PROBABILITY
& PARTNERS

AI Fairness in Financial Services

Trust at the Center

By Alexandru Giurca

July 2020

www.probability.nl

Summary

Financial institutions are rapidly embracing Artificial Intelligence (AI) in their day to day operations in many different areas ranging from HR and fraud detection to credit scoring and trading. Some applications of AI, such as hiring and loan decisions, can be highly sensitive to the so-called protected attributes such as gender or race. Particularly in these areas, where AI is increasingly evolving, fair treatment of individuals must be assured.

It is well known that AI applications introduce new risks by exhibiting significant biases – possibly unintentionally discriminating against individuals. Examples of such biases have recently reached the mainstream media and a powerful consensus is emerging among policymakers regarding principles for the ethical and fair application of AI. Algorithms might amplify societal inequities represented in the data. To avoid reputational damage, as well as to comply with recent regulation regarding AI fairness, financial institutions must ensure they recognize and mitigate these biases when applying AI.

This paper is intended to be a starting point in understanding why AI ethics has to be regarded also as a board level issue and why assuring fairness in AI models is a complex process that requires both, technical as well as business considerations. Financial institutions must establish a cross-functional AI governance framework covering the notion of fairness in the entire AI lifecycle. Further, concepts of ethics and fairness need to be incorporated in existing Model Risk Management practices.

Probability & Partners has a long-standing operational expertise in building and deployment of AI models. We provide guidance on ethical considerations in AI model development and model risk management and advise you on the optimal balance between model performance and fairness. Further, we assist in reviewing your AI governance framework and help you to embrace trustworthy AI in order to harness the full potential of these powerful new technologies.

Introduction

The financial services industry is going through profound change and disruption. Artificial Intelligence rapidly found its way into the industry in the last years and there is no denying that this disruptive technology offers immeasurable benefits. AI can increase operational and cost efficiencies and enhance business transformation including more tailored customer engagement. Possible applications of AI range from credit decision making and fraud detection to customer service, marketing, wealth management, risk management and trading. Momentum for AI will continue to build as data becomes more available, algorithms improve, developer skill sets broaden and computing power accelerates.

Both on a European level and on a local level, the use of AI has caught the attention of regulators and supervisory authorities. The European Central Bank[1] and the European Banking Authority[2] as well as De Nederlandsche Bank[3] have recently taken an active interest in AI, and while recognizing its benefits, they are increasingly raising their voice on potential unforeseen risks and unintended consequences of its use. Especially in financial services – a data driven business on the one hand, a regulated and privacy-concerned business on the other hand – the request for assurance, interpretability and **fairness of AI algorithms** rises.

Over the past few years, society has become increasingly aware of unintended discrimination of artificial intelligence systems – with harmful results. Recognizing these risks and reducing them is an urgent priority. Ethical issues arise from the fact that AI has difficulty distinguishing between "good" and "bad", not being able to take into consideration conscience or ethical values. Unless properly supervised, AI – learning from historical data – is prone to exacerbate existing ethnic stereotypes and bias-filled decisions made by humans.

In this paper we highlight the principle of **Fairness** – being crucial for financial institutes given the potential financial and regulatory risks, but even more concerning tremendous reputational risks – and why it should be on top of mind for senior executives. We give guidance on how organizations can address **AI ethics** more proactively from a business perspective as well as in the model validation and model risk management.

Use of AI in Financial Services: The Road Ahead

According to the French Prudential Supervision and Resolution Authority (ACPR) financial services is one of the sectors that invested most heavily in Artificial Intelligence. Its implementation is notably more advanced in the banking sector than in the insurance sector[4].

Although AI is used mainly for enhancing operational processes (e.g. text mining to extract useful information from documents), the trend indicates that **applications material to customers are expanding**[5]. Among these – according to the DnB in the Netherlands – there are customer interacting **chatbots and virtual assistants** (e.g. ASR's website), real-time **transaction monitoring** (ABN AMRO's "Grip") and **Fraud Detection** (e.g. "FRISS"), **credit decision making** as well as Customer Relationship Management with regard to **product recommendations** (Salesforce's "Einstein") [3].

