing, re-calibration, and optimizing with fairness in mind. Even potential threats of abuse of these corrections resulted in research regarding the ethical implications of using different types of corrective algorithms, pointing out that the choice of technique implicitly tilts the ethical evaluative angle. More recent work also asks when residual discriminatory outcomes are, in fact, fair when acquiring the training data is expensive and discriminates against wealthy individuals.

9.4.3. Human Bias in Data Collection

Another potential source of bias in information available for credit scoring is that some information is collected from individuals involved in the credit granting process. The required applicants could be biased towards information collectors such as loan officers or branch managers, and these collectors themselves could have consumer biases that affect selection. They find evidence that internal loan officers with more ethnically diverse borrower pools are less likely to overstate the number of dependents in a credit application and less likely to classify non-self-employed individuals.

Trust in credit information might be bounded rationally. Findings on selection suggest this could occur when collectors are statistically discriminatory, that is, the probability of biased information collection is higher for certain types of loans. The analysis of this coincides with the findings using the same unique loan-level data set containing detailed credit report contents. Individuals moving along the financial inclusion line are more likely to withdraw their consent so that their data can be utilized in lending determinations. While this could be consistent with a decrease in trust with an increase

in data use, an alternative explanation is that data disclosure demonstrates a borrower's creditworthiness by revealing new information.

9.5. Ethical Considerations

The literature shows that the governance of data usage and credit decision models is an important topic due to fears arising from the use of artificial intelligence, data protection, and consumer discrimination. The agent applicants may never understand the data decision process. Credit decision-making should include not only profitability goals but also factors such as transparency, accountability, non-discrimination, and responsibility involving the use and governance of data. Large differences are observed when people judge fairness and when the algorithm evaluates and suggests proposals. This is an aspect indicating that the profit-maximizing credit scoring algorithm can be led astray with massively wrong recommendations in lending decision-making. It predicts not only the default value in their loans but also creates scenarios in which they will be bad business. Furthermore, and more relevantly for discriminatory issues in the present study, the public-private information environment provides a type of

9.5.1. Fairness and Equity

Optimal credit policy should balance goals that are often in tension. It should treat like borrowers similarly, to be transparent and fair; it should aim to charge more to those who are likelier to default or whose default costs are particularly high, to be equitable and to allow for more worthwhile investments. These goals can come into conflict. It is precisely in addressing these sorts of trade-offs that ethical and policy considerations should center. They don't substitute for proper empirical validation of effects or for the crafting of appropriate competition policies, but in a credit decision-making system that has these features, the people making decisions and asking 'why' can do so more knowledgeably. They are in a better position to design the compensation scheme and supervisory structures so as to align business incentives with the combined desiderate of greater precision, fewer errors, greater equity, and greater transparency.

Balancing these ethically significant elements is likely to require balancing interests in competition in the supply of business-focused machine learning models with attention to ethics, including by: pursuit by business more generally of corporate social responsibility; provision of fiduciary advice privately and in the form of regulatory rate-making to find an optimal trade-off among particular fairness, accountability, transparency, and debiasing goals; and protection of the bodies best situated for machine learning, both privacy and data protection bodies, and competition authorities. Any regulatory body will, of course, face intense challenges, including litigation, were it to

implement these suggestions, but the public-private negotiations could allow for decisions that optimize more finely across many of the relevant goals and work to ensure that more of the factors, including ethical concerns, that are relevant in a tradeoff-laden credit decision-making context are considered and given appropriate weight.

Fig 9 . 2 : Fairness and equity

9.5.2. Transparency and Accountability

1. Average Predictive Power of Data Elements: Organizations employing credit scoring models use complex models that combine thousands to well over 100,000 data elements, with the number of data elements generally inversely related to the degree of predictive accuracy. Lack of transparency regarding the content and degree of predictive accuracy of models limits the ability of consumers, companies, and regulators to assess the fairness of their results. The predictive accuracy of individual data elements in the model may be a better indicator of the value of a source of discriminatory content than simply the presence of a relationship to the model's output.

2. Impact of Data Elements on Decision Outcomes for Various Groups: Organizations employing credit scoring models do not often know how an organization's policies would be applied in practice or even which groups the algorithms would treat differently. It is common not to know the precise means by which the physical and demographic files are combined, when links are established, or even when the systems are generating credit reports from this data. Given the wide array of data that is used and the expansive number of reporters, this creates not only the potential for identifying technical or perceptual accuracy problems but also calls into question the extent of reasonable access or correction rights.

