Explainable AI For Financial Risk Management

FRIL
Financial Regulation Innovation Lab

FinTech Scotland®

University of Glasgow

University of Strathclyde Glasgow

Financial Regulation Innovation Lab

Shaping the future of financial regulation

Who are we?

The Financial Regulation Innovation Lab (FRIL) is an industry-led collaborative research and innovation programme focused on leveraging new technologies to respond to, shape, and help evolve the future regulatory landscape in the UK and globally, helping to create new employment and business opportunities, and enabling the future talent.

FRIL provides an environment for participants to engage and collaborate on the dynamic demands of financial regulation, explore, test and experiment with new technologies, build confidence in solutions and demonstrate their ability to meet regulatory standards worldwide.

What is Actionable Research?

FRIL will integrate academic research with an industry relevant agenda, focused on enabling knowledge on cutting-edge topics such as generative and explainable AI, advanced analytics, advanced computing, and earth-intelligent data as applied to financial regulation. The approach fosters cross sector learning to produce a series of papers, actionable recommendations and strategic plans that can be tested in the innovation environment, in collaboration across industry and regulators.

FRIL White Paper Series

Explainable AI for Financial Risk Management

James Bowden*   Mark Cummins*   Daniel Dao*   Kushagra Jain*

* University of Strathclyde

March 2024

Abstract: We overview the opportunities that Explainable AI (XAI) offer to enhance financial risk management practice, which feeds into the objective of simplifying compliance for banking and financial services organisations. We provide a clear problem statement, which makes the case for explainability around AI systems from the business and the regulatory perspective. A comprehensive literature review positions the study and informs the solution framework proposed. The solution framework sets out the key considerations of an organisation in terms of setting strategic priorities around the explainability of AI systems, the institution of appropriate model governance structures, the technical considerations in XAI analytics, and the imperative to evaluate explanations. The use case demonstration brings the XAI discussion to life through an application to AI based credit risk management, with focus on credit default prediction.

Table of Contents

1. Problem Statement

Recent years have seen a seismic shift in the development and deployment of Artificial Intelligence (AI)[1] systems within financial services to support a range of functions and activities. AI systems, with the support of cloud computing and high-performance computing infrastructure, bring value through the ability to process vast amounts of data at unprecedented speeds to deliver actionable insights for practitioners. One of the key functions within which AI is applied in decision support tools is financial risk management. Various approaches have been proposed to support the management of core risk pillars, including credit risk, market risk, liquidity risk and operational risk. AI systems allow for improved financial risk management procedures informed by a wider pool of structured and unstructured data sources, offering greater accuracy in forecasting risk exposures and facilitating higher frequency risk monitoring and management practices.

The use of AI systems, however, brings its own unique set of risks to a financial services organisation. Notable amongst these risks is the lack of transparency around how AI systems operate and, in particular, the lack of explainability around AI system outputs. Indeed, in the pursuit of AI system performance accuracy, developments have advanced in the direction of deep learning, which brings us further away from explainability. Figure 1 from Yang et al. (2022) visualises the inverse relation that exists between AI model performance and explainability. Here, one can see that linear models and rules-based models provide the highest levels of explainability but the lowest levels of performance, while, in contrast, deep learning models provide the highest levels of performance but the lowest levels of explainability. In deploying AI systems, there is therefore an inevitable trade-off to be made between performance and explainability,

which depends considerably on the use case in question and the associated materiality for the organisation.
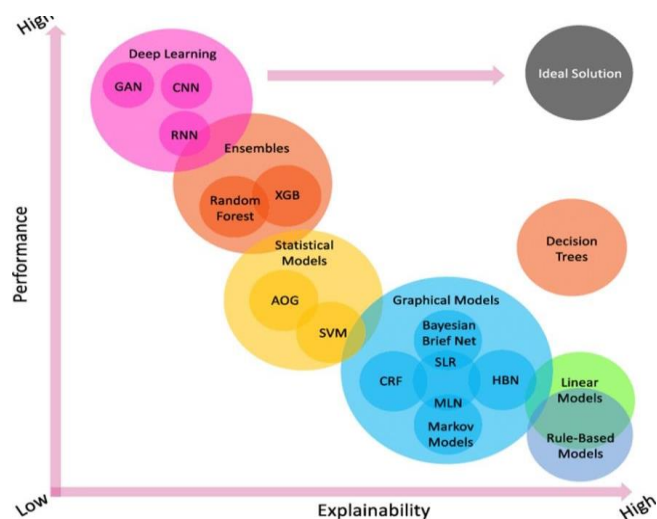
Source: Yang et al. (2022)

Figure 1: The Performance-Explainability Relation

The inability to explain AI system outputs creates a significant barrier to wider AI systems adoption within financial services. In particular, the lack of explainability leads to an understandable distrust in AI systems, fuels the challenge of articulating and communicating the value proposition of AI systems internally within a financial services organisation, creates difficulties in adhering to external regulatory and supervisory compliance and oversight, and threatens good consumer outcomes in respect of the right to explainability. Against this backdrop, the OECD in its assessment of AI opportunities and challenges in finance, explicitly calls out explainability as a significant challenge (OECD 2021). The OECD notes the lack of explainability in AI systems impedes micro-level prudential supervision, which in turn creates macro-level risk for the financial system. So, the challenges pertaining to the

---

[1] For convenience, we use the generic term AI generally throughout this white paper in the knowledge that AI is a wide concept that incorporates machine learning and deep learning, which the latter recognised as a subset of machine learning.

generation of AI system explanations directly impacts on financial stability.

The need for transparency and explainability is reinforced by emerging regulations in the UK and EU. The EU AI Act has advanced considerably with political agreement at the Parliament, Council and Commission levels. The finalised text is due to be completed in 2024 and the full application of the act will be phased in over a 24-month period thereafter. The [EU AI Act](#) sets out a number of priority principles, with transparency being central to these. This transparency principle calls explicitly for explainability around AI systems deployment. In particular, decision outcomes need to be explainable with associated transparency required around training data and accuracy performance. The UK is taking a different approach in that it is not developing a single AI act but instead plans to leverage numerous regulatory frameworks. The [UK's approach to AI regulation](#) is one of pro-innovation. Five principles are set out under this approach, one of which is *Appropriate Transparency and Explainability*. This principle mirrors that of the EU AI Act in its call for explainability around AI systems deployment. In this context, this white paper is particularly pertinent.

The discussion thus far reinforces the imperative for AI system explainability to receive the same level of attention that AI system performance receives. This is especially true when considering financial risk management, given this this function is premised on enabling financial services innovation through controlling risk exposure. In this white paper, we tackle the problem of explainability in AI systems applied for financial risk management. We propose the use of innovative Explainable AI (XAI) techniques that allow financial risk analysts and managers to leverage AI, while providing explanations that can be linked back to existing financial theory and evidence.

2. Literature Review

To overcome crucial weaknesses of black boxes in traditional machine learning model, Explainable Artificial Intelligence (XAI) /

Machine Learning has been developed in decision support system (Guidotti et al., 2018). It is imperative that all individuals understand "meaningful explanations of the logic involved" in decision-making models, following recent General Data Protection Regulation (GDPR) law of European Parliament.

A new field of XAI research has recently emerged. Mueller et al. (2019) classifies XAI into three generations. First-generation focuses on internal working process using expert knowledge and natural language processing. Second-generation emphasizes cognitive assistance, while Thirdgeneration shifts to black-box systems to explain inner workings like First-generation. The development of computers and technology systems currently enables unpacking various explainable choices, thus contributing to transparent decision-making driven by responsible and trustworthy processes.

Explainable Artificial Intelligence (XAI) yields numerous advantages. It aims to reveal data correlations, describe the process of inferring causality, and establish more harmonious human– machine links (Haefner et al., 2021). XAI reduces the likelihood of inaccurate informational and biased decisions, increasing the credibility and consistency of financial processes (Rudin and Radin, 2019). Interpretability's concepts of dependability and trustworthiness can significantly improve the user experience and increase their confidence in operational integrity (Adadi and Berrada, 2018).

However, XAI does not come with no cost. Line of literature (Miller, 2019; Ali et al., 2023) underscores the inherent trade-off between accuracy and explainability. While high-complicated machine learning and AI models highly likely provides better accurate results, they would suffer from low explainability, and vice versa. It becomes essential to improve the explicability of results while simultaneously ensuring a satisfactory level of accuracy.

Given the substantial potentials, costs, and advantages associated with XAI in decision support systems, a considerable body of literature has emerged to advance

understanding in XAI applications within financial services, with a particular focus on financial risk management, where the significance of decision-making is pronounced.

2.1 XAI Applications in Financial Services

Recent applications of artificial intelligence (AI) in the financial sectors aim to bolster decision making for key stakeholders, such as financial institutions, companies, and investors (Goodell et al., 2021; Padmanabhan et al., 2022). Nevertheless, the inherent black-box nature of AI models gives rise to concerns regarding their effectiveness, trustworthiness, and untapped potential. Therefore, XAI has been deliberated in finance literature as a prospective and viable remedy to address these concerns.

