INNOVATING WITH CONFIDENCE: EMBEDDING AI GOVERNANCE AND FAIRNESS IN A FINANCIAL SERVICES RISK MANAGEMENT FRAMEWORK

Michelle Seng Ah Lee,[†] *Luciano Floridi*[††] *& Alexander Denev*[†††]

ABSTRACT

An increasing number of financial services (FS) companies are adopting solutions driven by artificial intelligence (AI) to gain operational efficiencies, derive strategic insights, and improve customer engagement. However, the rate of adoption has been low, in part due to the apprehension around its complexity and self-learning capability, which makes auditability a challenge in a highly regulated industry. There is limited literature on how FS companies can implement the governance and controls specific to AI-driven solutions. AI auditing cannot be performed in a vacuum; the risks are not confined to the algorithm itself, but rather permeates the entire organization. Using the risk of unfairness as an example, this paper will introduce the overarching governance strategy and control framework to address the practical challenges in mitigating risks AI introduces. With regulatory implications and industry use cases, this framework will enable leaders to innovate with confidence.

† University of Oxford, Deloitte UK; michelle.lee@oii.ox.ac.uk.

†† University of Oxford, Alan Turing Institute; luciano.floridi@oii.ox.ac.uk.

††† University of Oxford, Deloitte UK; adenev@deloitte.co.uk. This research was partially supported by Deloitte (MSAL and AD); and by Privacy and Trust Stream - Social lead of the PETRAS Internet of Things research hub - PETRAS is funded by the Engineering and Physical Sciences Research Council (EPSRC), grant agreement no. EP/N023013/1, and Google UK Limited (LF).

TABLE OF CONTENTS

I. INTRODUCTION

Adoption of AI in the FS sector is still in its infancy, according to a recent survey of more than 3,000 C-suite executives conducted by Deloitte and the European Financial Management Association (EFMA). The survey results show that 11% had not started any activities in AI, and 40% were still learning how AI could be deployed in their organizations.[1]

For the purpose of this discussion, we use the term AI generally to refer to the collection of techniques that leverage machine learning to perform tasks that normally require human intelligence, including natural language processing, speech recognition, and decision-making under uncertainty. Traditional approaches to tasks were either a people-based process or a systemic

1. Louise Brett et al., *AI and You: Perceptions of Artificial Intelligence from the EMEA financial services industry*, DELOITTE 9 (Apr. 2017), https://www2.deloitte.com/content/dam/Deloitte/cn/Documents/technology/deloitte-cn-tech-ai-and-you-en-170801.pdf [https://perma.cc/R688-FSQS].

rules-based process. Loans were either granted at the discretion of the bank manager or by using a scorecard to calculate a customer's risk level. The unprecedented availability of affordable computer power and the rise in volume and variety of data gave rise to new and advanced algorithms to analyze more information faster. These AI tools can be static and periodically updated, e.g., a revenue forecasting model that is updated per fiscal quarter, or live and continuously evolving with a real-time feedback cycle, e.g., a chatbot that learns in real-time from the user's input.

Despite the slow adoption rate, FS firms are exploring how to leverage AI to drive cost efficiencies and maintain competitiveness. Most banking executives (65%) see the highest potential impact of AI in customer service, while most insurance executives (78%) view back office and operations as the best part of the value chain for AI use.[2]

FS is a highly regulated industry, comprising a wide variety of complex business lines and products. Given the history of regulatory penalties levied for non-compliance or misconduct in the FS industry and the growing regulatory scrutiny around the use of AI, the conservatism in its adoption is understandable.

In the past year, the U.K. Financial Conduct Authority (FCA) and Information Commissioner's Office (ICO) have been actively issuing opinions on AI and machine learning. While regulators are not proactively designing regulation for AI, they are formulating their expectations with their recent publications on algorithmic trading,[3] supervision of internal models,[4] and Senior Managers and Certification Schemes (SM&CS).[5] Conscious of the lag between the pace at which new technologies evolve and the speed at which new regulations can be developed, regulators have historically adopted the principle of "technological neutrality." Therefore, the same regulatory principles in the aforementioned publications apply to firms regardless of the technology they use to perform a regulated activity. They can also be seen as indicators as to how AI may be regulated in the future.

---

2. *Id.* at 7, 12.

3. *See Algorithmic Trading Compliance in Wholesale Markets*, FIN. CONDUCT AUTHORITY (Feb. 2018), https://www.fca.org.uk/publication/multi-firm-reviews/algorithmic-trading-compliance-wholesale-markets.pdf [https://perma.cc/WWS2-UERJ] [hereinafter *Algorithmic Trading Compliance*].