9.5.3. Informed Consent

In informed consent, each individual who will be the subject of the credit decision, by his or her action, explicitly allows and understands the factors and data being used to make the decision, the outcomes and biases, and the risk of harm. This approach assumes that the tiered approach of exception existence, use, and understanding are overcome. As a result, an important question when dealing with those in the denied population or the one that may be harmed is: "If the subject being reviewed truly knew what was being used and how it was being used and the outcome, would they consent?" This approach, which does not ameliorate the inherent biases and proxies being used but rather requires that each subject accept and understand the outcome and potential for harm, has some strong support in the algorithms, especially when a score is being interpreted in a way that requires qualification under the new general data protection regulations.

For financial decisions, especially around credit assessment, informed consent seems ethically complex and perhaps not just. For instance, there are possible and potentially serious negative consequences both for the individual and for others if major purchases are delayed or choices made based on fewer or more expensive credit options lead to inferior results. And weighing the individual financial cost and potential negative outcomes, are people really in a position to suffer diminution for non-participation? Even if informed consent were given voluntarily, the valid will of the thought process would act in a way to overcome the margin condition; an implication that the individual would not be hurt by their participation but rather revealed and visible from the beginning. Would it be considered ethical to make a decision on whether to approve based on invalid or inaccurate data? Unfortunately, life can sometimes be difficult, contradictory, and have few good choices.

9.6. Regulatory Frameworks

This article has focused on the calculation of credit scores, the treatment of borrower characteristics, and statistical accuracy. But these important outcomes only reflect part of the social and ethical concerns linked to algorithmic credit decisions. Because there is limited regulatory oversight of AI algorithms and production data in private settings, differentiating laws and guidelines in both contractual and execution domains has strong potential to impact decision-making through considerations of compliance close to the source of the critical ethical concerns. Legal penalties for breach and workplace norms are discussed briefly; this is in part because sociological theory identifies the possibility that private actors will self-regulate through workplace norms to protect against legal penalties, as well as the potential impacts on design within a specific social context where actors are in direct communication. In finance, data and technology are informally regulated. Organization-specific contractual arrangements with third-party vendors can

be initiated for specific subareas, each with their own terms and conditions. The contracting process often involves a significant amount of negotiation, which has the potential to impact social norms at the individual institutional level and promote workplace norms for ethical concerns. Do governmental entities have an expanded role in algorithmic decision-making in the manner of a landlord that has evolved potent capabilities in land use governance? Given organizational, regulatory, and research needs surrounding the design and use of AI algorithms, what is the optimal role of AI in both functions?

9.6.1. Current Regulations

While regulations such as the Equal Credit Opportunity Act and the Fair Credit Reporting Act protect people from being discriminated against in credit decisions, they do not touch upon the use of algorithms to arrive at such decisions. As a result, digits have enabled magazines, rather than make up their minds. Several industry and media reports have noted that the current regulatory environment is not fit to cover the use of AI in credit decision making and that not all of the enumerated factors may be divulged by everyone generating credit scores either. The present regulatory guidance, however, is incredibly strict in some situations and incredibly vague in many others.

The Federal Trade Commission's Policy Statement on Unfairness states that an act is unfair if it causes or is likely to cause substantial injury to consumers, if the harm is not outweighed by countervailing benefits to the consumer or to competition, and if the consumer could not reasonably have avoided the injury. Yet, this definition has not challenged any specific credit algorithms that we know of. Furthermore, it is noted that although credit scoring algorithms use over 1,000 factors, credit reports only show a handful of those. This discrepancy between the public and private sectors makes it difficult for credit score recipients to figure out their errors and improve their scores.

9.6.2. Proposed Changes

It is accepted that the process of building credit scoring models and the types of data utilized are based on statistical procedures and are likely to increase the importance of information about members of different groups who have similar objectives. The result is that some individuals with similar potential credit profiles are given disparate credit opportunities based on their group status or the characteristics of the neighborhoods where they live. The common counterargument is that objective information is utilized in the development of credit models, and this is thought to remove any fault or bias stemming from the social connections between the variables and group status.