In the insurance sector AI is used for **underwriting**, e.g. in car insurance for the analysis of driving behaviour (e.g. Fairzekering), providing better and faster car damage estimates[7] or supports managers to **predict coming churn** by providing a prioritized list of clients who show signs of considering cancelling their policy[9]. Furthermore, the World Economic Forum (WEF) foresees AI applications to advise clients on prevention strategies to lower their risk profiles and to provide predictive analytics to clients that help them better understand their risks[8].

Use Case: Fraud Modelling

1 One application of AI is Fraud detection and prevention, replacing or enhancing solutions based on predictive and prescriptive analytics.

2 This type of application utilizes a continuous stream of data (e.g. banking transactions, loan applications)

3 The software can then notify a human of any deviations from the normal pattern, so they may review it.

4 The human agent can accept or reject this alert, which signals to the AI model whether or not it was correct to indicate fraud.

5 This further trains the AI model to "understand" that the deviation it found was either fraud or a new, acceptable type of deviation.

In several areas of finance Natural Language Processing, Speech and Image Recognition are evolving. Examples include analysing credit applications to gauge the creditworthiness of a customer (**credit approval or prepayment modelling**) from their digital footprints[3] or estimating car damage based on photos[5]. Banks and insurance companies increasingly use AI to **screen job applicants** and chatbots to make better and faster hiring decisions[10], thus reducing the cost of human workers in searching for the right candidates by up to 90%[12]. Furthermore, although not material to customers, AI is present in algorithmic trading (e.g. ING Katana for bond trading).

Regulatory Developments regarding AI Fairness

AI is increasingly used in many high-stakes decisions in financial services. However, it gives rise to unwanted, and often serious, negative consequences. Besides the already existing regulatory landscape – including cybersecurity and data privacy considerations – the responsible implementation of AI has been a growing topic of discussion in regulatory circles. These put their focus on potential hidden risks, which are difficult to anticipate, identify and quantify. These risks are mainly attributed to **ethics** and **fairness** of outcomes of AI applications[13].

> "**Discrimination** (intentional/unintentional) occurs when a group of people (with certain shared characteristics) is systematically more adversely affected by a decision (e.g. output of an AI model) than another group, in an inappropriate way."
>
> *European Banking Authority*[3]

In the context of AI, ethics represents all ethical issues in the development, deployment and use of AI applications. Its central concern is the impact on people's lives and society as a whole[1]. An approach to AI ethics should be based on the fundamental human rights and democratic values enshrined in the EU Charter[14] and Dutch Equal Treatment Act (AWGB, Section 5) [15] – described by reference to human dignity, freedom, non-discrimination and equality as well as social justice and internationally recognised labour rights.

We observe that the calls for the ethical, fair and responsible use of AI are at the centre of the regulatory expectations:

ECB: "Trustworthy AI"[1]

Trustworthy AI has three components, which should be met throughout the system's entire life cycle:

(1) it should be **lawful**, complying with all applicable laws and regulations
(2) it should be **ethical**, ensuring adherence to ethical principles and values and
(3) it should be **robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm.

- OECD: human-centred AI guidelines in May 2019, promoting the responsible stewardship of Trustworthy AI while insuring respect for human rights and democratic values[16].
- G20: AI principles mostly focused on the benefits for people and society[17].
- ECB: Guidelines for "Trustworthy AI" to put forward human-values based approach to AI[1].
- De Nederlandsche Bank: "General principles for the use of Artificial Intelligence in the financial sector" with 6 principles ("SAFEST") for the use of AI by financial institutions[3].

To date, regulatory initiatives related to fairness have focussed on appropriate safeguards and human oversight to avoid discriminatory outcomes. The focus is currently shifting towards a quantitative assessment of bias.

"SAFEST" Principles (De Nederlandsche Bank)[3]

1 Soundness

AI applications should be reliable and accurate, behave predictably and operate within the limits of regulations.

2 Accountability

Financial undertakings need to be accountable for the use of AI applications. Firms should demonstrate that they understand their responsibility for AI applications.

3 Skills

Organisations should ensure that senior management, risk management and compliance functions have an adequate level of expertise, especially in understanding AI limitations.

4 Fairness

AI should not unintentionally disadvantage certain groups of customers. Firms should define their concept of fairness and demonstrate how their systems behave accordingly.

5 Ethics

Firms should ensure that the outcomes of their AI systems do not violate own ethical standards. Stakeholders should not be mistreated or harmed.

6 Transparency

Firms should be able to explain how they use AI in their business processes, and how these models function (where reasonably appropriate).