4. *See ECB guide to internal models*, EUROPEAN CENT. BANK (Mar. 2018), https://www.bankingsupervision.europa.eu/legalframework/publiccons/pdf/internal_models/ssm.guidegeneraltopics.en.pdf [https://perma.cc/HV3T-HC6K].

5. *See Senior Managers Regime*, FIN. CONDUCT AUTHORITY 3 (Mar. 2019), https://www.fca.org.uk/publication/corporate/applying-smr-to-fca.pdf [https://perma.cc/E95F-FPVE].

Past literature on the use of AI has focused on the techniques, tools, and methodologies to ensure the fairness, accountability, and transparency of AI algorithms. However, there has been little effort to contextualize these findings within regulatory limitations, and the connection between the technical frameworks and the governance process of an organization has largely been overlooked. Despite the numerous competing mathematical formalizations of fairness, the practical implications for industry on how to implement fair algorithms are uncertain.

This paper will use the risk of discrimination as an example to discuss the practical FS challenges of managing risks introduced by AI. We will walk through an AI product lifecycle and reveal the process by which risks can be identified, assessed, controlled, and monitored in an FS company by deriving recommended practices and principles from past publications by regulators. While it may refer to external regulations, most examples will be drawn from the European Union and the United Kingdom.

II. FAIRNESS IN THE FINANCIAL SERVICES INDUSTRY

Machine learning is increasingly being used to make or aid decisions that are consequential to FS customers, from evaluating their credit worthiness to recommending investment products to pricing their insurance premiums. It also impacts employees, with CV screening algorithms and performance tracking measures.

Historically, FS companies have focused on limited types of data that directly relate to the desired outcome. For auto insurance, such metrics included past driving convictions and number of years of driving experience.[6] For credit risk, they included debt-to-income ratio and past payment histories.[7] With the advent of big data analytics, firms are beginning to incorporate non-traditional types of data into their algorithms as proxies of risk. Controversially, insurance pricing has been found to be influenced by an applicant's email domain name

6. *See, e.g.*, *How Is My Insurance Premium Calculated*, THINK INSURANCE, https://www.thinkinsurance.co.uk/personal/young-driver-insurance/how-is-my-insurance-premium-calculated [https://perma.cc/HRT4-KJAM] (last visited Oct. 12, 2019).

7. *See What risks do banks take*, BANK ENG., https://www.bankofengland.co.uk/knowledgebank/what-risks-do-banks-take [https://perma.cc/QHE4-QJ5A] (last visited Oct. 12, 2019).

and surname,[8] and credit lending decisions can depend on an individual's Internet browsing history.[9]

Prior to the availability of big data and machine learning algorithms, companies could avoid liability by showing that any unequal treatment of protected class was unintentional because the protected attributes were not considered in the decision-making process. AI-driven processes are less transparent than traditional systemic rules-based processes due to their ability to extract patterns from complex feature relationships. Recent legal rulings, however, have transferred the emphasis from discriminatory intent to discriminatory impact. The U.S. Supreme Court upheld "disparate impact" claims under the Fair Housing Act in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*[10] The case found unintentional discrimination to be illegal if the plaintiff can show a disproportionate impact on a protected group.[11] In the United Kingdom, *Essop v. Home Office* similarly found indirect discrimination to be unlawful in hiring practices.[12]

As discrimination gets embedded in such complex relationships in social data within "black box" algorithms, and as governments increasingly focus on impact rather than intent of discrimination, new approaches to identifying the harm in these automated decision tools are required. Given a bias, people-based processes may arrive at different decisions. AI, by contrast, can replicate an identical bias at-scale, crystalizing the bias and removing the outcome ambiguity associated with human decision-making. This is especially concerning in domain areas with documented historical discrimination, as AI can exacerbate any underlying societal problems and inequalities. Even if AI is designed to augment human decision-making rather than completely replace it, the business users may not comprehend the confidence intervals provided and may

8. John Leonard, *Admiral Insurance found to give higher quotes to Hotmail users and people called Mohammed*, COMPUTING (Jan. 24, 2018), https://www.computing.co.uk/ctg/news/3025139/admiral-insurance-found-to-give-higher-quotes-to-hotmail-users-and-people-called-mohammed [https://perma.cc/7793-U9SX].

9. James Rufus Koren, *What does that Web search say about your credit?*, L.A. TIMES (July 17, 2016), https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html [https://perma.cc/T2M3-WZ5M].

10. Deborah B. Baum et al., *Supreme Court Affirms FHA Disparate Impact Claims*, PILLSBURY WINTHROP SHAW PITTMAN LLP (July 21, 2015), https://www.pillsburylaw.com/en/news-and-insights/supreme-court-affirms-fha-disparate-impact-claims.html [https://perma.cc/7J85-7AMP].