One application of XAI within the financial services sector involves the augmentation of the asset pricing process through a deep investigation of machine learning techniques for return prediction (Gu et al., 2020). This entails the utilization of both conventional regularized linear methods, such as regressions, and advanced nonlinear methodologies, including boosted regression trees (such as Extreme Gradient Boosting) and Random Forest Regressions, among others. They show substantial gains by including machine learning for estimating expected returns. Gu et al. (2020) see R^2 improvements, and big gains for strategies harnessing machine learning predictions. Their empirical analysis identifies the most informative predictor variables, allowing further investigation into economic mechanisms, and XAI can be used in an analogous fashion to this in other contexts. Machine learning also makes it possible to improve expected return estimates using predictive information in complex and unstructured data sets (Giglio et al., 2022). There are of course, drawbacks to using such models. For instance, the return prediction literature using them delves little into understanding economic mechanisms (such as risk-return trade-offs, market frictions, or behavioural biases) potentially responsible for observed predictability (Giglio et al., 2022). Distinguishing between risk premia and mispricing in this context requires a more structured modelling approach, and factor models are the dominant tool researchers have used in this pursuit (Giglio et al., 2022).

The evaluation of fund performance stands as a pertinent domain that contemporary literature on XAI in the investment services is currently investigating. Kovvuri et al. (2023) employ the XGBoost model as a machine learning framework for evaluating the performance of global equity fund, while they utilize Shapley values as an XAI method to elaborate on and extend explanations regarding predictors. More recently, the use of XAI (specifically variable importance for neural networks) has been used to assess the skill of mutual fund managers and ascertain which fund characteristics differentiate out-of-sample mutual fund performance, before and after fees, and the significance of their interaction effects using neural networks (Kaniel et al., 2023). In a similar vein, XAI methods (Shapley values for elastic net, random forests and gradient boosting) have been shown to allow one to distinguish between positive and negative alpha mutual funds outof-sample net of transaction costs, based on their characteristics and their interactions (DeMiguel et al., 2023). One recent work of Babaei et al. (2022) investigates how XAI can elevate the practices of portfolio management. Specifically, they adopt XAI technique (Shapley values) to elucidate the rationale behind the chosen portfolio weights.

Numerous research endeavours have delved into the integration of XAI in financing. While XAI methods (global Shapley value and Shapley–Lorenz) has been harnessed to counter racial discrimination in an algorithmic loan decision making setting (Agarwal et al., 2023), Lu and Calabrese (2023) adopt the Cohort Shapley value to assess the fairness in financing small and medium enterprises in the UK. The literature on the application of XAI to highlight discrimination or promote fairness while concurrently optimizing performance is vast (e.g., Martin, 2023; Wan et al., 2023; Chen et al. , 2022; Blattner et al., 2022; Karimi et al., 2022; Kozodoi et al., 2022; Bartlett et al., 2022; Fuster et al., 2022; Castelnovo et al., 2020; Dudik, et al., 2020; Bird, et al., 2020; Bellamy, et al., 2018).

Other fields of financial services can benefit from the application of XAI, including household consumption (Zhou et al., 2023), corporate governance (Scott, 2015), and customer relations (Coussement and De Bock, 2013). While the potential applications of XAI in financial services are broad, our specific emphasis lies in the domain of financial risk management.

2.2 XAI Applications in Financial Risk Management

Risk management (e.g., default and bankruptcy prediction, fraud detection) is concerned with identifying, measuring, and controlling financial risks (Zheng et al., 2019). Financial institutions continuously perform it, and regulators require it (Adams and Hagras, 2020). The application of XAI to open the black box is financial risk management is becoming more common in literature. XAI is being applied to unravel aspects of credit risk management, including default and bankruptcy prediction (Sigrist and Hirnschall, 2019; Zheng et al., 2019; Zhang et al., 2023) and fraud detection (Jarovsky et al., 2018). Additionally, there is a recent focus on employing XAI to unveil insights into Environmental, Social, and Governance (ESG) risk management.

Default and bankruptcy prediction, as discussed by Sigrist and Hirnschall (2019), involves assessing the likelihood of corporate failure. The focus of default prediction lies in estimating the probability of debtors, such as credit card holders and financial institutions, defaulting based on available information, including profiles, loan history, and repayment history, while bankruptcy prediction utilizes publicly accessible information to evaluate the potential for a company to go bankrupt (Sigrist and Hirnschall, 2019; Zheng et al., 2019). Sigrist and Hirnschall (2019) extend the AI model to assess default prediction by two model-agnostic post-hoc XAI tools (variable importance measures and partial dependence plots. In their recent work, by focusing on Chinese listed manufacturing companies spanning the years 2012 to 2021, Zhang et al. (2023) develop a financial risk early warning model using the D-S Evidence theory-XGBoost (DS-XGBoost) framework and conducte an analysis of model interpretability through SHAP (Shapley Additive Explanations). The identification of fraudulent transactions is a crucial aspect of fraud detection, involving the exposure of unauthorized activities on various accounts (Jarovsky et al., 2018). While AI already play a pivotal role in supporting these efforts, XAI serves to enhance decisionmaking process by providing transparent and non-discriminatory justifications, thereby making AI applications more industry-applicable (Park et al., 2021).

XAI can be used in credit risk management and, in particular, in measuring the risks that arise when assessing credit in peer to peer lending platforms with Shapley values (Bussmann et al., 2020); assess the impact of financial and non-financial factors on a firm's ex-ante cost of capital, a measure that reflects the perception of investors on a firm's riskiness with Shapley values and Lorenz Zonoids (Bussmann et al., 2023). Lin and Bai (2022) gather data from 40 listed enterprises in the mining, steel, and power industries, encompassing 224 financial and non-financial indicators, to predict long-term debt. Employing the XGBoost method for feature selection in the context of high dimensionality, the study identifies the top six indicators within subsets that demonstrate significant efficacy in predicting long-term debt of firms. The selected indicators' predictive capabilities were further elucidated through the Shapley additive explanation value. In a related investigation, Tron et al. (2023) scrutinize the capacity of corporate governance features in non-listed companies to discern instances of corporate defaults using XAI techniques.

Explainable AI (XAI) is also adopted to assess the risk associated with Environmental, Social, and Governance (ESG), which currently stands out as a prominent and trending area in sustainable finance. The demonstration of the use of XAI in the case of ESG Regulation Compliance is specifically motivated by past work demonstrating the utility of such methods in such contexts. More specifically, ESG rating transparency has been scrutinized with the aid of explainable artificial intelligence algorithms, lending interpretability (with

Shapley values) to and shedding light on ESG scores derived from proprietary models with satisfactory accuracy levels (Del Vitto et al., 2023). Specifically, their interpretability method allows a fuller understanding of the rating system of an issuing agency, and better integration of information provided by the sustainability performance indicator in decision making. Further, through local interpretability Shapley values application the ratings associated of any company can be explained and motivated. Comparisons can also be facilitated across different ratings providers to reconcile disagreements driven by differences in feature relevance in their methodological assessment. Further, the challenges posed by the rapidly evolving landscape of ESG Regulation are significant. The use of these techniques could assuage concerns businesses have in ensuring compliance, and at the same time assuring regulators of the fulfilment of the objectives sought to be achieved by legislation. These regulations might be especially suited for applying these concepts, particularly in the reporting case, as all forms, at least for EU disclosure, are standardized.

Given the potential transparent and interpretability, XAI facilitates comprehensive analysis of the decision-making processes in ESG, including reducing biases against social or demographic groups in machine learning models (Seele, 2017; Lacoste et al., 2019; Hoepner et al., 2021; FritzMorgenthal et al., 2022; Sætra, 2023). Specifically, Seele (2017) explores the application of predictive policing in corporate sustainability management, elucidating its value to shareholders and financial analysts. Lacoste et al. (2019) use XAI to develop a tool to quantify the carbon emissions for corporate practitioners. Hoepner et al. (2021) underscore the importance of addressing explainability challenges in financial data science research. Fritz-Morgenthal et al. (2022) also propose responsible, trustworthy, explainable, auditable, and manageable AI to

investigate governance concerns. Additionally, Sætra (2023) contribute to this line of literature by formulating an ESG protocol for companies, aiming to enhance corporate governance and stakeholder communication regarding AI capabilities, assets, and activities.

3. Solution Framework

3.1 Corporate Strategy

In considering XAI integration into AI system deployment within a financial services firm, it is necessary to consider the importance of explainability strategically and to connect this explicitly with the firm's overall digital strategy. Grennan et al. (2022), in a McKinsey[2] outline the business case for explainable AI. In particular, the following benefits are identified:
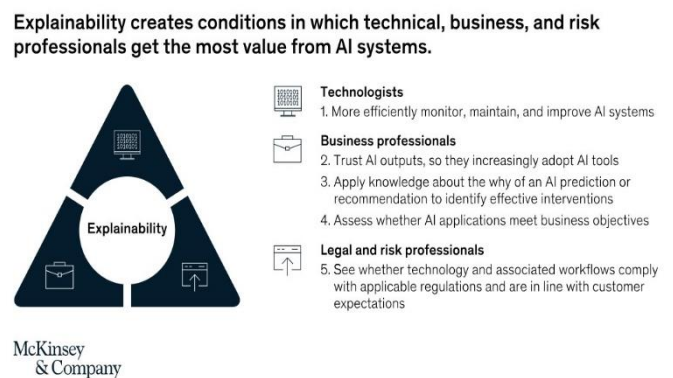
Source: Grennan et al. (2022) [McKinsey]
Figure 2: Impact of Explainability on AI System Users

- Increased productivity through better monitoring, maintenance and enhancement of AI systems;
- Building trust and adoption rates among key stakeholders through the transparency that explanations provide;
- Identifying new value creation opportunities from the insights that explanations provide;

---

[2] "Why businesses need explainable AI – and how to deliver it" by Liz Grennan, Andreas Kremer, Alex Single, and Peter Zipparo. Report available at https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainablehttps://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-itai-and-how-to-deliver-it.

- Articulating the business value of AI systems through explanations that connect investment to outcomes more closely.
- Better risk mitigation and regulatory compliance outcomes afforded by AI system explanations.