We believe that this line of argument requires reconsideration. We have demonstrated that some algorithmic models rely on data that is likely influenced by societal prejudices and is accepted as unchallenged or as an authentic proxy for creditworthiness, even in the absence of a significant link, weakness, or absence of links between credit risk and the group status which is deemed to arouse prejudices. Non-personally identifiable information such as the marital status of applicants, for instance, is used to create mostly distinct group statuses in favor of a favored group without this serving as an indicator of creditworthiness. We propose algorithms to comply with a moral compass in terms of its design, which does not allow for imaging that is in conflict with existing credit policies.

9.7. Mitigating Bias in Algorithmic Credit Decisions

Lending decisions made by automated algorithms can exhibit bias against certain groups. These biases often result from the data upon which these algorithms are trained. More critically, the outputs of these algorithms can reinforce and cause ongoing harm to existing inequality in lending markets. Evidence shows that it reacts to the gender of the applicant and her beauty. Specific loans are found favorable for certain races. Lenders assess a loan applicant's potential based on personal names, not only on credit history, income, financial stability, and related facts.

Given the potential for unfair and deleterious outcomes, society has a compelling interest in reducing bias in amount-allocating algorithms. In other words, without accurate, reliable data, the outputs will be biased and the application of the technology can further harm society. These issues present an important question: how can we lessen these effects? There are a number of ways to mitigate algorithmic bias in credit decisions and ensure that the outcomes of such algorithms are ethical. The following outline approaches can be applied to mitigate the impact of potential bias: 1. Fairness in Law and Data Regulations 2. User-Centered Human-Machine Collaboration 3. Adapted Data Quality Concepts 4. Transparently Understanding Data Quality Requirements

9.7.1. Techniques for Reducing Bias

While a comprehensive review of solutions to reduce the bias in the decision-making process is beyond the scope of this chapter, we discuss some common techniques that can be adapted to consumer credit decisions too. Fair machine learning uses different mathematical definitions of fairness that align with specific substantive concerns in different application areas and closely related algorithms to optimize the fairness of model outcomes. These techniques can be combined to reduce bias for different groups of consumers while preserving the financial viability of the underwriting process itself.

Model compression makes a larger, more accurate model act like a smaller, simpler model, letting us take advantage of the performance gains while mitigating end effects, such as different levels of transparency, that can creep in with large gate models. Lastly, post hoc calibration allows for ease of model interpretation after the model has been standardized. This can lead to a form of public confidence in the model through procurement transparency. All of these methods are described initially in a more generic context, but are then transformed into a credit underwriting setting with realistic financial requirements and data considerations in mind.

9.7.2. Best Practices for Data Collection

The benefits of algorithmic decision-making for purposes such as credit and lending depend critically on the quality of the underlying data. As a result, the financial services industry already applies a wide range of data quality processes in the credit risk assessment process, both for internal and regulatory reporting processes, as well as in relation to broader privacy, data protection, and data security concerns. In the case of negative financial outcomes attributed to biases in algorithmic decision-making, the lack of any consensus on best practices for data acquisition and treatment stands in sharp contrast to financial services industry practices for credit risk assessment of individuals. Currently, the field of AI ethics emphasizes the potential negative biases in these applications of AI and attempts to develop technical tools to detect and reduce them.

While we agree with the analysis that data characteristics, such as potential bias and discrimination, have a first-order impact on the appropriateness of algorithmic decision-making applications, we believe more proactive development of best practices for data collection and treatment (and of processes for supervising the production of data) will help to address the perceived weaknesses embedded in today's algorithmic decision-making processes. As a result, in this section we focus on such best practices, starting with an in-depth discussion on the relevance and applicability of data protection-based data quality requirements for data-driven algorithmic decision-making; potentially helpful regulatory guidance exists in other areas as well.

9.8. The Future of Algorithmic Credit Decisions

The increasing sophistication of algorithmic decision processes offers much promise specifically for credit scoring. The sector for personal loans alone amounts to billions of dollars. The use of alternative data, including nontraditional sources of personal information, could further improve the predictive value of credit scores and enable more informed credit decisions. The historical development and the mandatory inclusion of

specific types of data in the credit scoring process, like information from renting a personal living space, show that societal developments can indeed affect credit scoring.