11. *Id.*

12. Tom Lowenthal, *Essop v Home Office: Proving Indirect Discrimination*, OXFORD HUM. RTS. HUB (Apr. 6, 2017), http://ohrh.law.ox.ac.uk/essop-v-home-office-proving-indirect-discrimination [https://perma.cc/VN5K-Q6XP].

not feel comfortable overriding the algorithm in practice, given the complexity of how it reached the decision.

On the other hand, the rulings stipulate that if the accused can prove a legitimate business necessity, this treatment can be deemed lawful; however, the required evidence for this justification is unclear and has not yet been studied. In the United States, the "business necessity clause" states disparate impact can be justified to meet performance-related constraints, provided the least possible disparate impact is incurred given the constraints.[13] In the United Kingdom, following the Supreme Court ruling of *Essop v Home Office*, a provision, criterion, or practice (PCP) can be justified by showing it is a "proportionate means of achieving a legitimate aim."[14]

In July 2018, the FCA wrote that while traditionally they have focused on procedural fairness in assessing firms' conduct, there are cases for intervention to ensure distributive fairness in pricing discrimination.[15] The FCA lists six evidential questions to assess whether an intervention is required:

- customer vulnerability;
- scale of adverse effect;
- number of people affected;
- lack of transparency in pricing methodologies;
- essential nature of product or service; and
- societal views of unfairness.[16]

This suggests a step further in the regulators' focus on impact over intent, and organizations will need to shift to an outcome-based analysis of whether their processes are fair.

This paper will use the risk of unlawful discrimination as an example in exploring how an FS company would manage this risk throughout an AI solution's product lifecycle.

III. MANAGING RISKS OF AI THROUGH ITS LIFECYCLE

Academic research has focused on model and algorithmic risks, such as bias and accuracy, in isolation. In reality, model design and performance must

13. Baum, *supra* note 10.
14. Lowenthal, *supra* note 12.
15. Mary Starks et al., *Price discrimination in financial services*, FIN. CONDUCT AUTHORITY 1 (July 2018), https://www.fca.org.uk/publication/research/price_discrimination_in_financial_services.pdf [https://perma.cc/3WK8-LT34].
16. *Id.* at 6.

also consider non-model risk domains, such as regulatory and compliance risk, technology risk, people risk, supplier risk, conduct risk, and market risk.

For example, assessing a model for fairness is not a purely mathematical or computational problem. The appropriate definition, assessment, and remediation has to consider the regulations guiding the use case, the potential technological limitations in its implementation, and the alignment with the company's risk appetite, ethics, and core values.

The adoption of AI does not require an overhaul of the existing enterprise Risk Management Framework (RMF), but rather an awareness of how AI may complicate the detection of risks as they manifest themselves in unfamiliar ways. The volume and speed of data processed may require a much faster reaction speed for any errors, and the complexity of a machine learning algorithm may hinder its explainability and auditability.

Supervisors will expect firms to have robust and effective governance in place, including RMF, to identify, reduce, and control any of the risks associated with the development and ongoing use of each AI application across the business. The RMF should be approved by the board.[17]

A. DESIGN

1. Definition of Scope

A recent FCA report[18] outlines the requirement for firms to define algorithmic trading, with the objective to ensure that firms establish an appropriate process to identify and manage its usage. The FCA can require firms to provide a description of their algorithmic trading strategies within fourteen days.[19] Similarly, FS firms will need to define the scope for what constitutes an AI technology or solution.[20] The difference between AI and rules-based systems, robotic process automation, and static mathematical models should be clear to both management and employees.[21]

The scope should reflect the regulatory implications around the firm's use of AI.[22] Increasingly, AI solutions in industry leverage third-party machine learning algorithms as accelerators for development. While the build process may have been outsourced, the FS firm is still liable for all associated risks.[23] A

17. Tom Bigham et al., *AI and risk management*, DELOITTE 18 (2018), https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/deloitte-gx-ai-and-risk-management.pdf [https://perma.cc/D3BT-3VP5].

18. *See* Algorithmic Trading Compliance, *supra* note 3.

19. *Id.* at 8.

20. *Id.* at 8–9.

21. *See* Bigham et al., *supra* note 17.

