Opportunities in Real Time Fraud Detection: An Explainable Artificial Intelligence (XAI) Research Agenda

Eleanor Mill, Wolfgang Garn, Nick Ryman-Tubb, Chris Turner
Surrey Business School, University of Surrey, Guildford, GU2 7XH

***Abstract*—Regulatory and technological changes have recently transformed the digital footprint of credit card transactions, providing at least ten times the amount of data available for fraud detection practices that were previously available for analysis. This newly enhanced dataset challenges the scalability of traditional rule-based fraud detection methods and creates an opportunity for wider adoption of artificial intelligence (AI) techniques. However, the opacity of AI models, combined with the high stakes involved in the finance industry, means practitioners have been slow to adapt. In response, this paper argues for more researchers to engage with investigations into the use of Explainable Artificial Intelligence (XAI) techniques for credit card fraud detection. Firstly, it sheds light on recent regulatory changes which are pivotal in driving the adoption of new machine learning (ML) techniques. Secondly, it examines the operating environment for credit card transactions, an understanding of which is crucial for the ability to operationalise solutions. Finally, it proposes a research agenda comprised of four key areas of investigation for XAI, arguing that further work would contribute towards a step-change in fraud detection practices.**

Keywords*—Artificial intelligence; explainable AI; machine learning; credit card fraud

I. Introduction

Europol's Serious and Organised Crime Threat Assessment identifies non-cash payment fraud as one of the most concerning criminal activities in the European Union [1]. In the UK alone, fraud losses on UK issued cards totalled £567 million in 2020 [2]. UK losses, however, are dwarfed in comparison to global losses which were estimated to be $32.39 billion in 2020, extending to over $40 billion by 2027 [3]. It is argued that as the use of non-cash payment cards increases year on year, perpetrators of these frauds are likely to see a continual increase in their illegal funding unless industry and academics can come together to create a significant step-change in the way in which fraudulent transactions are intercepted.

A. Changing Landscape

The volumes and velocity of credit card transactions means that financial institutions cannot rely on human expertise alone to identify fraudulent transactions. Fraud Management Systems (FMS) complement other internal processes to help automate fraud detection and decision-making. FMSs are traditionally rule based, meaning every single transaction is checked against a catalogue of pre-determined rules. This is an approach favoured by industry fraud experts because of the ease with which they can understand the inputs, modify the rules and interpret the results. However, whilst the relative simplicity of rule-based systems ensures the results are easily understood, this fixed approach does not scale well and limits the ability of the FMS to recognise or adapt to evolving patterns of fraud. Moreover, recent regulatory and technological developments threaten the effectiveness of traditional rule-based fraud management systems. As a consequence the payments industry, and therefore payment card fraud detection, is facing a once-in-a-generation need for radical change.

1) Regulatory developments: As part of the Payment Services Directive 2 (PSD2) regulation, Strong Customer Authentication (SCA) has recently been enforced in Europe and the United Kingdom [4]. SCA employs new Regulatory Technical Standards (implemented through an initiative called 3-D Secure 2.0) which enhance the current practices of processing customer transaction data. One of the pre-SCA challenges for issuers in fraud detection was the limited amount of data they received from the retailer – typically less than 10 variables per transaction. In contrast, the new Regulatory Technical Standards describe "Authentication Enrichment" data that a retailer should now provide to an issuer in addition to the usual transaction data. The Authentication Enrichment data increases the original 10 variables to over 100 variables (known as "security features") [5], [6] as shown in Fig. 1 .

The ten-fold increase in the security features necessitates a step-change in traditional rule-based fraud detection methodologies. Whilst rule-based engines will continue to perform initial screening of transactions to eliminate the most common fraud approaches, machine learning (ML) will be required to perform the majority of the analysis. Synergistically, the results of those ML models must be easily translated by the fraud analysts and management teams in order to interpret and act upon any newly derived insights.

Additionally, the Regulatory Technical Standards dictate the need to perform the analysis of transactions using these data points in real-time. The adoption of Authentication Enrichment data and enforcement of real-time analysis makes improvements to the automated processing of transactions increasingly urgent: It is claimed that "Approximately 80% of issuers plan to invest in machine-learning (ML) and rule-based engines to facilitate SCA processes by the end of 2021" [7].

2) Technology developments: Technology is revolutionising the way society pays for its goods and services. Contactless technology has become mainstream [8] and digital wallets such as Apple Pay, Google Pay or Samsung Pay have significantly increased their user base, especially in the younger generations

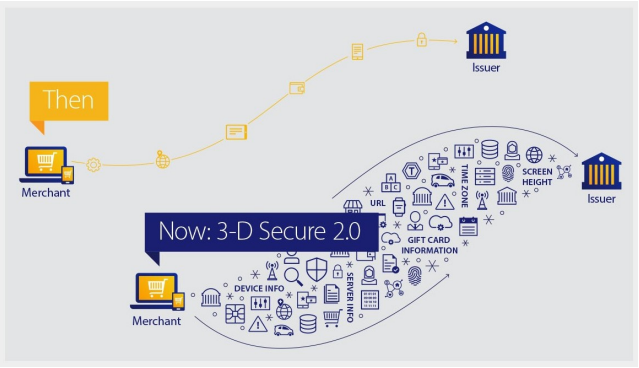


Fig. 1. Additional security features to be provided from the merchant (retailer) to the issuer as a result of new SCA regulatory technical standards (implemented through an initiative called 3-D Secure 2.0) [6].

[9]. Using this technology, payments can now be made through physical cards, mobile phones and even jewellery such as rings or watches which use Near Field Communication (NFC) technology. In addition, Open Banking has facilitated the entrance of a myriad of new payments service providers and the introduction of account-to-account payments [10].

These innovations not only transform the footprint of a traditional payment transaction but also highlight the flexibility needed to address the future transaction landscape. Traditional rule-based fraud detection methodologies which rely on a user's consistent and repetitive behaviours are less effective when payments can be made through any device, piece of clothing or jewellery in any location and at any time. Similarly, changes to the way payments are conducted means a re-evaluation of a retailer's payment infrastructure [11] which, in turn, will also affect their traditional fraud detection methodologies.

Finally, recent advances in technology enable fraudsters to work en masse, causing disruption at an ever faster pace. In [12], Dvorsky reported on the already-present ability of criminals to launch AI-based attacks, enabling much faster and more widespread disruption than previous human manual led strikes. Fraudsters are agile. They do not have the restraints of customer privacy, regulation and legacy applications to accommodate. In a recent industry report [13], Mike Haley, CIFAS CEO said "raud is ever evolving, and criminals continue to collaborate. As a community, we must do the same".

Hence the need for the FMS to be able to adapt at pace becomes even more critical. Rule-based systems may have been sufficiently refined over the past 30 years to effectively seek out known fraud patterns or traits, yet it is suggested that they are no match for the dynamics of this modern fraud landscape. As a consequence, the accuracy of the traditional FMS over the medium-to-long term will decline.