Placing strategic importance on the explainability of AI systems has the potential to impact various key users across an organisation. Figure 2 from Grennan et al. (2022) summarises this impact for several professional roles – technologists, business professionals and legal and risk professionals. It can be seen that XAI can benefit users through delivering efficiencies, building trust, facilitating human-in-the-loop interventions, aligning with business objectives and complying with regulations. This latter point is extremely important in the context of, on the one hand, using AI towards simplifying compliance, and, on the other hand, complying with regulation pertaining to AI systems usage within financial services organisations.

Source: Grennan et al. (2022) [McKinsey]

Figure 2: Impact of Explainability on AI System Users

3.2 Model Governance

Once the strategic priority has been approved in respect of explainability of AI systems, a financial services firm then needs to ensure appropriate model governance structures are in place. AI systems represent a new form of model usage for financial services firms. This novel suite of models creates unique model risk exposure for the organisation – given the black box nature of AI systems – that must be controlled through existing, but suitably adapted, model risk management structures.

Deloitte (2022)[3] outline three lines of defence in respect of the governance of XAI models. These lines of defence are summarised in Table 1. The first line of defence relates to model developers within an organisation, who must ensure that explainability is built into AI model deployment. Model developers are required to embed XAI as required to deliver on the explainability standards agreed within the organisation, which applies whether the AI system has been developed in-house or acquired from an external third-party provider. The second line of defence relates to the model validators and model risk managers within an organisation, who have responsibility for validating developed AI models from an explainability perspective (among other considerations) and assigning usage conditions based on explainability levels (among other conditions). The third line of defence relates to the audit and compliance functions within an organisation, who have responsibility for ensuring that the explanations delivered by XAI are fit for purpose, understood by users and can be justified to external auditors.

Line of defense	AI governance and controls	XAI inclusions
First	Model developers working with data governance, tech, and risk teams to curate data, establish standards for the AI application, and determine risk thresholds that should activate a "kill switch"	Identifying XAI risks, assessing opacity/complexity constraints in vendor models, documenting potential drawbacks and steps taken to alleviate them
Second	Model validators and model risk managers monitor AI functions and outputs, conduct ongoing bias testing, and make risk management standards consistent across business units	Quants within the bank check developers' work for XAI shortcomings; verify that resulting levels of interpretability are sufficient for model use
Third	AI ethicists use independent tools and techniques to test models for safety and robustness, and serve as the final point of accountability	Audit and compliance perform final check on XAI, confirming whether explanations serve the intended purpose and can be understood and acted upon by relevant users

Source: Deloitte (2022)

Table 1: XAI Governance and Lines of Defence

3.3 Approaches to Explanation Generation

With strategy and governance structures in place, the organisation needs to then focus on engineering explainability into AI systems through the choice of specific XAI approaches. This choice may depend on the nature of the problem space and the materiality attached to this. We provide an overview of the main considerations in respect of XAI techniques.

---

[3] Report available at https://www.deloitte.com/an/en/our-thinking/insights/industry/financialhttps://www.deloitte.com/an/en/our-thinking/insights/industry/financial-services/explainable-ai-in-banking.htmlservices/explainable-ai-in-banking.html.

What is an explanation, and what are its properties? (Molnar, 2020)

An explanation usually relates feature values of an instance to its model prediction in a humanly understandable way. To explain an ML model's predictions, some explanation method is relied on, such as an algorithm that generates explanations. Other explanation types consist of a set of data instances (e.g., for the k-nearest neighbour model). For example, a support vector machine can be used to predict cancer risk, and explain predictions with the local surrogate method, that generates decision trees as explanations. Alternately, a linear regression model may be used that is already equipped with an explanation method (interpreting weights). Certain properties have been identified for explanation methods, and explanations. These may be used to assess how good they are. It is unclear how these properties may be measured correctly, so formalizing how they could be calculated is a vicissitude. (Molnar, 2020).

Properties of Explanation Methods (Molnar, 2020):

• Expressive Power - "Language" or structure of explanations the method generates. An explanation method may generate natural language, a weighted sum, decision trees, IFTHEN rules, or something else (Molnar, 2020).
• Translucency - Describes the extent of reliance on the explanation method to look into the ML model, like its parameters. E.g., Intrinsically interpretable models like the linear regression model (model-specific) with explanations reliant on them are highly translucent. Conversely, methods solely dependent on manipulating inputs and observing predictions have zero translucency. Different scenario-dependent translucency levels may be desirable. High translucency methods can rely on more information to generate explanations. Meanwhile, low translucency explanation methods are more portable (Molnar, 2020).

• Portability - Describes how many ML models with which the explanation method may be used. Low translucency methods have higher portability: they treat ML models as black boxes. Surrogate models may be the explanation method with highest portability. Model specific methods (only work for that model e.g., recurrent neural networks) have low portability (Molnar, 2020).
• Algorithmic Complexity - Describes computational complexity of the explanation generating method. Important when computation time bottlenecks generating explanations (Molnar, 2020).

Properties of Individual Explanations (Molnar, 2020):

• Accuracy: How well is unseen data predicted? High accuracy is particularly valuable if the explanation is used for predictions in place of the ML model. Low accuracy may be acceptable if the ML model's accuracy is also low, and if the goal is to explain what the black box model does. In this case, only fidelity is important (Molnar, 2020).
• Fidelity: How well is the black-box model's prediction approximated? High fidelity is one of the most important explanation properties, as low fidelity explanations have no value in explaining ML models. Accuracy and fidelity are closely related. If the black box model has high accuracy its explanation also usually has high fidelity and accuracy. Some explanations only offer local fidelity, i.e., explanation only approximates well to model prediction for a data subset (e.g., local surrogate models) or individual data instance (e.g., local Shapley Values) (Molnar, 2020).
• Consistency: Differences between models trained on the same task and producing similar predictions? E.g., assume a support vector machine and a linear regression model are trained on the same task and both produce very similar predictions. Using a method of choice, if the explanations are very similar, they are highly consistent. This is somewhat subtle, as two models may use different features, with similar predictions (also called "Rashomon Effect"). A high consistency is undesirable here as the

explanations must be very different. High consistency is desirable if models really rely on similar relationships (Molnar, 2020).

• Stability: Similarity across similar instances. Stability juxtaposes explanations between similar instances for a model, whereas consistency contrasts explanations between models. High stability means slight variations in an instance's features do not substantially change the explanation (unless these slight variations also strongly change the prediction). Instability may be due to high variance of the explanation method. Put differently, strong effects on explanations are seen from slight changes to feature values of the instance to be explained. Non-deterministic components of the explanation method may also drive instability, like a data sampling step, which the local surrogate method uses. High stability is always desirable (Molnar, 2020).

• Comprehensibility: How well do humans understand? While seemingly like the other properties, this one is particularly important. It is difficult to measure and define, but very crucial to get right. Comprehensibility is broadly agreed to depend on the audience. Measurement ideas include measuring the explanation size (number of features with non-zero weights in a linear model, number of decision rules, etc.) or testing how well people predict ML model behaviour from explanations. Comprehensibility of features used in explanations also should be considered. Complex feature transformations may be less comprehensible than the originals (Molnar, 2020).

• Certainty: Is the certainty of the ML model reflected? Many ML models only predict without stating the confidence of correct predictions. If a 4% cancer probability is predicted for a patient, is it as certain as the 4% probability another patient, with different feature values, received? Explanation incorporating model certainty is very useful (Molnar, 2020).

• Degree of Importance: How well is importance of features or parts of the explanation reflected? If a decision rule explanation for instance is generated for an individual prediction, is it clear which rule

conditions were the most important (Molnar, 2020)?

• Novelty: Is it evident if a data instance to be explained is sampled from a "new" region, far removed from the training data's distribution? If not, the model may be inaccurate and explanation useless. Novelty is conceptually related to certainty. Higher novelty, Implied higher likelihood of low model certainty due to lack of data (Molnar, 2020).

• Representativeness: How many instances are covered? Explanations may cover an entire model (e.g., linear regression model weights interpretation) or represent individual predictions (e.g., local Shapley Values) (Molnar, 2020).

What are good or human-friendly explanations? (Molnar, 2020)

There can be far-reaching consequences for interpretable machine learning based on "good" explanations, as defined by humans. Concise and single (or at most double) cause explanations which juxtapose the treatment group with a counterfactual group are preferred by humans. Good explanations are provided particularly by abnormal causes. Explanations are also "social interactions between the explainer and explanation recipient", where a human being or a machine is the explainer. This implies the actual content of the explanation is significantly impacted by the social context. Alternately, they may refer to "the social and cognitive process of explaining, but also to the product of these processes". Furthermore, a careful distinction needs to be made when comparing explanations that are "human-friendly", and complete causal attribution, where all factors for a particular prediction or behaviour need explaining. The latter may be preferred in legal contexts, where one is mandated to debug an ML model or indicate all influencing sources (Molnar, 2020).

Conversely, where non-experts or time-starved individuals are the explanation's target audience, an alternative definition applies, which defines an explanation as "the answer to

a whyquestion", which can be answered with an "everyday"-explanation. Instances of such questions may include why a loan was rejected, or why a treatment lacked efficacy for a patient. Such "why" questions can usually be reformulated as questions beginning with "how" as well (Molnar, 2020).

A deeper dive into what constitutes a "good" explanation yields certain criteria which have definite implications for interpretable ML. These can be listed as follows - for more detail on these and their implications with examples, interested readers are referred to (Molnar, 2020):

• Contrastive – Answers why this prediction was made *instead of another prediction*. The implication is a requirement for application-dependent explanations because a point of reference for comparison is needed. This may depend on the data point to be explained, but also on the user receiving the explanation. The solution for automated creation of contrastive explanations might also involve finding proto/archetypes in the data (Molnar, 2020).