Costs for producing accurate credit scores have already been reduced considerably. When the technology is further developed and the attention to bias, ethics, and data quality increases, it could even lead to an improved assessment of creditworthiness and fewer mismatches between lenders and borrowers—reducing the black box effect. However, achieving these welfare-enhancing results is far from straightforward without accompanying policy measures. Since natural persons will also interact more frequently with sophisticated technology in the future, they should be more involved in the construction of value-based algorithms—similar to entities profiting from externalities. A counterweight to the overwhelming power of channeling companies is recommended. Reformulating the prohibition to refuse credit on an arbitrary basis to a prohibition to refuse credit based on arbitrary features of a prospective borrower provides guidance without hampering all innovation. The human level of the proposed resolution will increase.

It is important to consider that an increasing sophistication of credit scoring is not inevitable and does not automatically produce welfare-enhancing results. While machine learning technologies are able to process large amounts of heterogeneous data, which could make the assessment of creditworthiness more accurate, it is often stated that the improvement should benefit both lenders and borrowers. However, not all loan applicants will benefit and the assessment of one's creditworthiness is only one part of extending credit. As illustrated by information asymmetries and distributional effects, credit markets are not only malfunctioning, but they are also subject to distributional effects, which implies that the use of sophisticated techniques may lead to negative effects for some potential borrowers.

9.8.1. Emerging Technologies

Support for the production of this chapter was provided by the Sloan Research Foundation. I am grateful for the financial support and to Gideon Mann and Solon Baracas for valuable input. All opinions are mine. I acknowledge the guidance and support received from all at the Fintech reading group and NYU School of Law.

Split data and model segregation is a technique to train an AI algorithm using sensitive attribute data without creating a service built on it. In essence, a company deploys an architecture with three parts: 1. A model that predicts an attribute from only information not linked to the sensitive field; 2. A segmenter that subsequently executes non-sensitive attribute information into multiple class sub-models solely trained on it; and 3. A classifier that assigns the individual segments. Although the prediction error can increase

in the model because of the loose linkage between the classifier and the segmenter, a looser linkage helps satisfy the model's demands. If client care primarily involves utility rather than privacy defending the consumer, this technology produces a more privacy-friendly tool. Ideally, firms operating in domains with worries about bias in the training data could utilize it to develop client-side instruments. The asymmetric segregation of data and model can also influence the type of architectures that developers are inclined to use.

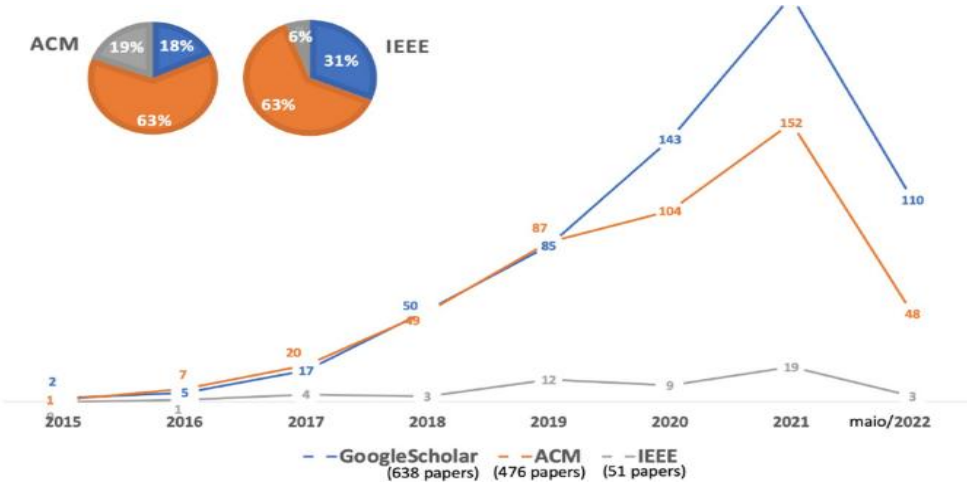

Fig 9 . 3 : Algorithmic discrimination in the credit domain

9.8.2. Trends in Regulation

This decision comes against a backdrop of increasing regulatory, market, and social attention paid to algorithmic bias in credit scoring and AI systems more broadly. In the United States, certain users of consumer reports, such as mortgage lenders, must take "adverse action" notices under certain circumstances. Moreover, federal banking agencies are required to make sure that the automated underwriting system used to generate the reports will be evaluated and validated under established standards. This includes ensuring that automated systems are designed and maintained in a manner that promotes the use of credit scores to support safe and sound or efficient lending, credit, and insurance, in addition to no adverse by, or differential impacts on, applicants or borrowers against the applicant or borrower on any prohibited basis.