Why Fairness Matters?

Ethical issues with AI can have a broad, major impact. Developing AI that accurately performs on the whole spectrum of human diversity is more likely to benefit a broad group of potential customers. Besides regulatory risk, ethical flaws can damage a more intangible but priceless asset – the company's reputation – as well as the lives of employees, customers and society as a whole. In the last years there have been numerous instances of AI discrimination highlighted in the news, resulting in fines and legal cases[19], but most importantly in loss of credibility:

In October 2018, Liberty Mutual Insurance Europe was fined £ 5.2 million by the Financial Conduct Authority (FCA) for poor oversight of a third-party supplier, which enabled them to provide mobile phone insurance including all claims handling functions. The insurer's overreliance on voice analytics AI software led to some claims being unfairly declined[20].

In November 2019, the Apple Card received a lot of criticism as women received significantly less credit than their spouses even though they shared the same characteristics[21]. This problem tied back to Apple's AI credit decision algorithm that determined an individual's credit limit – without proper fairness and explainability measures – based on an analysis of people living in the same area who shop at the same stores. Besides leading to tremendous investigations, the incident caused major reputational damage, both to the Apple Card as well as to Goldman Sachs – the bank that backed the Apple Card.

Other striking examples of discrimination include Amazon realizing in 2018 that their AI based recruiting tool was not rating candidates for technical jobs in a gender-neutral way by considering certain hobbies associated to men more promising for good candidates. Ultimately the tool was temporary abandoned, but Amazon is continuously working to introduce ethical standards in the long term[22].

Ethics

AI has to be built with good intentions aiming for positive benefits for the society by not harming vulnerable populations.

Legal

Being in line with existing legal non-discrimination frameworks and regulatory requirements is a top priority due to extensive penalties.

Profit

Next to the performance, fairness has to be considered in order to target all customers adequately and not miss profits.

Trust

Biased algorithms can reduce customers' satisfaction, and cause lasting brand damage. It is difficult to measure if trust can be recovered.

Bias reduces the potential of AI for business and society by encouraging mistrust and producing distorted results. Business and organizational leaders need to ensure that AI systems improve on human decision-making – by taking care of traditionally disadvantaged groups.

Bias in AI is a Human-made Problem

Human decisions are difficult to objectify[23], since humans may lie about the factors they considered or they may be unconscious about their biases[24]. AI decision systems can be used to augment human judgement and reduce both conscious and unconscious biases. Unlike human decisions, decisions made by AI can be examined, and interrogated. However, since AI systems heavily rely on large amounts of data when learning patterns ("training"), they can embed and amplify human and societal inequities represented in this data. As AI algorithms are only optimization programmes, without morality and ethical values, the underlying chosen **data rather than the algorithms themselves is most often the main source of discrimination**.

> "A major problem is that minority groups by nature are often underrepresented in datasets, which means algorithms can reach inaccurate conclusions for those populations."
>
> *Sorelle Friedler, Haverford College*[27]

For example, extracting information from news through natural language processing may exhibit gender stereotypes found in society[25]. AI based hiring algorithms used to scan job applications might mistakenly screen out female applicants, if the historical data used to train it reflects past decisions that resulted in few women being hired. If a general credit lending model is trained on data, which is not representative for the whole population (low amount of young people), and finds that younger individuals have a higher likelihood of defaulting, it might reduce lending based on age. Society and legal institutions may consider this not only to be illegal age discrimination, but also a strong demotivation of young entrepreneurship.

Addressing Bias is Complex

Teaching the AI system what is right or wrong is very complex, since the removal of existing societal prejudices present in the data is a challenging task. Especially when using Natural Language Processing and Image Recognition involving customers, leaders should be well aware of discrimination[28]. To address potential discrimination, the first step is to question **what unintentional biases might exist** in a modelling use case and how they might manifest themselves in the data.

A system is not always unfair if a protected class is favoured in its outcomes. For example, if a bank is aiming to target only students and young persons with a special loan type, then age bias towards elder persons is not meaningful to assess, however racial or gender bias might occur. In this case loan decision outcomes are spread intentionally (through the modelling task) only across young persons.

Protected Attributes (NL)

Gender, Age, Race, Pregnancy, Religion, Political Opinion, Nationality, Citizenship, Sexual Orientation, Civil (marital) Status, Disability Status.