B. Current Status

To address the challenges of escalating transaction volumes, changes in regulation, technological advancements and a more sophisticated and technology-savvy criminal fraternity, researchers are exploring the opportunities of employing ML techniques in credit card fraud detection. However, adoption of ML techniques in financial settings have been slow to materialise [14]. The running hypothesis is that organisations perceive ML techniques as "black box" solutions which lack transparency and are therefore difficult to trust. Some domains, for example movie recommendation engines, are able to tolerate the opacity which accompanies black box solutions since the consequences of an incorrect outcome (for example a poor movie recommendation), whilst potentially irritating, present a low risk to the user.

In financial domains the consequences of an incorrect decision on a data subject are more impactful. In the case of credit card fraud detection, a consumer is likely to have the transaction rejected, and potentially the credit card subsequently withheld or cancelled. At the very least this will result in annoyance or embarrassment, but it may also impact the consumer's ability to buy groceries or keep up with payments on more substantial items. The existence of these risks places a much stronger onus on practitioners to ensure they can trust in the outputs of these ML models. For the finance industry, the inability to understand or justify the outcomes of the black box ML models has consequentially become a strong barrier to change.

To counter this challenge, scholars have begun investigating ways in which ML techniques can be leveraged whilst simultaneously providing transparency to engender trust in the models and therefore encourage more ubiquitous adoption. An emerging and increasingly popular technique to create this transparency is Explainable Artificial Intelligence (XAI).

C. Terminology

Scholarly research of nascent fields often begins with the difficulty of achieving a consensus on normative terminology. This is especially pertinent for the discourse surrounding XAI. As noted by both [15] and [16], many authors avoid committing themselves to a definition of an XAI system. This may be because, as a nascent field, the community have yet to come together to agree upon a clear definition. Yet without open discussion, how can consensus be reached? Those same authors suggest that this avoidance exposes the discipline to criticism that the field lacks rigour, noting that the community cannot justify claims of delivering XAI without agreement as to what XAI is.

To complicate matters further, there is also discord between authors regarding use of the terms "explainable" (usually followed by "artificial intelligence" and denoted XAI) and "interpretable" (usually followed by "machine learning" and denoted IML) with some authors considering the two terms analogous [17], [18] and other authors seeing a clear distinction between them.

One suggestion [19] is that the term "explainable" should be considered an umbrella term which has the goal to "... summarise the reasons for neural network behaviour, gain the trust of users, or produce insights about the causes of their decisions". The authors then perceive interpretability as a sub-goal to shed light on "what a model did or might have done" – answering the question of "how" the system came to its conclusion, yet stopping short of providing the complete response which a system audit may require. An explainable model is therefore, by definition, inherently interpretable yet

the reverse is not true – an interpretable model does not necessarily satisfy all the requirements of being explainable.

Similarly, [20] provide an holistic definition of XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future." Their concept of interpretability, analogous with the perspective of [19], is also subservient to the concept of being explainable. However the authors are more precise in their description, suggesting that "Interpretable models are machine learning techniques that learn more structured, interpretable, or causal models."

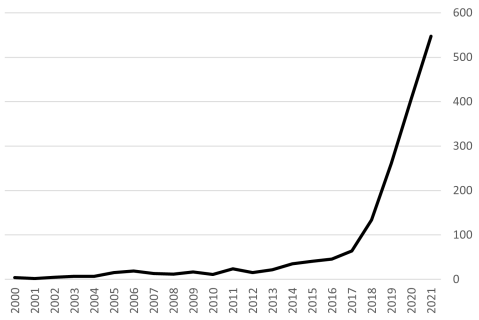
This concept of interpretability representing models that can be decomposed by an appropriately skilled audience is becoming more widely recognised amongst contemporary authors. Specifically, authors identify linear models, decision trees, rule-based models and constrained variants of black-box models as interpretable models [21]–[25]. Such models are often referred to as "inherently" interpretable [22], [23], [26], [27] or "intrinsic" [28], with the advantage that they are able to provide accurate and undistorted [26] explanations for the model output.

In contrast, black box models are often defined as models which are not interpretable, that is their complexity is so acute that the intended audience are unable to unravel their inner workings. When presented with such a model, it is increasingly commonplace for those seeking an explanation of the output to implement a subsequent interpretable model in a post-hoc fashion, the purpose of which is to find an approximate and human-understandable explanation to the original model's output. Obsfucating the holistic definitions of both [19] and [20], authors frequently refer to these post-hoc models as explainable models [15], [22], [29] or explainable AI [23], [27], although others employ the term post-hoc interpretability [21], [30].

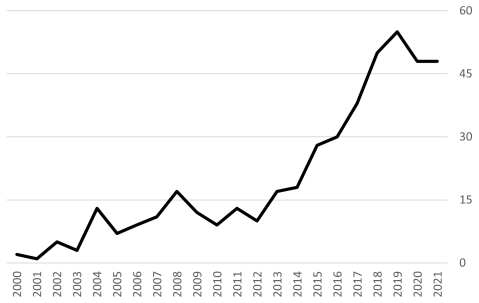
In an effort to reconcile the discourse, this paper leverages the holistic perspectives of both [19] and [20] to suggest XAI should be considered as an umbrella term. Specifically, it adopts the definition put forward by [20] (see above) which emphasises the importance of producing an explanation that is human-understandable, including transparency of the working parameters of the system. Where necessary, it differentiates XAI models through use of the terms "intrinsic" and "post-hoc". The former term describes models that are inherently interpretable. Decision trees and linear regressions are well studied examples of intrinsic models. Section III-B discusses intrinsic models in more detail and highlights some of their perceived challenges. Other models are built to prioritise alternative desiderata such as precision, accuracy or speed. In that circumstance, explanations are obtained "post-hoc", i.e. derived as part of an additional process after the model has delivered the outcome.

D. Scholarly Focus

Despite the ongoing debate as to the exact terminology and definitions pertaining to XAI, many scholars are undeterred in their investigations. XAI models are an increasingly popular research topic within the ML community (see Fig. 2a), and the techniques to develop, present and categorise the explanations



(a) Web of science core collection articles or proceedings papers focusing on XAI 2000 to 2021.



(b) Web of science core collection articles or proceedings papers focusing on credit card fraud 2000 to 2021.

Fig. 2. Scholarly focus for XAI and credit card fraud 2000 to 2021.

are many and varied. Likewise, investigations into credit card fraud detection are enjoying renewed attention (Fig. 2b). However, analysis of this joint population reveals just one paper published over the past 21 years which specifically investigates the application of XAI within a credit card fraud context [31].

Within that paper the authors initially propose a black box solution to distinguish between fraudulent and legitimate transactions. They subsequently acknowledge the difficulty that a human being would have in understanding the resulting output and propose an overlay to translate the results into human-understandable format. The explanation is therefore positioned as an afterthought, rather than central to the paper.

One additional paper of note explores the ability to extract generalised rules from a neural network within the domain of credit card fraud [32]. Despite no specific reference to XAI, it makes an early contribution to the field by introducing SOAR (Sparse Oracle-based Adaptive Rule extraction) which makes complex rule-sets more comprehensible by exploiting key decision boundaries.