22. *Id.*

23. *See* Algorithmic Trading Compliance, *supra* note 3, at 5, 16, 26.

retail banking chatbot powered by a Natural Language Processing application programming interface (API) provided by a third-party company should still fall under the scope of AI RMF because accountability for any legal or regulatory breach still lies with the firm. The FCA advises, where there is technical outsourcing, "the firm remains fully responsible for its regulatory obligations."[24]

2. Risk Identification and Assessment

For firms to identify and assess the impact of AI use cases on their risk appetite, they should first develop a set of clear and consistent assessment criteria to apply to all such cases. The firm should identify relevant risk domains for a solution as well as specific product risks and then assess whether the level of residual risk is acceptable given the existing controls. It is critical that the risk assessment and management process do not constrict creativity. The main objective is to ensure that risks are identified early and properly managed to create a safe setting for innovation. Relevant considerations include:

- **External vs. internal:** The intended audience of the AI solution will determine the conduct risk implications as well as the threshold confidence level and performance the solution is required to reach prior to deployment. For example, an insurance pricing model with a customer user interface has a higher risk of unfair outcomes than an income validation model being used by employees.
- **Use of personal information:** Under the General Data Protection Regulation (GDPR) in Europe, Privacy Impact Assessment should be performed if the organization plans to process personal data. An algorithm using personal information for decision-making should be assessed for fairness. In addition, GDPR gives consumers additional rights to understand and take control of how firms are using their personal data. The UK Information Commissioner pointed out that "where a decision has been made by a machine that has a significant impact on an individual, the GDPR requires that they have the right to challenge the decision and a right to have it explained to them."[25] While there have been disagreements among academics on the definition of and the legal basis for this

24. *Id.* at 5.

25. Science and Technology Committee, *Oral evidence: Algorithms in decision-making, HC 351*, HOUSE OF COMMONS (Jan. 23, 2018), http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-%20decisionmaking/oral/77536.html [https://perma.cc/W4SY-WXYQ].

"right to explanation,"[26] firms should nonetheless have a process in place to respond to customers' inquiries in a meaningful, transparent, and understandable manner and be able to demonstrate that an algorithm is compliant with data protection requirements.

- **Data accuracy and quality:** All input and training data into the machine learning model should be high quality and fit for its intended purpose. This includes a review of the data collection methodology for potential selection bias and an evaluation of the distribution of outcomes for possible biases against protected classes.

Societal views of unfairness, aside from being an FCA criterion for intervention, can lead to reputational damage. When an investigation revealed that motorists named Mohammed are being charged up to £919 more in car insurance than men with typically white, English names, it led to public outcry and calls for a boycott.[27] To avoid such scrutiny, features that are input into the model should be assessed for appropriateness.

Figure 1 visualizes a possible decision boundary for whether or not an input variable should be used in a model. Given the decisions in *Essop v Home Office* and *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, even if a feature is correlated to a protected characteristic, there may be reasonable grounds to use it for business objectives. For example, among loan applicants, income levels may differ between men and women because more women work part-time. Income may still be used in a lending decision due to its high relevance to the risk of default. In contrast, an email domain name may be predictive of risk, but if it is highly correlated to race, it may need to be removed from the model due to the lack of foundation of a causal link to risk.

26. Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 76–99 (2017).

27. Lester Holloway, *Boycott car insurance firms that discriminate*, OPERATION BLACK VOTE (Jan. 25, 2018), https://www.obv.org.uk/news-blogs/boycott-car-insurance-firms-discriminate [https://perma.cc/9QWW-G7FN].