As there is not a single notion of fairness[29], the **suitable definition has to be chosen in the context of the specific use case** and **bias has to be measured according to them.** Once potential biases are identified, companies can mitigate them by using specific algorithms to debias the data or by regarding fairness in the training procedure.

Unfortunately, there is a trade-off between algorithm performance and fairness. Different bias mitigating methods have different negative impact on ML algorithm performance, depending on the amount of data available. However, if the right debiasing technique is chosen specific to the use case and data at hand, the negative impact can be greatly decreased. In general, collecting more data may increase algorithm performance further, while also improving fairness. Besides, the consideration of fairness in algorithms has a broader positive business impact that cannot be easily quantified. Therefore, business leaders need to balance this trade-off.

The simple removal of protected attributes does not guarantee the fairness of a model, since many other attributes might correlate with it. This makes it easy to reconstruct a protected attribute such as race even if it is ommited in the training of the AI algorithm. For example, it is known that certain living areas are mostly populated by certain ethnical minorities. Including data about residency may lead to indirect discrimination.

AI Ethics as a Board-Level Issue

Any AI-related ethical issue can carry broad and long-term risks – **reputational, financial and strategic** ones. Therefore, it is prudent to engage the board to address AI risks. Ideally, the task should fall to a technology or data committee of the board. Such a committee can help to maintain and assess how far an organisation is in assuring everyone is treated fairly. It evaluates whether an application could be prone to potential bias and ensures necessary documentation of the AI and data used.

Instead of avoiding the use of AI, organizations should ensure it is **built and used responsibly** – not only complying with applicable laws, but also with the companies' ethical code, principles and goals – and balance performance with fairness. Fairness should not be only the responsibility of data scientists. It is important for business leaders and data scientists to work together to select the right notion of fairness for the particular decision that is to be made and design the algorithm that best meets this notion.

> "Organizations need to assign the **accountability for AI applications** and the management of associated risks at the board of directors level."
>
> *De Nederlandsche Bank*[3]

Given the interconnected nature of AI risks, financial institutions should establish a **dedicated cross-functional AI governance** with clear understanding of roles and responsibilities (including first-line accountability) and coverage across multiple independent risk management functions (e.g., model risk management). AI governance frameworks should take into account the whole AI lifecycle, from design and training, to implementation and continuous testing of the final system. This includes emergency controls in place to intervene against, switch off or roll-back an AI model. **Human monitoring** of outcomes and processes should remain aligned to the risk appetite and objectives of the firm.

Leaders need to frame and ensure adoption of a thoughtful process around "**fairness by design**"[1]. They should emphasize their organization's **diversity values** and make sure they are communicated into analytics teams. Heterogeneity in AI oversight brings multi-disciplinary perspectives, thus helping mitigating inherent societal biases.

New implications for Model Risk Management

Existing Model Risk Management (MRM) practices need to be modified for AI models and concepts of **bias and fairness have to be included**. These considerations must be addressed across AI model development, implementation, validation and use. Fairness has to be assessed only for applications that are material to stakeholders and is highly dependent on the use case. As bias is predominantly attributed to the underlying data, particular focus has to be put on the **input data**. Existing data management frameworks have to be enhanced to assess the scope of data, improve data quality and strengthen data monitoring processes. Validators need to check whether model developers have taken steps to ensure fairness – through qualitative as well as quantitative investigation.

In order to achieve fairness, bias has to be measured and mitigated according to a chosen fairness notion at each stage of the **model-development process**. Next to a strict assessment of the **used attributes**, the AI **model interpretability** has to be guaranteed depending on the type of decision informed by the algorithm and the potential impact on the customers concerned. Since certain AI models modify their parameters dynamically with new incoming data, it has to be decided whether a **dynamic calibration** is appropriate and if yes, whether fairness is still assured. Often, vendor and **third-party models** are available as "black-box". In this case the model validation models consist of outcomes analysis, sensitivity analysis and benchmarking – especially in terms of fairness.