Hence fraud – XAI cross-disciplinary research has so far lacked focus. This paper seeks to address the gap by arguing that techniques attributed to the field of XAI have the ability to accelerate a step-change in the detection of fraud in the credit card industry. This research agenda suggests ways in which XAI can improve the adoption of complex models, such as neural networks, in credit card fraud detection. Section II begins with a discussion of the credit card fraud operating landscape and key challenges which must be overcome for

successful model adoption. Section III then lists four significant focus areas which would benefit from increased scholarly attention. Finally, Section IV provides concluding remarks.

II. Fundamental Concepts and Background

A. Credit Card Operating Environment

Credit card transactions are bifurcated into Cardholder Present (CP) and Cardholder Not Present (CNP) transactions. For CP transactions the customer is physically present at the purchase point and offers a physical card to the retailer for payment. For CNP transactions the purchase is carried out remotely, for example over an e-commerce website. It is this latter scenario which will be the focus of this paper.

The speed and simplicity with which an individual can execute a credit card transaction disguises the complexity of its operating environment. There are multiple key organisations which have to interact seamlessly to deliver a smooth consumer experience. Fig. 3 shows the five key parties involved and the general timings used to execute and settle a credit card transaction.

The customer initiates the process by providing credit card payment details to the retailer in exchange for a product or service (step (1)). In real-time, the retailer requests permission from the issuer through both the acquirer and the payment card association [steps (2) to (4)] and receives an authorisation code back [steps (5) to (7)], at which point the transaction is either authorised or declined. Readers will be familiar with this entire request and response process being completed in a matter of seconds.

Following the transaction approval, the retailer receives funds from the issuer up to three days later [steps (8) to (13)]. The issuer then places the transaction on the credit card statement and issues the statement up to thirty days post transaction [step (14)]. The cardholder then has up to another thirty days to settle the bill either in full or through the use of a credit facility [step (15)].

Real-time fraud analysis focuses on confirming the authenticity of a single credit card transaction before the transaction is completed (see step (1) to step (7) in Fig. 3). The retailer, acquirer, card association and issuer all have roles to play. They perform similar types of analyses in order to ensure they are comfortable with the validity of the transaction, yet their fraud detection datasets are substantially different (Table I), enabling a multi-dimensional view of both the transaction and the context within which the transaction is being executed [33].

TABLE I. Organisations and their Credit Card Fraud Detection Datasets

Organisation	Fraud Dataset
Retailer	Previous customers and purchases
Acquirer	Transactions from all retailers who bank with them
Card Association	Transactions using the card association brand
Issuer	Transactions from all customers using issuer cards

To minimise repetition, this paper will assume the perspective of the retailer / e-commerce gateway in its discussions of fraud identification strategies and where XAI can improve the status quo. However, the strategies discussed are equally as relevant to acquirers, card associations and issuers in the real-time environment.

B. Key Challenges

FMS which enable the retailer's detection of illegitimate credit card transactions are hindered by four key challenges which will be described below. These challenges complicate the fraud identification process yet must be catered for in order to provide an operationally effective solution. Since intrinsic XAI models need to incorporate both the underlying ML algorithm and the explanation, any intrinsic XAI model will have to accommodate for all of these challenges in order to deliver an effective fraud detection explanation. In contrast, the first challenge is the only challenge relevant for a post-hoc XAI model, since its underlying AI model should operationally satisfy all key challenges.

1) Real-time analysis: Modern technology allows for the accumulation of hundreds of security features to provide information about the legitimacy of a transaction, as illustrated in Fig. 1. However, to ensure adherence to new regulations, deliver a smooth checkout experience for the customer and to minimise losses at the e-Commerce gateway those security features also need to be processed in real time. The real-time credit card transaction process illustrated by points (1) to (7) in Fig. 3 typically takes less than two seconds [34].

The foremost concern for the retailer is the provision of a seamless checkout experience for all legitimate transactions. A recent survey indicated that almost 20% of online shopping cart abandonment experiences were as a result of a "sticky" checkout experience [35]. The negative experience also reduces the likelihood of individuals visiting the store in the future thereby also impacting future sales revenue. Retailers' determination to protect their seamless checkout process is one of the key drivers behind the slow adoption of 3D Secure[1] checkouts [36].

2) Concept drift: A further advantage of real-time fraud analysis and explanation is the ability to detect emerging fraud trends and enable timely decision-making. Historically, the behaviour of fraudsters has been moderately consistent, enabling the cataloguing of fraud vectors which allows for rule-based analysis [37]. However, recent technological advances have enabled a more sophisticated and agile offender. *Concept drift* is the term used to describe this changing circumstance. Unforeseen, changing patterns in the fraud vectors results in the rules catalogue becoming either outdated or unmanageably large as more rules are added to try to keep pace with the new patterns of fraud. As a consequence, the fraud identification becomes less effective.

Examples of XAI models addressing concept drift in the domain of financial fraud are scant. However, the field could benefit from advances made in other fields. In particular, recent years have cemented the importance of addressing concept drift in the medical field of pandemic / epidemic response. In this field, authors have proposed various explainable models to support a real-time decision support model. Notably, [38] analyse Covid-19 symptomatic data using their DeepCOVID post-hoc XAI model. The result is a real-time graphical representation

[1]3D Secure (3DS) requires customers to complete an additional verification step with the card issuer when paying, for example being directed to an authentication page on their bank's website, where they enter a password associated with the card or a code sent to their phone.

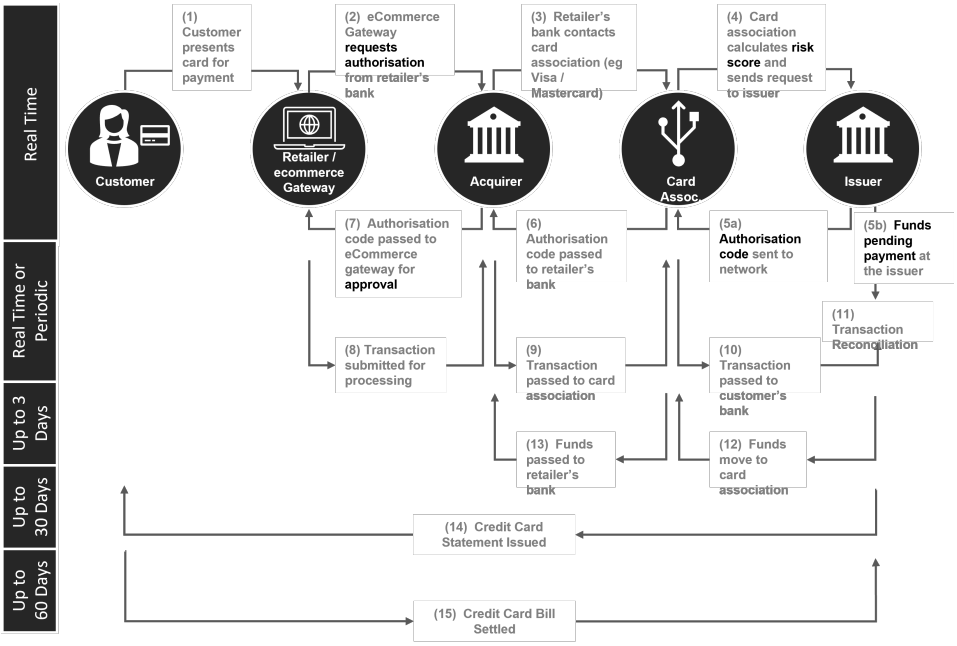


Fig. 3. CNP Credit card transaction life-cycle.

of the variables providing the most significant contributions to the prediction of Covid-19 diagnosis.