Figure 1: Decision boundary for acceptable use of input variables

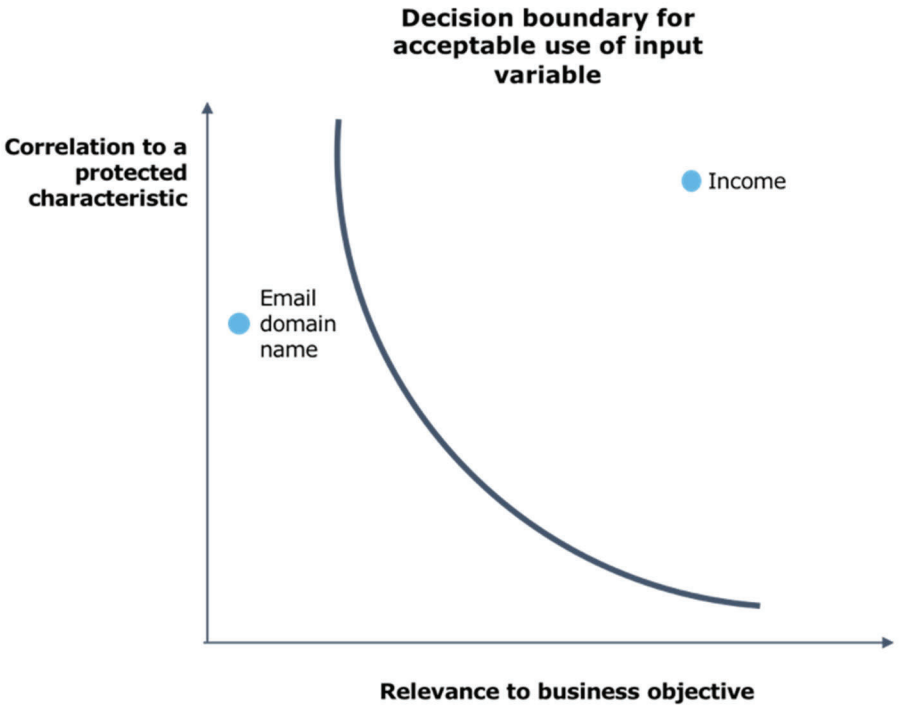
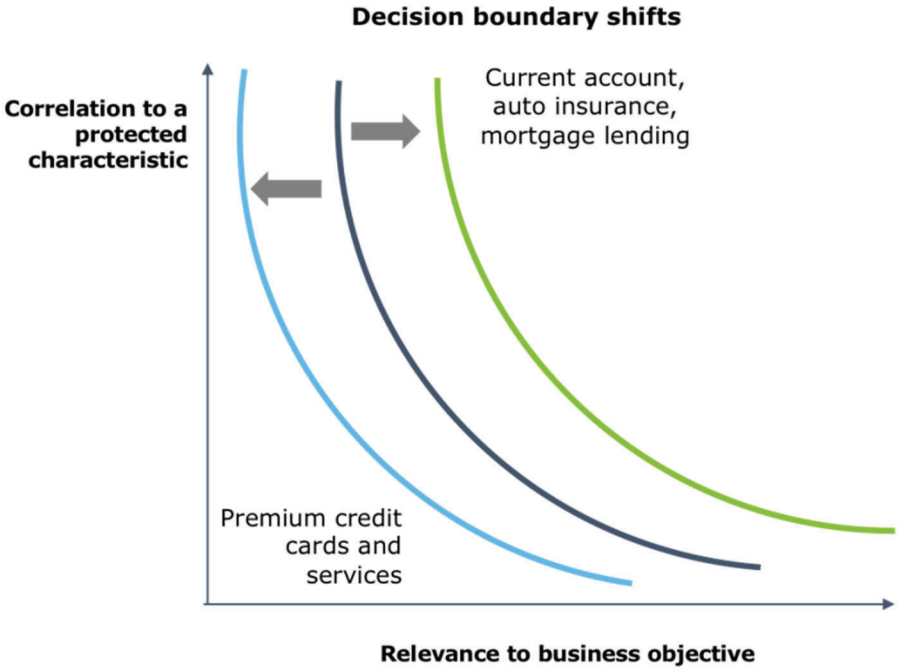


Figure 2: Decision boundary shifts



This decision boundary may shift depending on the conditions outlined by the FCA for possible intervention. The drivers of decision-making in providing essential products, such as checking account, car insurance, or mortgage, may be subject to higher scrutiny than the rationale for offering premium credit cards. This is also related to the greater number of people and a higher proportion of vulnerable customers in essential financial products.

This pre-processing step ensures that the decision to include features correlated to protected characteristics is carefully considered within the context of the regulated domain and the potential impact on consumers.

3. Risk Management Plan and Control Design

Risk management plans should mitigate the risks identified in the assessment, and the residual risk should be in line with the given overall risk appetite. This includes the appropriate controls and testing methodologies, which may vary depending on the FS domain.

Considering the example of the risk of unlawful discrimination, the appropriate control would be to test the algorithm for fairness. Yet, the numerous competing mathematical definitions of fairness only obfuscate its criteria, hindering the ability of business leaders to enforce its implementation. In order to formulate a risk management plan, an appropriate and actionable definition of fairness should be assigned for each use case.

Fairness through Unawareness: This model attempts to avoid discrimination by excluding protected attributes from the model build. Given the power of machine learning algorithms to deduce complex patterns from other features, this does not guarantee a fair outcome. One example of this is the impact of the controversial EU ruling to prohibit car insurance companies from discriminating based on gender in order to counteract the fact that men paid more for insurance than women. Rather than the gap between men and women's insurance premiums narrowing, it has widened from £27 to £101, as insurance companies have turned to gender-correlated proxies for risk measurement, such as occupation and average length of driving history.[28]

While this may be considered more fair if we believe the new prices are reflective of true risk differences between men and women, it is less equitable and would not meet some of the constraints of other definitions of fairness. The model may still be discriminating based on gender through its proxies. In a 2018 study of 1 million insurance quotes in the United Kingdom, the median price was the highest for laborers (e.g., construction workers) and barbers—

28. Patrick Collinson, *How an EU gender equality ruling widened inequality*, GUARDIAN (Jan. 14, 2017), https://www.theguardian.com/money/blog/2017/jan/14/eu-gender-ruling-car-insurance-inequality-worse [https://perma.cc/6BV8-334Z].

stereotypically male jobs—and the lowest for personal assistants and secretaries—stereotypically female jobs.[29]

Defining A as the protected attribute, Y as the actual outcome, and $\hat{Y}$ as the predicted outcome, other fairness metrics in existing statistics literature include:

Demographic parity:[30] Demographic parity (group fairness) is a population-level metric where the outcome is independent of the protected attribute. Formally:

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$$

As Gajane and Pechenizkiy argue, this metric is feasible where there is no reliable "ground truth" data, such as in credit risk and employment where historical discrimination against protected groups is well-documented.[31] They are, on the other hand, ineffective where disproportionality in outcomes can be justified by non-protected, non-proxy attributes, as this can lead to reverse discrimination and inaccurate predictions.[32] It is also not stipulated to select the most optimal outcome.[33] In these cases, the tradeoff between accuracy and demographic parity may be too significant for application in business-critical usage, such as pricing. In employment, where there is an additional interest in increasing the diversity of the workforce, demographic parity may be a useful metric to ensure an equitable representation of all protected classes.

Counterfactual fairness:[34] This model posits that given a causal model (U, V, F) with a set of observable variables (V), a set of latent background variables (U) not caused by V, and a set of functions (F), the counterfactual of belonging to a protected class is independent of the outcome. Where X represents the remaining attributes, A represents the binary protected attribute, and Y is the actual outcome, and $\hat{Y}$ is the predicted outcome, formally:

$$P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a}{}'(U) = y|X = x, A = a)$$

29. Rebecca Rutt, *How much does your job cost in car insurance*, THIS IS MONEY (Apr. 26, 2018), http://www.thisismoney.co.uk/money/bills/article-5637979/The-jobs-expensive-car-insurance.html [https://perma.cc/YE2F-PJV3].

30. Nina Grgic-Hlaca et al., *The case for process fairness in learning: Feature selection for fair decision making*, NIPS SYMP. ON MACHINE LEARNING & L. (2016) [https://perma.cc/D5XT-FZJV].

31. Pratik Gajane & Mykola Pechenizkiy, *On formalizing fairness in prediction with machine learning*, ARXIV (May 28, 2018) https://arxiv.org/pdf/1710.03184.pdf [https://perma.cc/7TPK-HKFM].

32. *Id.*

33. *Id.*

34. Matt Kusner et al., *Counterfactual Fairness*, ARXIV (Mar. 8, 2018) https://arxiv.org/pdf/1703.06856.pdf [https://perma.cc/82PV-BF6Z].

While the methodology of causal inference is robust, causal links are often difficult to hypothesize in complex FS domains. An FCA report acknowledges the challenge of disentangling the differences in actuarial risk (cost-based pricing) from different willingness to pay (price discrimination).[35] The metric is additionally prone to hindsight bias and outcome bias.[36]

Individual fairness:[37] Individual fairness states that similar individuals get similar outputs. Formally, for similar individuals i and j:

$$\hat{Y}\left(X^{(i)}, A^{(i)}\right) \approx \hat{Y}\left(X^{(j)}, A^{(j)}\right)$$

This criterion has a high dependency on the measurement of "similarity" between individuals that does not correlate to the protected characteristics. It is also more computationally intensive than population-level metrics, which could be a challenge for any real-time solutions with Big Data.

Equalized odds / equalized opportunity:[38] Equalized odds imply that predicted outcome given actual outcome is independent on predicted protected attribute given actual outcome. This guarantees that the predicted outcome has equal true positive rates across protected characteristics. Equalized opportunity focuses on the true positives: given a positive outcome, the prediction is independent of the protected attribute. Defining A as the protected attribute, Y as the actual outcome, and $\hat{Y}$ as the predicted outcome, formalization of equalized odds is:

$$P\left(\hat{Y} = 1 \,\middle|\, A = 0, Y = y\right) = P\left(\hat{Y} = 1 \,\middle|\, A = 1, Y = y\right), y \in \{0,1\}$$

Similarly, equalized opportunity meets the following condition:

$$P\left(\hat{Y} = 1 \,\middle|\, A = 0, Y = 1\right) = P\left(\hat{Y} = 1 \,\middle|\, A = 1, Y = 1\right)$$

The relative importance of the accuracy metrics can differ across FS use cases. For example, a mortgage lending company may be most concerned about the algorithm's false positive rates (i.e., approving loans that lead to default). A retail bank with an algorithm to predict expected churn may focus on the false negative rates (i.e., was offered a better rate, but left anyway). The equalized odds and equalized opportunity metrics fail to address discrimination that may already be embedded in the data.[39]

35. Starks et al., *supra* note 15.
36. Gajane & Pechenizkiy, *supra* note 31.
37. Grgic-Hlaca et al., *supra* note 30.
38. Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, ARXIV (Oct. 7, 2016), https://arxiv.org/pdf/1610.02413.pdf [https://perma.cc/3C7W-YESH].
39. Gajane & Pechenizkiy, *supra* note 31.