References

1. High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Brussels: European Commission.
2. European Banking Authority (2020). EBA Report on Big Data and Artificial Intelligence. Paris.
3. Joost van der Burgt (2019). General principles for the use of Artificial Intelligence in the financial sector. Amsterdam: De Nederlandsche Bank.
4. ACPR (2018). Artificial Intelligence: challenges for the financial sector. Paris: Autorité de Contrôle Prudentiel et de Résolution.
5. FSB (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Basel: Financial Stability Board.
6. Tractica (2019). Report: Artificial Intelligence Market Forecasts. Available at: https://tractica.omdia.com/research/artificial-intelligence-market-forecasts/
7. DnB and AFM (2019). Artificial intelligence in the insurance sector: an exploratory study.
8. WEF (2018). The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem. World Economic Forum.
9. Mutanen, T., Nousiainen, S. & Ahola, J. (2010). Customer churn prediction - A case study in retail banking.
10. Feloni, R. (2017). Consumer-goods giant Unilever has been hiring employees using brain games and artificial intelligence — and it's a huge success. Businesss Insider. Available at: https://www.businessinsider.nl/unilever-artificial-intelligence-hiring-process-2017-6
11. HireVue (2019). AI in Recruiting: What it Means for Talent Acquisition in 2019. Available at: https://www.hirevue.com/blog/ai-in-recruiting-what-it-means-for-talent-acquisition
12. Tammenga A. (2020). The application of Artificial Intelligence in banks in the context of the three lines of defence model. Maandblad Voor Accountancy en Bedrijfseconomie 94(5/6): 219-230. https://doi.org/10.5117/mab.94.47158
13. European Parliamentary Research Service (2020). The ethics of artificial intelligence: Issues and initiatives.
14. European Parliament (2012). Charter of Fundamental Rights of the European Union 2012/C 326/02.
15. College voor de Rechten van de Mens (2005). Equal Treatment Act (AWGB). Utrecht.
16. OECD (2019). The OECD AI Principles. Available at https://www.oecd.org/going-digital/ai/principles
17. G20 (2019). G20 Ministerial Statement on Trade and Digital Economy. Available at: https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf
18. General Data Protection Regulation (2018). GDPR Recital 71. Available at https://gdpr-info.eu/recitals/no-71/
19. Borgesius, F.Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Council of Europe
20. Financial Conduct Authority (2018). The FCA has fined Liberty Mutual Insurance Europe SE £5.2 million for failures in its oversight of mobile phone insurance claims and complaints handling. Available at: https://www.fca.org.uk/news/press-releases/liberty-mutual-insurance-europe-se-fined
21. Vincent, J. (2019). Apple's credit card is being investigated for discriminating against women. The Verge. Available at https://www.theverge.com/2019/11/11/20958953/apple-credit-card-gender-discrimination-algorithms-black-box-investigation
22. Dastin J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
23. Kleinberg, Jon and Ludwig, Jens and Mullainathan, Sendhil and Sunstein, Cass R., Discrimination in the Age of Algorithms (February 5, 2019). Available at SSRN: https://ssrn.com/abstract=3329669.
24. Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." American Economic Review, 94 (4): 991-1013.
25. Google Developers (2018). Text Embedding Models Contain Bias. Here's Why That Matters. Available at: https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html
26. Healey, K., Woods Jr., R. H. (2020). Ethics and Religion in the Age of Social Media. New York: Routledge, https://doi.org/10.4324/9780367824181
27. Levin, S. (2016). A beauty contest was judged by AI and the robots didn't like dark skin. The Guardian. Available at: https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people
28. Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding-Royer, E.M., Chang, K., & Wang, W.Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. ArXiv, abs/1906.08976.
29. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. ArXiv, abs/1908.09635.
30. US Equal Employment Opportunity Commission (1998). Uniform Guidelines on Employee Selection Procedures

How can Probability & Partners help?

Probability & Partners has a long-standing operational expertise in building and deployment of ML models. This gives us a unique edge also in ability to validate such models.

We can help your organization with full stack validation of AI or ML solutions irrespective of the area of application. We create a tailormade validation that best suits the materiality, importance for business and stage of deployment of your AI/ML model(s).

We can help you not only with analysing algorithms/data for biases, but also advising on the optimal trade-off between performance and bias. Probability & Partners have experience with a so called “black-box” validations that are performed without obtaining the full model documentation or source-code. Such set-up is ideal for IP-sensitive validation assignments.

Dr. Svetlana Borovkova

Head Quantitative Modelling

svetlana.borovkova@probability.nl

Pim Poppe

Managing Partner

pim.poppe@probability.nl

PROBABILITY
& PARTNERS