3) Minimising false positives: A model will sometimes incorrectly indicate a positive (e.g. fraudulent) result. This is known as a false positive result. Maximising the opportunities for a seamless checkout experience requires fraud investigators to minimise the occurrence of false positives when identifying fraudulent transactions.

False positives create friction in the process by either slowing down the real-time approval whilst manual assessment is required or cancelling the valid transaction altogether. In the latter case, the retailer loses both the goodwill of the customer and the value of the sale [39]. In addition to the negative experience of the customer, the occurrence of false positives creates further expense for the retailer as manual intervention is required to investigate the queried transactions. In an empirical survey of contemporaneous neural networks applied in credit card fraud detection, [14] suggested all but eight of the fifty-one (51%) ML methods in their literature population would be operationally ineffective. This is due to the high numbers of false positives in the results, requiring costly and inefficient manual oversight.

Whilst obtaining the proportion of false positive results on a test dataset helps to understand the efficacy of an AI model, it does little to provide transparency as to why incorrect predictions are being made. In contrast, XAI solutions have the advantage of being able to provide transparency to explain the reasoning for a false positive result. Saliency plots, for example, have been used by researchers to understand why an image-processing model was mistaking the picture of a husky for a wolf on a test dataset despite working with good accuracy on the training dataset [21].

4) Dealing with class imbalance: Fraudulent transactions are anomalous data points which exist within a large population of genuine transactions. Mark Nelson, Visa's Senior Vice President of Risk Products and Business Intelligence, reports that Visa operates at a fraud rate of 0.1% of transactions [40]. Having an unbalanced dataset such as this creates difficulties for training ML models with the data since many algorithms assume an equal distribution of each class. When the minority class is the most important class, as it is in fraud detection, it typically results in a poor predictive performance.

A variety of approaches are available to scholars working with class imbalance. One option is to employ a weighted loss function which penalises the misclassification of the minority class thereby boosting its performance. Other popular approaches involve either undersampling the majority class or oversampling the minority class. Undersampling involves removing a proportion of the majority class in order to create a more balanced population. This is either done through random sampling or in a more structured way, often using nearest neighbour techniques. In contrast, oversampling the minority class increases the occurrence of the minority class in the dataset. This can either be done through making copies of existing minority transactions or creating additional synthetic transactions. SMOTE (Synthetic Minority Oversampling Technique) [41] remains a popular oversampling approach which has spawned an array of derivative oversampling techniques.

C. Fraud Risk Scoring

Fig. 1 illustrates the many data points which are available to the retailer for the purposes of performing a transaction fraud assessment. These data points are employed in a number of AI profiling algorithms to be used as inputs to generate an aggregated risk score. Fig. 4 represents a drill-down into the fraud detection process for a retailer / ecommerce gateway and highlights some of the most common inputs to the risk score such as product profiling, customer profiling, geo-location profiling and analysis of spending patterns [42]. The aggregated

fraud risk score is then compared to a fraud risk threshold determined by the retailer. Scores over the threshold identify transactions which the retailer considers worthy of challenge.

1) Product profiling: When retailers list a product for sale, they make an assessment of how appealing the product is likely to be to a fraudster. Typically, fraudsters steal products which are high value and high demand and can easily be resold on a secondary market. Retailers would identify these products in their portfolio as "High-Risk" and therefore apply a high-risk score to any sale of this product. The risk score is magnified when the number of high-risk products in a single transaction increase. In a recent survey of over 1,000 retail fraud professionals, the product profile (also referred to as the "Order Content") was the key fraud indicator for 34% of survey respondents [42].

Shopping trends are in constant flux, depending upon the availability of new technology releases, changes due to seasonal trends, product availability and even media or social media influences. Consequently, it is difficult to implement an effective rule-based solution for determining high-risk products. However, ML provides retailers with an ability to adapt to new trends in a timely manner. Analogous with the discussion on *concept drift* in the paragraphs above, XAI solutions will provide real-time transparency of emerging trends enabling a retailer to understand why specific products are designated as high-risk. Graph Convolutional Networks are a popular tool in the detection of emerging trends due to their interpretability, enhanced performance and flexibility [43].

2) Customer profiling: It is important that retailers know their customer. This is not only relevant from a loyalty perspective, building a strong customer-retailer relationship, but it also provides useful knowledge in the fight against fraud. The above mentioned survey [42] identified the customer profile as the second most important fraud indicator for the survey respondents.

In respect of CNP transactions, the retailer needs to have confidence that the customer is genuine, and that they are dispatching the product to the right person at the correct address. This is much easier if they already have a prior transaction history with the customer, and far more difficult if the customer is new onto their platform. In order to establish a customer profile, they reference a number of key pieces of information which includes, but is not restricted to:

- Name and delivery address
- Usual mode of ordering (e.g. mobile or desktop)
- Frequently used IP Addresses
- Frequently used payment details
- History of returns or disputes
- Email address
- Email account history

Changes to any of the above profile factors can increase the customer's risk score.

The lowest risk for the retailer is a customer with whom they have a regular transaction history, no reported disputes, consistent behavioural patterns (e.g. mode of ordering and use of IP Address) and delivery to the same dispatch address. Any transactions with a customer in this category would be given a low-risk score for their customer profiling.

The highest risk for the retailer is a new customer. In this case they have no prior relationship data to build a customer profile. Instead, they leverage existing banking protocols alongside using other available data. At a minimum they ensure the shipping address reconciles with the billing address provided at checkout. Any deviations further increase the risk score of the customer profile. Other tactics involve ensuring the email address is not duplicated across their systems and looking at the account history of the email address.

The author in [44] explored user profiling to detect fraudulent cellular usage. Their work used an intrinsic XAI rule-learning technique to determine whether or not a customer was making a phone call from a cloned or genuine account. However, the flexibility and adaptability of clustering and classification ML algorithms have become increasingly popular in recent profiling studies. In particular, [45] demonstrated the effectiveness of the WIBL (Weighted Instance Based Learning) algorithm compared to more traditional clustering methods. WIBL improves explainability over existing clustering methods by using weighted features to indicate feature importance.