• Selected - Select one or two causes from various possible causes as "THE" explanation, rather than covering an actual complete list of event causes (Molnar, 2020). This implies a preference for brevity in explanation with 1-3 reasons, even if reality is more complex (Molnar, 2020).

• Social - Part of a conversation or interaction between the explainer and explanation receiver. The implication is that attention to the social environment and intended recipients for explanations is needed. Getting this right may depend entirely on the specific application (Molnar, 2020).

• Focus on the abnormal - People focus more on abnormal causes in any sense (like a rare category of a categorical feature) to explain events, that had a small probability but nevertheless happened, without which the outcome would have greatly changed (counterfactual explanation) (Molnar, 2020). If an input feature for a prediction is abnormal, and it influenced the latter, it

should be included in an explanation, even if other 'normal' features have the same influence (Molnar, 2020).

• Truthful - Prove to be true in reality (i.e., in other situations), but selectiveness seems more important, which is troubling (Molnar, 2020). This implies events should be predicted as truthfully as possible (also called fidelity), with less relative importance given to it than contrast, social aspect, and selectivity (Molnar, 2020).

• Consistent with explainee's prior beliefs - Humans tend to devalue or ignore information inconsistent or in disagreement with prior beliefs, also called confirmation bias. Thus, this bias logically also extends to explanations (Molnar, 2020). This implies using specific ways to deal with inconsistent explanations, although difficult to integrate into ML, and may come at a heavy cost to predictive performance (Molnar, 2020).

• General and probable - A cause that can explain many events is very general and could be considered a good explanation. Although this contradicts the claim that abnormal causes make good explanations, as a rule of thumb, abnormal causes trump general causes, and in the absence of the former, the latter comes to the fore (Molnar, 2020). Implies measurement of generality should happen, which is easily achieved by the feature's support: the number of instances to which the explanation applies, divided by the total number of instances (Molnar, 2020).

Methods for machine learning interpretability can be classified according to various criteria (Molnar, 2020):

Intrinsic or post hoc: Criterion distinguishes based on how interpretability is achieved by restricting the model complexity (intrinsic) or analysing the model by applying methods after training (post hoc). Intrinsic interpretability describes models deemed interpretable owing to their simplicity, e.g., sparse linear models or short decision trees. Post hoc interpretability implies interpretability methods applied after model training, e.g., permutation feature importance. Post hoc methods may also be applied to intrinsically interpretability models, like computing permutation feature importance for decision trees (Molnar, 2020).

Model-specific or model-agnostic: Interpretability tools confined to specific model classes are considered model-specific. Linear regression model weights are interpreted this way, as their intrinsic interpretation is always model-specific. Similarly, tailored tools for interpreting machine learning models such as neural networks are also considered model-specific. In contrast, model-agnostic interpretability tools may be deployed on any model and are used post hoc, after model training. Generally, such agnostic methods function through feature input and output pairs' analysis. These methods cannot access model internals like weights or structural information (Molnar, 2020).

Scope of interpretability: Each algorithmic step in training a predictive model can be evaluated in terms of transparency and interpretability (Molnar, 2020):

- **Algorithm Transparency:** *Assesses how an algorithm creates the model.* This relates to how an algorithm learns a model from data and the relation types it is capable of learning. Using convolutional neural networks to classify images, one may explain the learning of edge detectors and filters on the lowest layers by the algorithm. This is comprehension of how the algorithm works, but not the specific model that learned in the end, and the individual prediction process. Such transparency only requires algorithmic knowledge rather than knowing data or the learned model. Algorithms like the least squares method are well studied and understood. They characterize high transparency. Deep learning approaches (pushing a gradient through a network with millions of weights) are in contrast less well understood. Research is ongoing on their inner workings and are thus opaquer (Molnar, 2020).
- **Global, Holistic Model Interpretability:** *This distinction focuses on how the trained model makes predictions.* A model may be called interpretable if it can be comprehended entirely at once. To explain the global model output, knowing the trained model,

algorithm and data are prerequisites. This interpretability level considers how the model decisions are made, from a holistic features' view, and each learned component e.g., weights, other parameters, and structures. Global interpretability answers the question: which features are important, and what kind of interactions between them take place? In other words, it helps comprehend the target outcome distribution based on features and is exceedingly difficult to achieve pragmatically. Any model beyond a limited number of parameters or weights cannot fit into an average human's short-term memory. One cannot imagine a five-feature linear model as it implies drawing the estimated hyperplane in a five-dimensional space. Any space over three dimensions cannot be conceived by humans. Thus, model comprehension by humans is generally limited to parts, such as linear model weights (Molnar, 2020).
- **Global Model Interpretability on a Modular Level:** *At a modular level global explanations determine how model parts impact predictions.* A Naive Bayes model with several hundreds of features is far too large for a human's working memory. Even with memorization, quick predictions for new data points would be impractical. The joint distribution of all features is needed over and above this to estimate each feature's importance and how they affect predictions on average, making it impossible. But a single weight is easily understood. Thus, understanding some models at a modular level is probable. Not all models can be interpreted at a parameter level. For linear models, the interpretable parts are weights, for trees they are splits (selected features + cut-off points) and leaf node predictions. Linear models may seem perfectly interpretable on a modular level, but a single weight's interpretation is inextricably linked with all other weights. This is why such an interpretation is prefaced by saying other input features remain the same, which is not realistic in most cases. A linear model predicting a house's value, accounts for both its size and number of rooms, and may negative weight the room feature. This is as it

is highly correlated with the house size feature. Where people prefer larger rooms, fewer rooms in a house may be valued over a house with more rooms, if both are of the same size. Weights only make sense after contextualizing other model features. But linear model weights may still be interpreted better than deep neural network weights (Molnar, 2020).

- **Local Interpretability for a Single Prediction:** *Investigates why the model made a certain prediction for a certain instance.* This entails homing in on a single instance and examining what the model predicts for it and explaining why. For individual predictions, an otherwise complex model might behave more accessibly. Locally, predictions may only be linearly or monotonically dependent on some features, rather than complexly so. Say a house's value depends nonlinearly on its size. But when examining one particular 100 square meter house, it is possible for that subset, prediction depends linearly on size. This can be deduced by simulating how predicted price changes upon increasing or decreasing size by 10 square meters. Local explanations may therefore be more accurate than global ones. (Molnar, 2020).

- **Local Interpretability for a Group of Predictions -** *Answers why a model made specific predictions for a group of instances.* Multiple instance predictions may be explained either with global (modular level) interpretation methods or with individual instances. Global methods can be applied by taking the group, treating them as the complete dataset, and using global methods with the subset. Individual explanation methods can be used on each instance, then listed or aggregated for the entire group (Molnar, 2020).

Interpretation method: Various interpretation methods can be broadly distinguished based on their results. These can be summarised as follows (Molnar, 2020):

- *Feature summary statistics -* Several methods give summary statistics for every feature, with some providing a single number per feature (like feature importance), or more complex output, (e.g., pairwise feature interaction strengths) (Molnar, 2020).

- *Feature summary visualization -* Most feature summary statistics may also be visualized. Certain summaries only become meaningful if visualized, and a table would be the wrong choice. A feature's partial dependence is such a case, where plots are curves depicting a feature and the average predicted outcome. Partial dependences are ideally presented with the drawn curve rather than printed coordinates (Molnar, 2020).

- *Model internals (e.g., learned weights) -* Intrinsically interpretable models fall into this category; for instance, linear models' weights or learned decision trees' structure (features, and thresholds for splits). There is no clear distinction between feature summary statistic and model internals in cases like linear models, as weights represent them simultaneously. Another method eliciting model internals is the feature detectors visualization in convolutional neural networks. Such methods are, by definition modelspecific (Molnar, 2020).

- *Data point -* This category comprises all methods with data points (already existent or newly created) as outputs to facilitate interpretability. One such method is counterfactual explanations. To explain a data instance forecast, the method changes some features where the predicted outcome changes accordingly (like a class prediction change), to find a similar data point. Another instance is identifying predicted class prototypes. For utility, interpretation methods returning new data points need data points that themselves are interpretable. This has limited relevance for tabular data with hundreds of features but works well for images and texts (Molnar, 2020).

- *Intrinsically interpretable model -* One black box model interpretation solution is (global or local) approximations with interpretable models. The model itself is interpreted through internal feature summary statistics or model parameters (Molnar, 2020).

3.3.1 Explainable AI Techniques

While a technical review of XAI techniques is beyond the scope of this white paper, Nagahisarchoghaei et al. (2023) in their survey paper provide a useful visualisation of existing XAI techniques across three broad categories: (i) self-explainability; (ii) global post hoc explainability and (iii) local post hoc explainability.

3.3.2. Implications for Financial Risk Management

From the discussion above, a number of decisions need to be made in respect of XAI systems deployment for financial risk management. While self-explainability is possible with some financial risk modelling approaches, many AI models applied for this purpose are likely to require some form of post hoc operation to generate the explainability. Furthermore, it is likely that a suite of models is being used for risk assessment and so model agnostic approaches may be preferable, which can then be applied consistently across the suite of models. Additionally, while a financial risk management team will be interested in global explainability, the focus of financial risk management on tail risk means that local explainability is likely to be of greater value. The decision on what specific XAI techniques to deploy depends on the specific AI modelling used and the form of explainability required. Of course, in practice, several forms of explainability will be required to complete a full risk assessment.

In Section 4 we provide a use case application of alternative XAI techniques applied to the problem of credit risk management purposes. These alternative XAI techniques help to formulate a more complete picture of what drives credit defaults.