For any AI with the risk of discrimination against protected classes, appropriate definition of fairness and justification should be required, taking into consideration the strengths and weaknesses of each option and the regulatory implications of its implementation in the FS domain. Once selected, the metric can be used to test the predictions for fairness as a part of the control process. The firm should also bring in stakeholders from legal risk and ethics teams to ensure the definitions are aligned to the companies' ethical values and risk appetite.

4. Defined Roles and Responsibilities

Given the potentially far-reaching implications of AI use on a business, FS firms may need to involve a wider set of stakeholders from first, second, and third lines of defense throughout the product lifecycle. Under SM&CS, senior management should be prepared to evidence an effective governance and risk framework for AI solutions. Good practice involves senior management's participation throughout the testing and development process and understanding of potential market conduct consequences.

The risk and compliance functions should be involved at each stage of the development, testing, and implementation process. In the publication on algorithmic trading, the FCA particularly noted that compliance staff should aim to have the required knowledge and skills to provide sufficient challenge to the development of algorithms, which may initially involve conducting a gap analysis of their ability to supervise algorithmic trading activity and establishing new roles and responsibilities where required.[40]

There should be close collaboration with the technical owner of the AI model and the business owner of the model outcome, with a gradual hand-off of accountability through the lifecycle from design to productionization.

B. BUILD

The AI development process can pose a challenge to traditional risk managers due to the agile approach often adopted by data science and AI teams. It is important to bridge the gap between traditional risk functions and technical teams, as the technical team is not always aware of business and risk challenges. For example, a team may use an open source tool without reviewing whether the license allows for its commercial use. Thus, controls for risks should be embedded into the development process. For example, use of an open source tool should trigger a required process to obtain approval from the legal team to proceed after reviewing the terms of the open source license. Below are analogous regulatory principles for other technologies that apply equally to AI.

40. *Algorithmic Trading Compliance*, *supra* note 3.

1. Development and Testing Process

By maintaining a robust, consistent, and well-understood development and testing framework, firms need to ensure that their development of algorithms is consistent with the risk appetite and behavioral expectations of the firm. The requirements are similar to those proposed by the FCA for algorithmic trading. Before sign-off, firms need to complete a comprehensive review and approval process, and all stakeholders need to confirm that their assigned tasks are completed, verified, and documented.

2. Governance and Oversight

Firms should aim to have an independent multi-disciplinary governance committee to review the documentation and completion of testing procedures and to verify that the algorithm is consistent with the original specifications. Its members should be trained to understand the risks associated with AI applications. The issue of fairness, for example, requires both domain knowledge and understanding of the mathematical trade-offs. These committees should establish the testing and assurance process and regularly review the performance to identify emerging issues.

3. Documented Change Management, Testing, and Approval

Throughout the development and testing process, firms should ensure they have adequate documentation and a comprehensive audit trail for all AI applications deployed throughout their organization, including the relevant owners and key compliance and risk controls in place. Should there be a change in the definition of fairness, for example, this falls under the category of material change and approvals by relevant stakeholders should be recorded.

4. Transparency and Explainability

An analysis of model drivers should reveal any features that should not be impacting the model. If a person's preferred email address provider is highlighted as a potential driver for insurance pricing, this should feed into the algorithm's fairness analysis to ensure this feature is not being used as a proxy for a protected characteristic. Methodological transparency was explicitly listed by the FCA[41] and the GDPR as a requirement for algorithmic decision-making. As the UK Information Commissioner stated earlier this year, "[w]e may need, as a regulator, to look under the hood or behind the curtain to see what data were used, what training data were used, what factors were programmed into the system, and what question the AI system was trained to answer."[42] GDPR

41. Starks et al., *supra* note 15.
42. Bigham et al., *supra* note 17.

will require a shift in relationships with regulators, requiring appropriately funded regulatory affairs teams to discuss any planned high-risk automated data processing.

C. PRODUCTIONIZE

Unlike robotic process automation and other rules-based and deterministic systems, risks in AI-driven solutions are dynamic and more challenging to detect. This requires a shift in mindset for risk managers, who will need to remain involved in the risk monitoring process. Prior to productionization, the solution should be safe to deploy at-scale by embedding automated controls.

1. Ensuring Solution Is Safe to Scale

High data processing volume and speed may require a much faster reaction speed for any errors because risk events can propagate much faster. There should be sufficient controls in place prior to go-live, with rules and thresholds programmed for when human intervention is required. The FCA, in its publication on algorithmic trading compliance, mandated a clear explanation of the conditions that need to be met before being implemented into a live environment.[43]

2. Review the Feedback Mechanism

For machine learning algorithms with live incoming data, there should be a control to flag unsuitable input. A chatbot, for example, should not learn from inflammatory or profane user comments. A pricing algorithm should not react erratically to external shocks.