3) Geo-location profiling: The IP address also enables the retailer to access location details from where the order originates. This information is useful to the retailer in a number of ways. First, there may be certain locations which the retailer knows from prior experience have high risk of fraudulent activity. Retailers are able to use rule-based filters to exclude sales to those areas if they wish. Second, the location given by the IP address can be reconciled against the shipping and billing addresses. Although not a conclusive assessment, incongruence may indicate a higher risk of fraudulent activity.

4) Spending patterns: Finally, the retailer can also look for unusual or tell-tale spending patterns. They do this both at an individual customer level, and also holistically across their customer base. As above, this is much easier at an individual level if they have an established relationship with the customer. In that case they may be concerned with behaviours such as cancellations of orders followed by purchases of high-risk items, large volumes of high-risk products in a single transaction or unusual purchases for the customer profile, for example, an 80-year-old suddenly purchasing five flat-screen televisions. Looking across their customer base, they may see an unusual volume of high-risk products being purchased by different people but delivered to the same address, a common tactic when using "mules" to disguise fraudulent purchases.

III. Research Agenda

The sections above articulate the motivation for change and describe the challenges encountered so far in the improvement of credit card fraud detection. In particular, Section II provides details regarding the context within which an effective fraud detection solution must operate. In this section we introduce a number of key concepts and developments within XAI that the authors argue would contribute towards a step-change in its adoption for credit card fraud investigations.

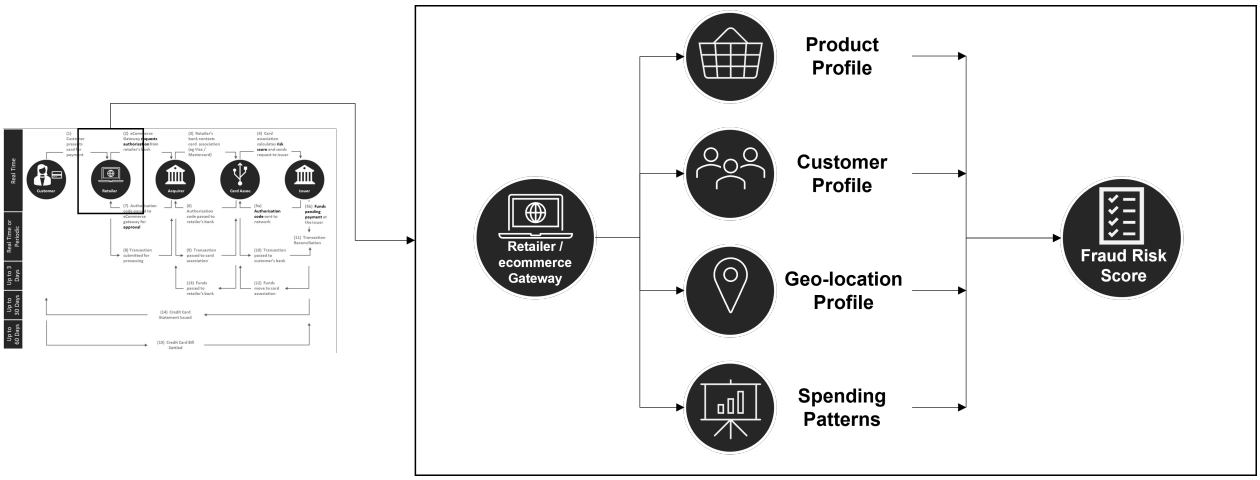


Fig. 4. Drill-down into the Retailer / e-commerce gateway process for fraud risk scoring.

To ascertain current trends in this domain, the Scopus database was employed as a primary source for gathering literature. A query identified computer science articles which were written in English and used the phrase "credit card fraud" in their key words. The resulting population of 181 articles were then filtered using reviews of the (1) title (2) abstract and (3) textual detail to focus on papers that are specifically concerned with the implementation of models in the domain of credit card fraud detection. In particular, papers which primarily focused on the generic development of models, only using a credit card fraud dataset as illustration of their techniques, were excluded from the survey. This filtering process resulted in a population of fifty-three papers, which subsequently grew to fifty-six following the addition of three papers identified by means of a snowballing technique.

Table II provides a selection of papers extracted from the full dataset. These papers consider at least two of the aforementioned operational challenges in their work. For completeness, the full table can be made available by contacting the authors of this paper. The table is complemented by Fig. 5a and 5b which summarise the full dataset.

TABLE II. LITERATURE COVERAGE OF REAL WORLD CREDIT CARD FRAUD CHALLENGES, BY PAPER - A SELECTION OF PAPERS WHICH CONSIDER AT LEAST TWO OF THE FOUR KEY CHALLENGES

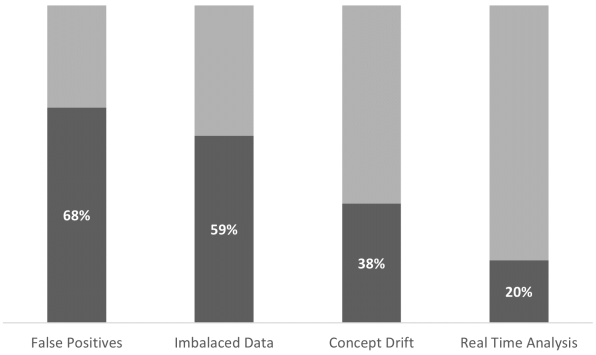
Reference	Managing False Positives	Imbalanced Data	Concept Drift	Real Time Analysis
[46]	✓	✓	✓	✓
[47]	✓	✓	✓	✓
[48]	✓	✓		✓
[49]	✓	✓		✓
[50]	✓	✓		✓
[51]	✓	✓	✓	
[52]	✓	✓	✓	
[53]	✓	✓	✓	
[54]	✓	✓	✓	
[55]	✓	✓	✓	
[56]		✓	✓	✓
[57]		✓	✓	✓
[58]		✓	✓	✓
[59]	✓		✓	✓
[60]	✓	✓		
[61]	✓	✓		
[62]	✓	✓		
[63]	✓	✓		
[64]	✓	✓		
[65]	✓	✓		
[66]	✓	✓		
[67]	✓	✓		
[68]	✓	✓		
[69]	✓	✓		
[70]	✓	✓		
[71]	✓	✓		
[72]	✓	✓		
[73]	✓		✓	
[74]	✓		✓	
[75]	✓		✓	
[76]	✓		✓	
[77]	✓		✓	
[78]		✓	✓	
[79]		✓	✓	
[80]		✓	✓	
[81]		✓		✓

A. Explanations within a Specific Context

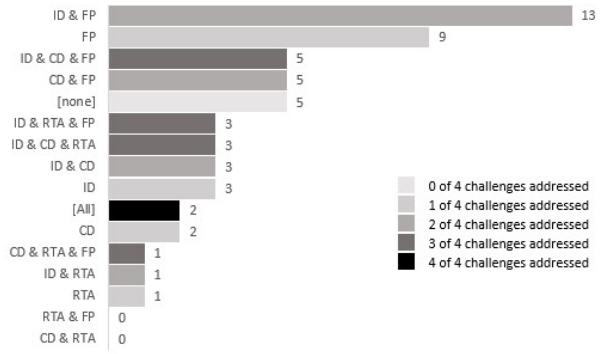
Section II-B introduces the practical constraints of real-time analysis, managing unbalanced data and concept drift and minimising false positives which need to be considered in order for a model to be operationalisable. These contextual requirements of credit card fraud detection are perhaps more complex and multi-faceted than many fields. Additionally, Section II-C highlights a variety of fraud investigation approaches which provide transparency on the root causes of the fraud detection. Unfortunately, it is rare for scholars to acknowledge or clarify the timing and perspective within which their model is intended to operate, and the field of credit card fraud detection is no exception.