Similar obligations to provide notice of automated decisions to individuals and to explain these decisions to them upon request exist, including the nature of the data used and the logic underlying these decisions. Indeed, others advocate that algorithmic bias should be

carefully regulated in credit scoring. These regulations follow well-established standards in law that disallow discrimination in consumer credit markets impacting protected classes. Moreover, when regulation imposes such obligations on credit reporting agencies and their business customers, they are compelled to spend resources on ensuring the quality of the data and models, even when there are financial incentives to act in a biased manner. These laws foreshadow a broader discussion of AI ethics that prioritize minimizing harm to society and avoiding the use of AI for indecent purposes, and not just lessons learned during the use of AI in credit markets.

9.9. Conclusion

In this paper, we studied the potential inclusiveness and credit impact insights of financial machine learning AI and the risk of credit risk prediction accuracy, algorithm-driven by the credit decision being influenced by data bias. Critically, the performance of a credit risk model importantly assesses data integrity and whether the model agrees more significantly with the rule of law than the use of defined weather variables. To address the data integrity of the financial machine learning model that we have considered as a benchmark, we employed common data preprocessing techniques and found critically significant improvements in accuracy and model performance over benchmark models that reflected practical challenges. Separately, we began analyzing the causality of historical financial inclusion and credit prediction models, highlighting potential biases and predictions as to which region and lender might be impacted.

We also tested data synthesis that generates an ambiguous credit score interpretation, unable to loan experienced human insights such as which data keywords contribute to the high likelihood of loan applications being declined. We introduce a new methodology that studies the model that ensures interpretations based on a concept that the model produces a causal effect of different data features versus the outcome, which we interpret as making us more inclusive. The method creates a new variable, called the explainer variable, whose value provides insights into the model's prediction. To demonstrate the importance of accepting models in finding systemic weaknesses in financial inclusion and credit decision models, we offer a transparent approach to interpreting black box machine-learning AI in financial applications.

9.9.1. Final Thoughts and Future Directions

Overall, conducting consideration of bias, fairness, and the ethical implications of different data-driven technologies is made possible by the fact that so many of them are currently in contexts where interpersonal human judgment has existed in the past. Where this is true, a wide variety of previous sociological and psychological studies are easily

reinterpreted as potentially being relevant to their creation, and the institutional and social structures to engage in their management and oversight simultaneously exist. That said, there are important differences between human decision making and decision making based on statistical algorithms -- no matter how ultimately dependent on the humans who create, deploy, and maintain those algorithms. Arguably, the biases in human decision making have simply gone understudied because their entanglement with broader social factors make them far more difficult to study and control for. Issues like those that are now raised in the context of algorithmic ethics -- manipulability, transparency, the fusion of predictive and prescriptive elements in judgements, and the relevance and justification of sub-group level judgements -- all already exist in deliberative concepts of stereotype and prejudice.

The precision and NPV/NPV style balance that makes ROC curve analysis so attractive is probably the fundamental reason why it has both a mathematical and formal policy utility. While their analogous curves have been proposed, the fact that neither the balance nor the ROC curve itself are consistent in a meaningful way greatly reduces their usefulness. Indeed, ROC and NPV/NPV curves considerations are so fundamental to the field that NPVs are written as continuous functions of the cut-point one uses to define different groups. The most serious error associated with this perspective is the fact that default-sensitive decisions like those involved in credit risk largely treat what is ultimately a continuous signal as something categorical. To change this point of view, however, introduces what is called the “loose wire problem” -- the point where one cuts the ROC/NPV or other curve that is only ultimately of importance by the presence of some other, unrecoverable, external signal.

References

Hurley, M., & Adebayo, J. (2017). *Credit Scoring in the Era of Big Data.* Yale Journal of Law and Technology, 18(1), 148–216. https://doi.org/10.2139/ssrn.3012535

Binns, R. (2018). *Fairness in Machine Learning: Lessons from Political Philosophy.* Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149–159. https://doi.org/10.1145/3287560.3287583

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning.* fairmlbook.org, 1–212. https://doi.org/10.48550/arXiv.1908.09635

Cowgill, B., Dell’Acqua, F., & Deng, S. (2021). *Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics.* Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1–20. https://doi.org/10.1145/3476061

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin’s Press. https://doi.org/10.5040/9781350987034