The appropriateness of the feedback loop should also be considered. In a credit risk algorithm, a bank is likely to lack data on the individuals who were denied a loan, even if they proceeded to get a loan elsewhere. The counterfactual of the decision, i.e., whether they would have paid back the loan had they been approved, is unknown. This missingness should be considered when evaluating the accuracy of the model. In the decision boundaries of the model, continuous experimentation to grant credit to those who were just outside the cut-off point can provide the business with evidence on whether the policy is appropriate.

3. "Kill Switch" and Business Continuity

Firms should document procedures and controls for a manual "kill switch" to stop an algorithm from operating once a critical error or abnormal behavior is detected. Business continuity plans may need to be redefined to provide a

43. *Algorithmic Trading Compliance*, *supra* note 3.

contingency plan for roll-back to manual processes with minimal disruption to critical business processes.

D. MONITOR

Due to the continuously-evolving nature of AI, a more dynamic monitoring approach will be required to ensure a model is still performing as intended for its specific use case. The Key Performance Indicators (KPIs), including non-functional requirements such as fairness, need to be continuously monitored for appropriateness, relevance, and accuracy. In addition, real-time measures of risk (KRIs) can help inform the second and third lines of defense. An example of this would be the number of complaints and appeals against an AI credit decision on the basis of perceived unfairness.

1. Automated Monitoring and Testing

AI-driven solutions can be leveraged for AI risk monitoring. For example, a machine learning-driven solution can monitor phone conversations between an insurance agent and a customer to predict the probability of mis-selling. In this tool called TrueVoice, subject matter experts in both insurance and conduct risk have developed and trained custom metrics such as customer vulnerability, dominance, and loss aversion, all of which indicate a higher likelihood of mis-selling.[44]

2. Vulnerable Customers

Another potential post-processing step may be needed to ensure fairness. If the model results in high variability in outcomes between protected classes, especially if vulnerable customers are involved, an organization may implement a rules-based approach to limit the variation. If a customer is rated as high risk due to the unusual circumstances surrounding his or her vulnerability, some flexibility is required. The FCA defines a vulnerable customer as "someone who, due to their personal circumstances, is especially susceptible to detriment, particularly when a firm is not acting with appropriate levels of care."[45] In particular, the FCA lists "lack of suitable affordable products for people in some non-standard situations" as a potential conduct risk and recommends that "[f]lexibility in the application of terms and conditions of products and

44. *TrueVoice*, DELOITTE UK, (2019), https://www2.deloitte.com/uk/en/pages/risk/solutions/truevoice.html [https://perma.cc/QEJ7-EKTV] (last visited Sept 18, 2019).

45. *Consumer Vulnerability*, FIN. CONDUCT AUTHORITY (Feb. 2015), https://www.fca.org.uk/publication/occasional-papers/occasional-paper-8-exec-summary.pdf [https://perma.cc/GK77-WAVK].

services play[] a significant role [to] ensur[e] the needs of consumers in vulnerable circumstances are met."[46] An FS organization may put guardrails in place to limit the level of variability if a customer is deemed to be vulnerable.

3. Periodic Re-validation

External and internal events can result in a change to the organization's risk profile. New legal and regulatory developments may require a change in the design of the model. Media scrutiny of a use case may make a solution nonviable. Legal teams should communicate any changes and their implications to business owners.

4. Internal Audit Planning

Internal Audit (IA) functions should receive training to acquire adequate expertise to properly understand the risks associated with each AI solution. AI components should be explicitly considered in their audit planning process, independent of the larger systems in which they sit. They should understand and handle compliance breaches and determine the frequency of the review required for each AI solution.

IV. CONCLUSION

While adoption rates have been slow, AI will increasingly become an integral component of FS firms' strategies to achieve operational efficiency, improve customer service, and gain insights for competitive advantage. It is imperative that organizations understand the implications of this adoption from a risk perspective, such that appropriate governance and controls are put in place to mitigate the new and exacerbated risks.

This paper explored the practical implications of risk management throughout an AI solution's product lifecycle. With a particular focus on the United Kingdom and the European Union, suggested approaches were coupled with regulatory principles and precedents. The primary highlighted example use case was the risk of discrimination against protected classes. While there has been a wide array of studies on the technical and theoretical definitions of fairness, further work is required to devise a framework to determine which definitions are most appropriate in the practical implementation of fairness metrics in FS industry.

Risks of AI are not confined to the algorithm itself, but rather affect the entire organization. AI-specific considerations should be integrated into existing RMFs to ensure they remain fit for purpose. Only then will FS firms feel

46. *Id.*

empowered to use AI, having the confidence that AI-related risks can be effectively identified and managed.