Fig. 5a and 5b show the resulting analysis of the literature population, with a view to understanding the extent of its coverage of the real world challenges discussed in Section II-B. Scholars demonstrate a strong awareness for incorporating the challenges of false positives and imbalanced data in their papers (Fig. 5a) yet the majority fail to account for the difficulties brought about by the need to consider concept drift and real time analysis.

Fig. 5b particularly draws attention to the fact that the literature has so far failed to address any of these challenges, or even combinations of these challenges, in a consistent manner. In fact, five papers within the literature corpus failed to recog-



(a) Summary analysis of literature coverage of real world credit card fraud challenges.



(b) Detailed analysis of literature coverage of real world credit card fraud challenges, using the following abbreviations: Imbalanced Data (ID); False Positives (FP); Concept Drift (CD) and Real Time Analysis (RTA).

Fig. 5. Literature coverage of real world credit card fraud challenges.

nise any of the aforementioned challenges, whilst professing to deliver an implementable solution. In contrast, only two papers addressed all four of the aforementioned challenges [46], [47], with forty-two papers (75%) acknowledging two or fewer than two of them.

This analysis supports the argument that the current population of literature fails to take the contextual requirements of the credit card fraud operating environment into account when designing AI solutions. To encourage more ubiquitous adoption, scholars need to demonstrate an understanding of operational challenges and incorporate innovative solutions into their models. Authors also suggest that more rigour can be achieved by partnering with practitioners to deliver a testing strategy that mimics the operational environment [16].

Whilst scholars seeking to apply ML techniques in this domain might choose to specialise on a single challenge such as having unbalanced data or concept drift, demonstrating that the model is implementable in an operational environment (i.e. meets *usability* requirements) is key to achieving rigour and therefore ensuring more widespread acceptance [16].

Just as the contextual considerations of ML models are necessary for improving organisational adoption, the overarching consensus for XAI is that explanations are also contextual [82]. That is, in order for an agent to deliver a successful explanation, the context of the question must first be determined, and then addressed within the explanation itself. But what is meant by context, in the field of XAI, and how can it be achieved?

Whilst the literature contains a panoply of papers suggesting frameworks for the context of an explanation [18], [83]–[86], few provide an initial definition of what context means in the domain of XAI. Yet it is clear that the domain would benefit from a common vocabulary in order to move forward [15], [16]. In the absence of a normative definition, this paper proposes the following:

Context in XAI is any information needed by the explanation system to satisfy the explanation goals, trust and usability expectations of the audience.

This definition brings together four key elements of context frequently discussed in the literature. First it leverages the centrality of the audience [25], [87]–[90] since it is the audience who determines whether the explanation is a good one or not [16]. Second it captures the importance of understanding the goals of the audience, [85], [91] since it is the goals that drive the ML model design [85], [91], [92]. Third it recognises the value of ensuring trust in the explanation [17], [21], [93] since trust enables the audience to decide whether or not to have confidence in the results [21], [22]. Finally, by acknowledging the importance of usability [15], [82], [94], [95] the definition ensures that the system is more likely to be successful in an operational context [16], [94], [96].

Hence, the first recommendation for contributing towards a step-change in credit card fraud is to ensure that XAI models are designed with the context in mind. Demonstrating adherence to usability constraints such as real-time delivery, minimising false positives and supporting concept drift will encourage practitioners to see the potential rewards that XAI can bring over extant rule-based methods.An understanding of the audience goals will help scholars to develop XAI models that target practitioner desiderata and reflect the needs of real problems.

B. Increase Focus on Intrinsic Models

Section I-C introduces the concepts of intrinsic and post-hoc XAI models. For credit card fraud, the determination of fraudulent transactions can have a significant impact on a person's life and well-being. A false positive result could cause emotional distress such as shame or embarrassment as well as practical difficulties such as being unable to purchase goods. On the other hand, a false negative result fails to identify a transaction as fraudulent and results in financial consequences for the credit card holder, retailer or issuer. The serious consequences that could arise as a result of the fraud detection model forces the need for absolute trust that the explanation correctly interprets the decision-making within the model. Some authors suggest that models used for high stakes circumstances such as these should employ an intrinsic rather than post-hoc design [23].

Arguments supporting the use of intrinsic models leverage their ability to overcome the difficulties associated with black box models and their post-hoc explanations. The overriding challenge of black box models is their inherent opacity which undermines the ability of an individual to decide whether or not they can trust the model's output. Moreover, the layering of a post-hoc explanation over the black box model introduces additional trust challenges. Since the post-hoc model, by definition, cannot provide a true 100% explanation of its underlying black box model, then there must be an element of uncertainty as to whether or not the explanation is correct. An individual faced with a black box model and post-hoc explanation therefore has two trust challenges to overcome:

1) Can the model be trusted to produce an accurate output?
2) Can the explanation be trusted to be faithful to the model?

In contrast, intrinsic models are sufficiently transparent that an individual can understand not only the most influential variables in the dataset, but also how those variables interact with other variables. Furthermore, the explanation, by design, directly reflects the model machinations, thereby enabling an easier decision as to whether or not to trust the output.

Unfortunately, there is a strong bias in the extant literature against the development of intrinsic models, meaning that focus is not forthcoming. Analysis of Guidotti's [97] comprehensive survey of explainability methods identifies a slim population of 10 papers devoted to this approach, compared to 130 using post-hoc methods. Those findings are consistent with the analysis of literature conducted in this survey. Only seven of the fifty-six papers focus on the development of models which are inherently interpretable. The remaining forty-nine either propose black box models, or complex ensemble models without any attempt to explain the resulting outcomes.

There may be many reasons for this. Some authors suggest intrinsic models sacrifice accuracy for interpretability, [20], although other authors vehemently contest the notion [23]. Perhaps some scholars take pride in the complexity of black-box models ignoring the practical advantages that a transparent model would bring. Alternatively, authors designing models without a specific use-case in mind may prefer the advantages of flexibility that accompany a post-hoc, model-agnostic design.

Despite the cloak of simplicity that accompanies intrinsic models, they have many operational challenges that would benefit from scholarly focus [23]. It is not the intention of this discourse to argue a preference for intrinsic models over post-hoc techniques but to highlight that the field would benefit from increased focus and visibility. More work needs to be done to investigate the opportunities of intrinsic models in the fields of high-stakes decision-making where faithfulness to the underlying model has both an academic and moral imperative.