3.4 Approaches to Evaluating Explanations

In advance of EU laws regulating AI and some associated standards, a careful evaluation of XAI is essential to outline specific desirable properties. Given that the overarching goal of XAI is to establish trust among humans, it is crucial to prioritize properties such as human-friendliness, privacy, and non-discrimination

(Robnik et al., 2018; Miller, 2019). Ali et al. (2023) document five aspects of XAI evaluations.

First, explanation evaluation can be built up on cognitive psychology theories to articulate a general formal system of how humans can interpret. By examining the cognitive state of human users, investigations can improve efficiency of explanations and enhance user understanding of AI systems. To determine what kinds of XAI are preferred, measures of understandability of users on AI agents and algorithms are imperative (Dodge et al., 2018; Penney et al., 2018; Rader and Gray, 2015). It is also essential to consider users' attention and expectation in the process of incorporating explainability into AI systems (Stumpf et al., 2018).

Satisfaction is the second aspect of XAI evaluations. A diverse array of metrics, encompassing both subjective and objective measures, has been adopted to assess the clarity and adequacy of explanations (Miller, 2019). Curran et al. (2012) utilize a method involving ranking and coding of user transcripts to evaluate the effectiveness of explanations within a computer vision challenge. Lage et al. (2019) illustrate the importance of complexity of XAI model (length, intricacy) in affecting satisfaction. Confalonieri et al. (2021) gauge users' perceived understanding of explanations through task performance metrics, including accuracy and response time, as well as subjective measures like confidence level of user's responses.

The next aspect of XAI evaluation is trust and transparency. Cahour and Forzy (2009) adopt three trust scales in trust assessment of users. Nothdurft et al. (2014) examine the relationship between user trust and AI decision explanations, particularly focusing on transparency. Bussone et al. (2015) utilize a Likert scale and think-aloud protocols to appraise user trust in a clinical decision-support system, revealing that factual explanations contribute to an enhancement in user trust. Recently, Stepin et al. (2022) employed Likert scales to measure human perceptions of the trustworthiness of automated counterfactual explanations.

Assessment of human-AI interface is one aspect to evaluate XAI. Myers et al. (2006) introduce a framework allowing users to pose "why" and "why not" questions for coherent responses Lim et al. (2009) assess human performance using AI systems with varied explanations, considering task completion time and success rates. Evaluating the human-AI interface helps verify model outputs and debug specific AI models (Kulesza et al., 2015). Visual analytics tools like TopicPanaroma, FairSight, DGMTracker, aid domain experts in evaluating and reducing biases for fair data-driven decision-making.

The last aspect that Ali et al. (2023) propose for XAi evaluation is computational assessment. Not only human assessment, but system transparency may also be prioritized. In response, Herman (2017) advocates for computational approaches to evaluate explanation fidelity, focusing on the accuracy of saliency maps as indicators. Various computational methods have emerged to assess the validity, consistency, and fidelity of explainability techniques compared to the original blackbox model. Zeiler and Fergus (2014) demonstrate improved prediction outcomes through evaluating a CNN visualization tool's fidelity in detecting model flaws. Ross et al. (2017) evaluates the consistency and computing cost of explanations using LIME as a baseline, while Schmidt and Biessmann (2019) introduce an explanation quality score based on human intuition.

4. Use Case Demonstration

4.1. XAI Applied to Credit Risk Management

In this section we provide a use case demonstration in the credit risk management space. Schmitt and Cummins (2023) consider the application of post hoc XAI techniques to AI based modelling of credit default prediction. Specifically, the study considers two modelling approaches currently receiving attention in the credit risk management literature: namely, deep learning (DL) and gradient boosting (GB).

The study tackles the black box issue surrounding much of the recent literature that applies AI modelling to credit default prediction. Using XAI techniques, the study is able to provide insights into the key feature inputs that are driving the default predictions, moving beyond accuracy as the sole measure of performance.

While the authors perform their analysis on both credit card data and personal loan data, we focus on the latter for illustrative purposes in this white paper. Table 2 provides a summary of the personal loan data used and the key features recorded for a base of 1000 German banking customers. 300 of these customers are recorded as having defaulted on this debt. The DL and GB models were applied for credit default prediction using an 80%-20% training-testing split. Specifics around the configuration of the DL and GB models can be found in Schmitt and Cummins (2023), while the interested reader is directed to the discussion therein around performance accuracy. Of note here is that for the German dataset of personal loans, the GB model demonstrates the lower performance (AUC 0.868) relative to the best performing DL model (AUC 0.930).

Three layers of XAI analysis were conducted. The first is global feature importance in default prediction, the second is local feature analysis (via the Shapley value approach) on default prediction, and the third is partial dependence plotting to ascertain marginal effects on the default prediction.

Dataset 2 - German	
Variable	Description
X1	Balance of checking account
X2	Duration in months
X3	Credit history
X4	For what was the loan taken
X5	Credit amount
X6	Savings account plus bonds
X7	Duration of current employment
X8	Installment rate as % of income
X9	Marital status and gender
X10	Other debtors/guarantors
X11	Present residence since
X12	Type of owned properties
X13	Age of applicant
X14	Housing (rent, own, free)
X15	Credits at other banks
X16	Existing credits at this bank
X17	Employment/Level of qualification
X18	The number of dependents
X19	Registered telephone or none
X20	Immigrant/foreign worker

Source: Schmitt and Cummins (2023)

Table 2: German Personal Loan Data Description

Figure 3 summarises the global feature importance, providing the top ten features identified under each of the DL and GB model specifications. Notably there is consistency observed between the selection of key features for both models. However, there is significant divergence in the ranking of these features in terms of importance for default prediction. Such an observation is useful in explaining what drives default predictions across the two models, while it emphasises how different AI models can weight different input features quite differently. This level of explainability offers insights for an organisation in terms of continual monitoring and ongoing model risk management of AI based credit risk systems.

Further to this, Schmitt and Cummins (2023) highlight an issue around global feature importance analysis for the DL model. The DL approach introduces randomness through its configuration that means that separate runs of the feature importance analysis lead to completely different rankings of the most important features. This means one cannot be confident in the explanations returned from such analysis for the DL model. The GB model, due to its configuration, does not suffer from this issue and so the feature importance returned is robust. In light of the above, we follow Schmitt and Cummins (2023) and return on their localised feature analysis for only the GB model.
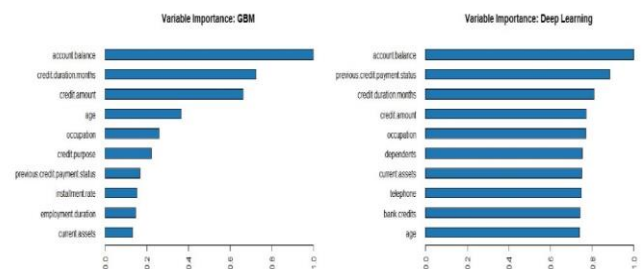
Source: Schmitt and Cummins (2023)

Figure 3: Global Feature Importance (Top 10 Rank)

Figure 4 presents the results of drilling down to establish the local effects that individual features have on the estimated probability of default. The approach used to determine this is the Shapley value approach, which leverages a sophisticated game theoretic framework. SHapley Additive exPlanation (SHAP) contribution analysis is a model-agnostic explanation approach that, like variable importance, identifies and ranks key features of a machine learning approach but also provides a summary of the impact that these features have on a localised basis. A local effect can thus be determined for each observation in the sample. Some effects are negative meaning that they decrease the likelihood of default estimate, while others are positive meaning that they increase the likelihood of default estimate.

The most important feature identified is a client's credit account balance. Our evidence

suggests that larger (smaller) credit account balances are associated with a negative (positive) impact on default prediction. This aligns with the intuition that larger (smaller) credit account balances are associated with clients with stronger (weaker) financial positions and repayment capacity.

The evidence around age further supports the above evidence pertaining to financial position and repayment capacity. The findings largely align with existing evidence. A general inverse relationship is observed between age and default risk, which when viewed via partial dependence plotting (Figure 5) suggests a decline in default likelihood as a borrower gets older, up until the age of 40 or so. Thereafter, the default rate increases somewhat again, although it stays below the default risk of younger age groups.

Final observations are made around the credit duration and credit amount. On the former, it is found via the SHAP explanations that credit default risk is higher for shorter-duration credit contracts. This tallies with the argument that the longer a credit line is in place then the more exposed the credit is to default. The SHAP values pertaining to the credit amount are somewhat mixed. We can see that high and low credit amounts are associated with increased default prediction. From a theoretical standpoint, arguments can be made for both directional observations.

Such local effects are very helpful for understanding the performance of the AI model on a localised basis, from which the organisation can then monitor and manage individual client exposures more closely.