1) Interpretable scoring systems: A noteworthy subgroup of intrinsic models are interpretable scoring systems, used in decision-making and risk evaluation. Decision-making typically involves the careful evaluation of a number of diverse facts in order to arrive at a balanced decision. For example, medical professionals often weigh-up a number of discrete facts about a patient before suggesting or even investigating a potential medical diagnosis, or finance professionals might weigh-up a number of different factors about a client before deciding on whether or not to offer them a loan. The accumulation and evaluation of these discrete facts are synonymous with domains requiring expert judgment. Heuristics are established through experience and expertise with simple techniques such as linear regression often being used to establish relationships between these pre-defined features and their classifier.

Physiological parameter	Score 3	2	1	0	1	2	3
Respiration rate (per minute)	≤8		9–11	12–20		21–24	≥25
SpO_2 Scale 1 (%)	≤91	92–93	94–95	≥96			
SpO_2 Scale 2 (%)	≤83	84–85	86–87	88–92 ≥93 on air	93–94 on oxygen	95–96 on oxygen	≥97 on oxygen
Air or oxygen?		Oxygen		Air			
Systolic blood pressure (mmHg)	≤90	91–100	101–110	111–219			≥220
Pulse (per minute)	≤40		41–50	51–90	91–110	111–130	≥131
Consciousness				Alert			CVPU
Temperature (°C)	≤35.0		35.1–36.0	36.1–38.0	38.1–39.0	≥39.1	

NEW score	Clinical risk	Response
Aggregate score 0–4	Low	Ward-based response
Red score Score of 3 in any individual parameter	Low–medium	Urgent ward-based response*
Aggregate score 5–6	Medium	Key threshold for urgent response*
Aggregate score 7 or more	High	Urgent or emergency response**

Fig. 6. National Early Warning Scores (NEWS2) for assessing and responding to acute illness severity in the NHS [100].

These aforementioned heuristics are known as "scoring systems". Their popularity stems from the fact that decision-makers find them easy to understand and interpret [98]. Moreover, the input variables can easily be flexed to reveal the consequential impact on the predictor variable, and the model presents a common language for standardisation of reporting and comparison of results. Fig. 6 shows the scoring system mandated by NHS England for the assessment of patients presenting to, or being monitored in hospital. The lower table indicates the response that a patient should receive depending upon the medical staff's assessment of the eight key variables in the upper table.

The transparency and uniformity of this approach has the added incentive of enabling the model to be transferable to other similar circumstances, as shown by [99] who demonstrated its effectiveness at also predicting short-term mortality as a result of Covid-19. However, the challenge of employing expert-led heuristic risk scores lies in the lack of a *formal guarantee* [101] that the heuristics are the right ones.

Recent experiences in AI demonstrate that oftentimes it is beneficial to ignore human experiences and instincts, and to instead be open to new discoveries and findings. One such example is the application of reinforcement learning to playing strategy games such as chess and Go. The initial approach was to use supervised learning techniques to "teach" the AI the strategies which had been learned by generations

of experts, but this only resulted in minimal improvements upon human levels of expertise. The step-change occurred when reinforcement learning techniques allowed the AI to learn for itself without human interference [102], resulting in significantly improved performance and the discovery of some novel game-winning strategies.

With this in mind, [103] introduce RiskSLIM (Risk-calibrated Supersparse Linear Integer Model) which learns from data, rather than experience and heuristics, to deliver a risk scoring system. The model not only works efficiently but is also able to be sensitive to organisational constraints such as minimising false positive results. Meanwhile it retains interpretability and enables expert decision-makers to flex the model prior to concluding on the overall risk assessment.

There are clear parallels to be drawn between the domains of medical risk assessment and credit card fraud detection. Both domains suffer from issues with unbalanced data, need to prioritise model efficiency and minimise false positives. Moreover, they require experts to have a full understanding of the drivers influencing the risk assessment.

Section II-C describes the four key dimensions which contribute to the holistic picture of a credit card transaction. Each dimension would be expected to have a risk score of its own and then be accumulated to produce an overall transaction risk score, in a similar manner to that presented in Fig. 6 [55], [57]. From the surveyed articles, six papers proposed a risk score as a decision-making tool as opposed to a binary classification approach. These papers were also more likely to have collaborated with industrial partners in their research, demonstrating the validity of risk scores being more aligned to a real-world perspective.

The survey also shows evidence that authors are increasingly looking beyond the single dimension of transaction spending patterns. Of the fifty-six surveyed papers, twenty-two of them incorporated customer profiling within their work. However, the inclusion of product profiles and geo-location profiles remains elusive.

Unfortunately, research into risk scoring systems which learn for themselves is scant. There are very few competitors to RiskSLIM to enable a sufficiently rigorous discourse. This is despite the successful practical applications which have been achieved by contemporary authors in the medical domain. For example, [104] collaborated with the World Health Organisation (WHO) to demonstrate its effectiveness in screening for adult attention-deficit/hyperactivity disorder and more recently [105] showed its effectiveness in screening for seizures in hospitalised patients. Given the ubiquity of scoring systems in use across multiple industries, and specifically their aforementioned relevance in fraud detection, the domain would benefit from more attention from scholars. In particular, it would be beneficial to explore applications for RiskSLIM outside of the medical domain, in addition to the development of alternative models to challenge the hedgemony of RiskSLIM as a self-learning risk scoring system.

C. Measure the Faithfulness of Explanations

Assuming a researcher chooses to engage in the development of a post-hoc explanation technique, then common sense dictates that the explanation must accurately represent the reasoning process behind the model's prediction. This close relationship between the explanation and the underlying reasoning process is often referred to as faithfulness [19], [21], [106] or fidelity [97].

It has been shown that without some measure of faithfulness of an explanation, an audience may be prone to over-trust and misuse explanation tools. This circumstance was exemplified by [107] who performed a contextual inquiry and survey of data scientists using the InterpretML implementation of Generalised Additive Models (GAMs) and the SHAP Python software package. Their investigation found that some users were using the tool to rationalise suspicious observations instead of just understanding the underlying model. Others were taking the visualisations at face value instead of using them to identify issues with the dataset. Moreover, the open-source nature of both tools led individuals to trust the explanations without fully understanding them.

Efforts to measure faithfulness are nascent, with few works in publication more than five years ago. In [108] the authors used a Natural Language Processing (NLP) model called NILE (Natural language Inference over Label-specific Explanations) to demonstrate that model faithfulness and model accuracy can co-exist. Their paper used a sensitivity analysis to evidence the faithfulness of their model. Building on that concept, [109] suggest that sensitivity should be accompanied by stability to determine whether or not an explanation is faithful.

In an effort to extract consistency from the diverse literature, [106] perform a review of faithfulness works. They identify (but do not necessarily endorse) three assumptions that they say researchers are making in order to determine faithfulness:

1) **The model assumption** Two models will make the same prediction if and only if they use the same reasoning process.
2) **The prediction assumption** On similar inputs, the model makes similar decisions if and only if it provides different interpretations for similar inputs and outputs.
3) **The linearity assumption** Certain parts of the input are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent from each other.