As a final comment, towards evaluating the quality of the explanations generated, Schmitt and Cummins (2023) dedicate considerable effort to carefully benchmarking the explanations obtained from the XAI analysis with existing financial theory and empirical evidence. Interestingly, given that much of the AI literature in recent years has focused on performance accuracy only, derived from black box AI implementations, the authors had to revert to much earlier literature that utilised transparent self-explaining logistic regression

approaches to credit default modelling. This again emphasises the importance of moving in the direction of explain ability around AI systems for financial risk management.
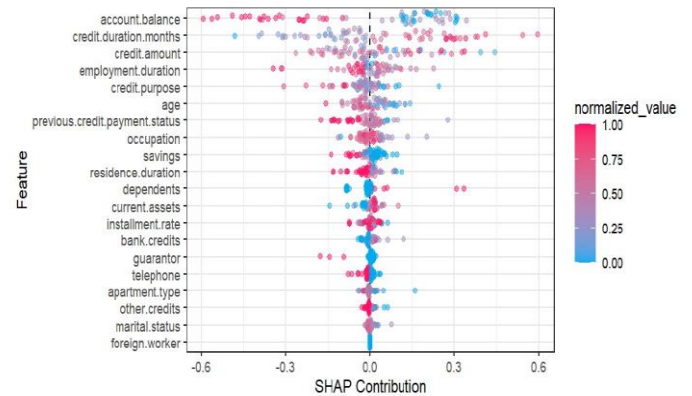
Source: Schmitt and Cummins (2023)

Figure 4: Local SHAP Contributions
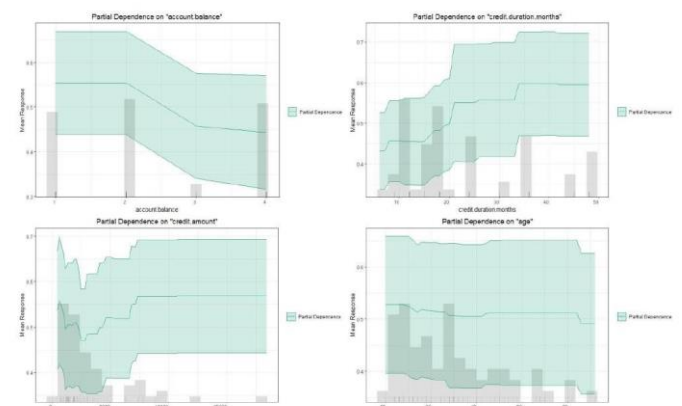
Source: Schmitt and Cummins (2023)

Figure 5: Partial Dependence (Selected Features)

5. Conclusion

In this white paper, we overview the opportunities that Explainable AI (XAI) offer to enhance financial risk management practice, which feeds into the objective of simplifying compliance for banking and financial services organisations. We provide a clear problem statement, which makes the case for explainability around AI systems from the business and the regulatory perspective. A comprehensive literature review positions the study and informs the solution framework

proposed. The solution framework sets out the key considerations of an organisation in terms of setting strategic priorities around the explainability of AI systems, the institution of appropriate model governance structures, the technical considerations in XAI analytics, and the imperative to evaluate explanations. The use case demonstration brings the XAI discussion to life through an application to AI based credit risk management, with focus on credit default prediction

.

Appendix A: Definitions

The technical terms and legislation discussed throughout this paper are defined, or elucidated upon here first to facilitate an easier understanding of the subject matter that follows. Here, we edify the relevant technical terms used, and subsequently elaborate upon the pertinent legislative definitions and legislation.

Technical Definitions

Explainable Artificial Intelligence (XAI) - Machine learning/artificial intelligence models often perform favourably relative to traditional econometric and linear models. They benefit by allowing for potentially complex, non-linear interactions among predictors. This renders them powerful but opaque. Thus, such models are often termed "black boxes," though they are easily analysed in many respects. Any exploration of these interaction effect is vexed by vast possibilities for identity and functional forms for predictors (Gu et al., 2020). Today, such complex black box algorithms are ubiquitously employed and deliver precise predictions and improved out-of-sample predictive performance, but frequently without explanations as to their decision making (Giglio et al., 2022), (Guidotti, et al., 2018). The use of so called XAI addresses this issue. It provides a framework which allows one to pinpoint what variables drive the performance or predictions of an algorithm, and what how important they are relative to each other.

Legislative Definitions and Pertinent Legislation

Next, we define the types of legislation as specified by the European Union (EU) or United Kingdom (UK) as applicable and available, as seen here, after which we elaborate upon the specific legislation pertinent to this study. We do not specify a specific definition when one cannot be found as defined by the EU or UK as applicable. The UK has or is in the process of developing sustainable finance and ESG legislation broadly equivalent to those of the EU. There are both distinctions and similarities in UK and EU legislation, which are elucidated

upon or highlighted in references across this article.

Roadmap - A roadmap is a plan that shows how a product or service is likely to develop over time. Roadmaps need to be easy to understand, and simple to adjust when priorities change - as often happens with agile ways of working. The definition of a roadmap is as specified on the UK Government's website here.

Standard - A standard is a document, established by consensus and approved by a recognised body. It provides rules, guidelines or characteristics for activities or their results so that they can be repeated. The aim is to achieve the greatest degree of order in a given context. The definition of a standard is as specified on the UK Government's website here.

Types of legislation - The aims set out in the EU treaties are achieved by several types of legal act. Some are binding, others are not. Some apply to all EU countries, others to just a few.

Regulations - A "regulation" is a binding legislative act. It must be applied in its entirety across the EU. For example, when the EU's regulation on ending roaming charges while travelling within the EU expired in 2022, the Parliament and the Council adopted a new regulation both to improve the clarity of the previous regulation and make sure a common approach on roaming charges is applied for another ten years.

Directives - A "directive" is a legislative act that sets out a goal that EU countries must achieve. However, it is up to the individual countries to devise their own laws on how to reach these goals. One example is the EU single-use plastics directive, which reduces the impact of certain single use plastics on the environment, for example by reducing or even banning the use of single-use plastics such as plates, straws and cups for beverages.

FinTech Research & Innovation Roadmap 2022-31 - A document aimed at providing a practical pathway to accelerate the

development of FinTech excellence, and to embrace opportunities across the financial services industry and the broader economy in Scotland and the UK. It aligns with the recommendations set out in the Kalifa Review of UK Fintech in February 2021, and supports the UK's national ambition to encourage growth by creating the right conditions for innovation. It was published with the objective of boosting economic recovery, driving growth, and creating jobs over the next ten years. Over the ten-year period, the ambition is to deliver in Scotland an additional 20,000 plus fintech related jobs as well as produce an increase in economic gross value add (GVA) through fintech innovation from £0.5bn GVA today to £2.1bn GVA by 2031. The industry led roadmap is the first of its kind in the UK and has been pulled together by the cluster body FinTech Scotland in collaboration with fintech entrepreneurs, the financial services sector, academia, regulators, Government bodies and consumer groups. The cross industry led collaboration has resulted in four key strategic innovation themes which provide the foundation for the roadmap, these are open finance data, climate finance, financial regulation and payments and transactions. The roadmap can be found in full on the FinTech Scotland website here, with an overview available here. The associated text above is sourced from these two web pages.

The Kalifa Review of UK FinTech - An independent report on the UK Fintech sector by Ron Kalifa OBE. At Budget 2020, the Chancellor asked Ron Kalifa OBE to conduct an independent review to identify priority areas to support the UK's fintech sector. The Review formally launched in July 2020 with objectives for supporting the growth and widespread adoption of UK fintech, and for maintaining the UK's global fintech reputation. The review can be found in full along with an executive summary on the UK Government's website here. The associated text above is sourced from this web page.

The EU's Corporate Sustainability Reporting Directive (CSRD), subject to European Sustainability Reporting Standards (ESRS) and the Task Force on Climate-related Financial Disclosures (TCFD), now supplanted by IFRS S1 and S2 - To better comprehend financial risks and opportunities, there is increasing scrutiny on sustainability and climate disclosures of enterprises by corporations, governments, and investors. Worldwide regulators and benchmark setters have introduced sustainability and climate reporting frameworks and rules to enhance the quality and quantity of the relevant information (Manifest-Climate, 2023).

The EU's CSRD and the TCFD are the world's leading climate reporting frameworks, both target improvements in entity disclosures of climate- and sustainability-related opportunities and risks. A further design intention is standardizing and harmonizing climate and sustainability reporting across companies and jurisdictions (Manifest-Climate, 2023).

Both entail company level disclosures and apply to large corporations and all listed organisations. The necessary disclosures under both are of sustainability related opportunities and risks. TCFD's scope is a subset of CSRD, as CSRD covers all sustainability topics (ESG), while TCFD is designed for ESG related to climate topics only (Manifest-Climate, 2023).

January 5, 2023, marked the date the EU's CSRD requirements came into effect. Roughly 50,000 companies — as defined above — compulsorily needed to disclose their sustainability risks and opportunities related to environmental and social issues under them (Manifest-Climate, 2023).

The directive mandates company reports based on the European Sustainability Reporting Standards, developed by EFRAG (previously the European Financial Reporting Advisory Group).Firms that fulfil the eligibility criteria must commence reporting by fiscal year 2024 (ManifestClimate, 2023).

There are many similarities between the CSRD and TCFD. A core one is that they both call for the robust companies' climate-related financial risks and opportunities reporting. The former's text on climate disclosures is in broad agreement with the four TCFD pillars —

governance, strategy, risk management, and metrics and targets. Further, both aim to promote capital market transparency and accountability, along with standardizing climate- and sustainability-related disclosures (Manifest-Climate, 2023).

Several other similarities exist in relation to governance, strategy, risk management, metrics and targets. Crucially, key distinctions also exist along the lines of scope (as discussed above), double materiality, compatibility with the 1.5°C transition, impact mitigation actions, strategic implementation, and effective disclosure preparedness for companies. For brevity, these are not expanded upon here, but interested readers are directed to (Manifest-Climate, 2023). Their publications that provide a comprehensive overview of the TCFD; guidance on metrics, targets, and transition plans; the TCFD's recommendations; and how to implement these recommendations can be found here on their website.

It is important to note (as stated on their website) that ``Concurrent with the release of its 2023 status report on October 12, 2023, the TCFD has fulfilled its remit and disbanded. The FSB has asked the IFRS Foundation to take over the monitoring of the progress of companies' climaterelated disclosures. As of November 2023, this website will no longer be updated or monitored but will remain available to serve as a resource for materials developed by the Task Force. The Task Force is deeply grateful to all parties involved for their input, support, and adoption of the TCFD recommendations.''

The IFRS has subsequently issued two inaugural global sustainability disclosure standards IFRS S1 General Requirements for Disclosure of Sustainability-related Financial Information, and IFRS S2 Climate-related Disclosures. Both fully incorporate the recommendations and are built on the framework of the TCFD. They consolidate the TCFD recommendations and framework with other standards and frameworks, including the SASB Standards, CDSB Framework, Integrated Reporting Framework and World Economic Forum metrics, to streamline sustainability disclosures. Details on these standards can be found on the IFRS website here, here, and here.

Overviews on them can also be found on the Cambridge Institute for Sustainability Leadership website here, and The Institute of Chartered Accountants in England and Wales website here for interested readers.

The EU's CSRD ESRS regulation's applicability begins with 2024 data and 2025 reporting (O'Connell, 2023). ESRS were initially adopted in July, 2023 and EFRAG released a draft data points list of the ESRS on the 25th of October, 2023.A high level info graphic of the ESRS and the data points from (O'Connell, 2023) can be seen in the Figure below from (O'Connell, 2023). For brevity, details are not expanded upon here, but interested readers are directed to (O'Connell, 2023).

The EU's Sustainable Finance Disclosure Regulation (SFDR) and the UK's Sustainability Disclosure Regulation (SDR) - Both these regulations, generally speaking can be categorised as product-level sustainability disclosures for financial market participants. These regulations are both designed with the aim of greater disclosure and transparency on sustainable finance investments and products, furnish further investment information to investors for informed decision making, and fight green washing through integrity and trust building in sustainable instruments (Vincent, 2023). More specifically, we shed light on each regulation below.

The European Commission-led EU SFDR, set out disclosure requirements on sustainability for financial market participants, for example investment firms, insurance, and reinsurance companies. It is applicable for EU domiciled firms, and for products marketed in the EU, regardless of business location (Vincent, 2023).

In contrast, improved sustainability information from issuers and investment managers that is more comparable, consistent, and comprehensive is the goal of the UK SDR. It is spearheaded by the UK Financial Conduct Authority (FCA). Its ambit extends to share and bond issuers that are regulated market listed or investment managers in the UK (Vincent, 2023).

Thus, the EU SFDR scope is restricted to EU-based companies and entities marketing products in the EU, whereas the UK SDR coverage extends to purely UK-based companies. A considerable number of firms may be required to comply with both regulations (Vincent, 2023). Similar to the case of the CSRD and TCFD, commonalities and differences exist for SFDR and SDR. For brevity, these are not expanded upon here, but interested readers are directed to (Vincent, 2023), (Simmons and Simmons, 2021). The full text of the SDR Policy Statement published in November, 2023 can be found here on the FCA's website.

UK's Climate-Related Financial Disclosure (Department for Energy Security and Net-Zero) – Climate-related financial disclosures for companies and limited liability partnerships (LLPs). Inscope companies and limited liability partnerships (LLPs) need to meet these new mandatory climate-related financial disclosure requirements under the Companies (Strategic Report) (Climate-related Financial Disclosure) Regulations 2022 and the Limited Liability Partnerships (Climate-related Financial Disclosure) Regulations 2022. The regulations were made on 17 January 2022 and apply to reporting for financial years starting on or after 6 April 2022. The guidance to help meet these disclosure requirements can be found in full here on the UK Government's website. The associated text above is sourced from this web page.

UK Sustainability Disclosure Standards (SDS) - Corporate disclosures on sustainability-related opportunities and risks that companies face are set out by the UK SDS. Subsequent UK regulation or legislation reporting requirements on opportunities and risks related to sustainability matters, including those stemming from climate change. They are anticipated to become effective on January 1, 2025, and published in July 2024 at the latest by the UK Department for Business and Trade (DBT). They will be based on the IFRS® Sustainability Disclosure Standards issued by the International Sustainability Standards Board (ISSB). They will be adopted into UK entities' legal and regulatory reporting requirements after their creation and publication (Brightest, 2023).

Interested readers are directed to (Brightest, 2023) for more detail

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138-52160.

Adams, J., & Hagras, H. (2020, July). A type-2 fuzzy logic approach to explainable AI for regulatory compliance, fair customer outcomes and market stability in the global financial sector. In *2020 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1-8). IEEE.

Agarwal, S., Muckley, C. B., & Neelakantan, P. (2023). Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods. *Economics Letters*, 111117.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805.

Babaei, G., Giudici, P., & Raffinetti, E. (2022). Explainable artificial intelligence for crypto asset allocation. *Finance Research Letters*, 47, 102941.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics*, 30-56.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., . . . others. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., . . . Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Blattner, L., Stark, P.-R., & Spiess, J. (2022). Machine Learning Explainability \& Fairness: Insights from Consumer Lending. *FinRegLab Whitepaper*.

Brightest. (2023, 08 05). *UK Sustainability Disclosure Standards (SDS) - Legislation Overview, Rules, & Requirements*. Retrieved from Brightest: https://www.brightest.io/uk-sustainability-disclosure-standardssustainability-disclosure-standards.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable machine learning in credit risk management. Computational Economics, 57, 203-216.

Bussmann, N., Giudici, P., Tanda, A., & Yu, E. P. Y. (2023). Explainable machine learning models to identify the key drivers of the implied cost of capital. *Available at SSRN 4173890*.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics* (pp. 160-169). IEEE.

Cahour, B., & Forzy, J. F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science*, 47(9), 1260-1270.

Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G., Naseer, A., Regoli, D., & Gonzalez, B. S. (2020). Befair: Addressing fairness in the banking sector. *2020 IEEE International Conference on Big Data (Big Data)*, 3652-3661.

Chen, Y., Giudici, P., Liu, K., & Raffinetti, E. (2022). Measuring Fairness in Credit Scoring. *Available at SSRN 4123413*.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(1), e1391.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. *Journal of Business Research*, 66(9), 1629-1636.

Curran, W., Moore, T., Kulesza, T., Wong, W. K., Todorovic, S., Stumpf, S., ... & Burnett, M. (2012, February). Towards recognizing" cool" can end users help computer vision recognize

subjective attributes of objects in images?. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces* (pp. 285-288).

Del Vitto, A., Marazzina, D., & Stocco, D. (2023). ESG ratings explainability through machine learning techniques. *Annals of Operations Research*, 1-30.

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., & Santos, A. A. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 103737.

Dodge, J., Penney, S., Anderson, A., & Burnett, M. M. (2018). What Should Be in an XAI Explanation? What IFT Reveals. In *IUI Workshops* (pp. 1-4).

Dudik, M., Chen, W., Barrocas, S., Inchiosa, M., Lewins, N., Oprescu, M., . . . Wallach, H. (2020). Assessing and mitigating unfairness in credit models with the Fairlearn toolkit. *Microsoft and EY White Paper*.

Fritz-Morgenthal, S., Hein, B., & Papenbrock, J. (2022). Financial risk management and explainable, trustworthy, responsible AI. *Frontiers in Artificial Intelligence*, 5, 779799.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, 5-47.

Giglio, S., Kelly, B., & Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 337-368.

Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, 32, 100577.

Grennan, L., Kremer, A., Singla, A., & Zipparo, P. (2022). Why businesses need explainable AI—and how to deliver it. *URL: Why businesses need explainable AI| McKinsey. Accessed*, *12*, 2022.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 2223-2273.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 1-42.

Haefner, N., Wincent, J., Parida, V., & Gassmann, O. (2021). Artificial intelligence and innovation management: A review, framework, and research agenda☆. *Technological Forecasting and Social Change*, 162, 120392.

Herman, B. (2017). The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*.

Hoepner, A. G., McMillan, D., Vivian, A., & Wese Simen, C. (2021). Significance, relevance and explainability in the machine learning age: an econometrics and financial data science perspective. *The European Journal of Finance*, 1-7.

Jarovsky, A., Milo, T., Novgorodov, S., & Tan, W. C. (2018, April). Rule sharing for fraud detection via adaptation. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)* (pp. 125-136). IEEE.

Kaniel, R., Lin, Z., Pelger, M., & Van Nieuwerburgh, S. (2023). Machine-learning the skill of mutual fund managers. *Journal of Financial Economics*, 94-138.

Karimi, H., Khan, M. F., Liu, H., Derr, T., & Liu, H. (2022). Enhancing individual fairness through propensity score matching. *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)*, 1-10.

Kovvuri, V. R. R., Fu, H., Fan, X., & Seisenberger, M. (2023). Fund performance evaluation with explainable artificial intelligence. *Finance Research Letters*, 58, 104419.

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 1083-1094.

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015, March). Principles of explanatory debugging to personalize interactive machine

learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 126-137).

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., & Doshi-Velez, F. (2019, October). Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, pp. 59-67).

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009, April). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119-2128).

Lin, B., & Bai, R. (2022). Machine learning approaches for explaining determinants of the debt financing in heavy-polluting enterprises. *Finance Research Letters*, *44*, 102094.

Lu, X., & Calabrese, R. (2023). The Cohort Shapley value to measure fairness in financing small and medium enterprises in the UK. *Finance Research Letters*, 104542.

Manifest-Climate. (2023, 05 11). *REGULATION How Does the CSRD Compare to the TCFD?*

Retrieved from Manifest Climate Blog:

https://www.manifestclimate.com/blog/comparing-csrd-

tcfd/#:~:text=The%20European%20Union's%20Corporate%20Sustainability,sustainability%2Drelated%20risks%20and%20opportunities.

Martin, K. (2023). Who Counts in Business Ethics. *Business Ethics Quarterly*, 216-243.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1-38.

Molnar, C. (2020). *Interpretable machine learning.* Lulu. com.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review,

synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876.

Myers, B. A., Weitzman, D. A., Ko, A. J., & Chau, D. H. (2006, April). Answering why and why not questions in user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 397-406).

Nagahisarchoghaei, M., Nur, N., Cummins, L., Nur, N., Karimi, M. M., Nandanwar, S., ... & Rahimi, S. (2023). An empirical survey on explainable AI technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics*, *12*(5), 1092.

Nothdurft, F., Richter, F., & Minker, W. (2014, June). Probabilistic human-computer trust handling. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (SIGDIAL) (pp. 51-59).

O'Connell, R. (2023, 10 26). *EFRAG's 1178 ESRS Data Points*. Retrieved from Nossdata Blog: https://www.nossadata.com/blog/efrag-1178-esrs-data-points.

OECD (2021), Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers, https://www.oecd.org/finance/financial-markets/Artificial-intelligence-machine-learning-big-data-in-finance.pdfmarkets/Artificial-intelligence-machine-learning-big-data-in-finance.pdf

Padmanabhan, B., Sahoo, N., & Burton-Jones, A. (2022). Machine learning in information systems research. *Management Information Systems Quarterly*, 46(1), iii-xix.

Park, M. S., Son, H., Hyun, C., & Hwang, H. J. (2021). Explainability of machine learning models for bankruptcy prediction. *IEEE Access*, *9*, 124887-124899.

Penney, S., Dodge, J., Hilderbrand, C., Anderson, A., Simpson, L., & Burnett, M. (2018, March). Toward foraging for understanding of StarCraft agents: An empirical study. In *23rd International Conference on Intelligent User Interfaces* (pp. 225-237).

Rader, E., & Gray, R. (2015, April). Understanding user beliefs about algorithmic curation in the Facebook news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 173-182).

Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-based explanations of prediction models. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, 159-175.

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717.*

Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to?

A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), 1-9.

Sætra, H. S. (2023). The AI ESG protocol: Evaluating and disclosing the environment, social, and governance implications of artificial intelligence capabilities, assets, and activities. *Sustainable Development*, 31(2), 1027-1037.

Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558.*

Schmitt, M., & Cummins, M. (2023). Beyond Accuracy in Artificial Intelligence Based Credit Scoring Systems: Explainability and Sustainability in Decision Support. *Available at SSRN 4536400*.

Scott, T. (2015). Does collaboration make any difference? Linking collaborative governance to environmental outcomes. *Journal of Policy Analysis and Management*, 34(3), 537-566.

Seele, P. (2017). Predictive Sustainability Control: A review assessing the potential to transfer big data driven 'predictive policing' to corporate sustainability management. *Journal of Cleaner Production*, 153, 673-686.

Sigrist, F., & Hirnschall, C. (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking & Finance*, 102, 177-192.

Simmons and Simmons. (2021, 10 22). *ESG: UK sets out its roadmap on sustainable finance*. Retrieved from Simmons and Simmons Publications: https://www.simmons-simmons.com/en/publications/ckv29tpyf196d0b97chfwb91r/esg-uk-sets-out-its-roadmap-on-sustainable-finance

Stepin, I., Alonso-Moral, J. M., Catala, A., & Pereira-Fariña, M. (2022). An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences*, 618, 379-399.

Stumpf, S., Skrebe, S., Aymer, G., & Hobson, J. (2018, March). Explaining smart heating systems to discourage fiddling with optimized behavior. In *CEUR Workshop Proceedings* (Vol. 2068).

Tron, A., Dallocchio, M., Ferri, S., & Colantoni, F. (2023). Corporate governance and financial distress: Lessons learned from an unconventional approach. *Journal of Management and Governance*, 27(2), 425-456.

Vincent, M.-A. (2023, 02 09). *The UK SDR vs EU SFDR – What financial organisations should know*. Retrieved from Sweep Blog: https://www.sweep.net/blog/the-uk-sdr-vs-eu-sfdr-whatfinancial-organisations-should-know

Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 1-27.

Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multimodal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77, 29-52.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In

Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13 (pp. 818-833). Springer International Publishing.

Zhang, T., Zhu, W., Wu, Y., Wu, Z., Zhang, C., & Hu, X. (2023). An explainable financial risk early

warning model based on the DS-XGBoost model. *Finance Research Letters*, 104045.

Zheng, X. L., Zhu, M. Y., Li, Q. B., Chen, C. C., & Tan, Y. C. (2019). FinBrain: when finance meets AI 2.0. *Frontiers of Information Technology & Electronic Engineering*, 20(7), 914-924.

Zhou, L., Shi, X., Bao, Y., Gao, L., & Ma, C. (2023). Explainable artificial intelligence for digital finance and consumption upgrading. *Finance Research Letters*, 58, 104489.

About the Authors

Dr. James Bowden is Lecturer in Financial Technology at the Strathclyde Business School, University of Strathclyde, where he is the programme director of the MSc Financial Technology. Prior to this, he gained experience as a Knowledge Transfer Partnership (KTP) Associate at Bangor Business School, and he has previous industry experience within the global financial index team at FTSE Russell. Dr Bowden's research focusses on different areas of financial technology (FinTech), and his published work involves the application of text analysis algorithms to financial disclosures, news reporting, and social media. More recently he has been working on projects incorporating audio analysis into existing financial text analysis models, and investigating the use cases of satellite imagery for the purpose of corporate environmental monitoring. Dr Bowden has published in respected international journals, such as the European Journal of Finance, the Journal of Comparative Economics, and the Journal of International Financial Markets, Institutions and Money. He has also contributed chapters to books including "Disruptive Technology in Banking and Finance", published by Palgrave Macmillan. His commentary on financial events has previously been published in The Conversation UK, the World Economic Forum, MarketWatch and Business Insider, and he has appeared on international TV stations to discuss financial innovations such as non-fungible tokens (NFTs).

Email: james.bowden@strath.ac.uk

Professor Mark Cummins is Professor of Financial Technology at the Strathclyde Business School, University of Strathclyde, where he leads the FinTech Cluster as part of the university's Technology and Innovation Zone leadership and connection into the Glasgow City Innovation District. As part of this role, he is driving collaboration between the FinTech Cluster and the other strategic clusters identified by the University of Strathclyde, in particular the Space, Quantum and Industrial Informatics Clusters. Professor Cummins is the lead investigator at the University of Strathclyde on the newly funded (via UK Government and Glasgow City Council) Financial Regulation Innovation Lab initiative, a novel industry project under the leadership of FinTech Scotland and in collaboration with the University of Glasgow. He previously held the posts of Professor of Finance at the Dublin City University (DCU) Business School and Director of the Irish Institute of Digital Business. Professor Cummins has research interests in the following areas: financial technology (FinTech), with particular interest in Explainable AI and Generative AI; quantitative finance; energy and commodity finance; sustainable finance; model risk management. Professor Cummins has over 50 publication outputs. He has published in leading international discipline journals such as: European Journal of Operational Research; Journal of Money, Credit and Banking; Journal of Banking and Finance; Journal of Financial Markets; Journal of Empirical Finance; and International Review of Financial Analysis. Professor Cummins is co-editor of the open access Palgrave title *Disrupting Finance: Fintech and Strategy in the 21st Century*. He is also co-author of the Wiley Finance title *Handbook of Multi-Commodity Markets and Products: Structuring, Trading and Risk Management*.

Email: mark.cummins@strath.ac.uk

Daniel Dao is a Research Associate at the Financial Regulation Innovation Lab (FRIL), University of Strathclyde. Besides, he is Doctoral Researcher in Fintech at Centre for Financial and Corporate Integrity, Coventry University, where his research topics focus on fintech (crowdfunding), sustainable finance and entrepreneurial finance. He is also working as an Economic Consultant at World Bank Group, Washington DC Headquarters, where he has been contributing to various policy publications and reports, including World Development Report 2024; Country Economic Memorandum of Latin American and Caribbean countries; Policy working papers of labor, growth, and policy reforms, etc.... Regarding professional qualifications and networks, he is CFA Charterholder and an active member of CFA UK. He has earned his MBA (2017) in Finance from Bangor University, UK, and his MSc (2022) in Financial Engineering from WorldQuant University, US. He has shown a strong commitment and passion for international development and high-impact policy research. His proficiency extends to data science techniques and advanced analytics, with a specific focus on artificial intelligence, machine learning, and natural language processing (NLP).

Email: daniel.dao@strath.ac.uk

Kushagra Jain is a Research Associate at the Financial Regulation Innovation Lab (FRIL), University of Strathclyde. His research interests include artificial intelligence, machine learning, financial/regulatory technology, textual analysis, international finance and risk management, among others. He is a recipient of doctoral scholarships from the Financial Mathematics and Computation Cluster (FMCC), Science Foundation Ireland (SFI), Higher Education Authority (HEA) and Michael Smurfit Graduate Business School, University College Dublin (UCD). Previously, he worked within wealth management and as a statutory auditor. He is due to complete his doctoral studies in Finance from UCD in 2023, and obtained his MSc in Finance from UCD, his Accounting Technician accreditation from the Institute of Chartered Accountants of India and his undergraduate degree from Bangalore University. He is a FMCC Database Management Group Data Manager, Research Assistant, PhD Representative and Teaching Assistant for undergraduate, graduate and MBA programmes.

Email: Kushagra.jain@strath.ac.uk

Get in touch

FRIL@FinTechscotland.com

FinTech Scotland®

University of Glasgow

University of Strathclyde Glasgow