In their discourse, [106] argue that the binary approach to determining faithfulness is fraught with difficulty since counter-examples will likely always exist. Instead, they suggest that authors should consider degrees of faithfulness to give an indication of how close an explanation is to the reasoning process of the underlying model.

Section III-B suggests there are two trust challenges that need to be overcome in order to be comfortable with the output of a black box model and its explanation. The issue of faithfulness is central to the second trust challenge. Nowhere is that trust more necessary than in high-stakes industries where the consequences of an incorrect or mis-interpreted explanation can be highly damaging. Whilst explanations may only be required under certain circumstances (for example in the event of an unexpected model outcome), there exists a moral

obligation to associate an explainable model in high stakes decision-making with some measure regarding the expected accuracy of the explanation to the ground truth.

D. Human Interaction with Explanations

In a recent call for closer integration between the Human-Computer Interaction (HCI) and ML communities, [90] cites the advantages to intelligible machine learning of leveraging the well-established human-centered research community. The cornerstone of HCI philosophy begins with understanding the needs of the audience, recognising that different audiences may have different requirements of the same system.

In the context of fraudulent transactions this paper adapts the work of [86] to suggest there are three key audiences for the explanation system:

1) The operator / executor i.e., the fraud analysts, whose role it is to determine the validity of the positive "red flag" transactions identified as potentially fraudulent.
2) The creator i.e., the technical support responsible for the internal operation of the system.
3) The examiners i.e., the senior management teams, who are focused on both the changing trends of fraud patterns and the integrity of the fraud identification process.

Critically, in the event of a transaction being deemed to be likely fraud, the cardholder should not be informed of the entire explanation without operator oversight, hence the omission of decision-subjects and data-subjects. This is because organisations within this process must take care not to advise fraudsters of the parameters in place to detect fraudulent transactions. It would therefore be incorrect to consider the cardholder as one of the parties requiring the direct explanation.

For the fraud analyst, the explanation is in place to ensure they fully understand, and agree with, the reasoning for the FMS to identify the transaction as fraudulent. They are detecting and looking for causal reasoning of an event which has already occurred. Hence their dialogue centres around local, causal explanations and the fitness of the attributes contributing towards each individual "red flag".

Technical specialists meanwhile are interested in "how", rather than "why" [110]. Their role is to ensure the system is operating effectively, for which they need transparency of the process rather than justification of an outcome. These teams will therefore look towards a causal attribution explanation in order to understand the internal workings of the explanation agent.

On the other hand, senior management teams are interested in fraud preventative measures [111]; explanations which shed light on predictive patterns. They may be searching for insight on emerging trends of fraud, in order to support future decision-making. Alternatively, they may be interested in validation that the models treat all data subjects equitably. Hence they need both local and global explanations; local to explain specific predictions and global to understand the model as a whole.

Identifying such diverse audiences and their corresponding perspectives provides a wealth of opportunities for researchers to explore a variety of targeted explanations in fraud detection. Yet the HCI community, and increasingly the ML community too, suggest that scholars should go a step further in their quest to satisfy audience desiderata. In particular, the explanation should also reflect contemporary understandings of how an audience engages with an explanation [87].

Miller's [87] seminal paper makes the case for ensuring researchers design explanations with an appreciation of human cognition in mind. It builds upon an earlier paper [82] which articulates the importance of comprehension in order to ensure the explanation is useful to the intended user in a practical setting. This view is widely held [16], [17], [93], [112].

Cognitive scientists claim that prior knowledge is widely recognised to have a profound influence on understanding new concepts [113]. Hence for an effective explanation, the explainer must first understand the audience's initial level of existing knowledge. Any subsequent new information then builds upon that baseline [114], [115], incrementally constructing a bridge to a new knowledge state. This individualised layering of new knowledge on old becomes synonymous with explanation as a dialogue, wherein the audience repeatedly questions the explanation agent until a point of understanding is reached.

However, building knowledge in this way only allows for the audience to learn from the explanation agent. In fields such as fraud detection, there are also likely to be instances where experts have more knowledge than the explainer, resulting in them outperforming the system-generated explanation [116]. In this circumstance, explanations should therefore be a two-way concept. Whilst we look to XAI to communicate unknown patterns and influences extracted from the prescribed data, the expert audience adds breadth, supplementing the explanation with their own peripheral knowledge and undocumented experiences. Hence in expert systems, designing explanation with an interactive dialogue in mind allows for the development of a "learning loop", which ultimately enhances the performance of both the XAI agent and the audience [116].

IV. Conclusion

Credit card fraud is widely acknowledged as a key contributor to the persistence of organised crime in the European Union. Moreover, the recent Covid-19 pandemic has accelerated the switch to digital payments and revealed the potential of a future cashless global society. As the use of payment cards continues to overtake the use of cash in our economy, the ability of payments providers to reduce the value and volume of fraudulent transactions becomes ever more crucial.

Regulators acknowledge this danger and are working to introduce increasingly stringent legislation to counteract the trend. In particular, they are leveraging the vast quantities of data available in our modern society to encourage more effective financial defences. As part of the PSD2 regulation, SCA has recently been enforced in Europe and the United Kingdom. SCA mandates real-time data analysis and the introduction of authentication enrichment data, both of which combine with recent developments in open banking and payment technologies to create an urgent need for change in the detection of fraudulent credit card transactions.

The overarching consensus is that established rule-based fraud detection methodologies are no longer scalable to the extent that modern society needs them to be. Moreover, they struggle to provide the flexibility or agility to adapt to either the rapidly changing operating environment or dynamic modus operandi of modern fraudsters. ML models have the ability to provide a solution to these challenges, yet their opacity has impeded their adoption in this domain.

In response, this paper argues for more researchers to engage with investigations into the use of XAI techniques for credit card fraud detection. It contributes to the discourse in three key ways:

1) It sheds light on recent regulatory changes which are pivotal in driving the adoption of new ML techniques.
2) It examines the operating environment pertaining to CNP credit card transactions, an understanding of which is crucial for the ability to operationalise ML solutions.
3) Using a survey of contemporary literature, it sets out a research agenda, arguing that further work would contribute towards a step-change in the adoption of ML into this industry.

The research agenda first suggests that the current literature fails to consistently accommodate the key contextual challenges of real-time analysis, concept drift, minimising false positives and dealing with class imbalance. These omissions lead to solutions which are not operationalisable, thereby undermining the relevancy of the work. Incorporating context fully into an XAI solution would support more wider adoption, yet recent papers in XAI have struggled to articulate the full meaning of context in this field. The first agenda point therefore provides a novel definition of the term ”context” in relation to XAI and goes on to suggest that researchers should always design XAI models with context in mind.

Second, it recommends that more work should be done to examine the utility of intrinsic models and in particular focus on the under-researched area of self-learning risk scoring systems. Contemporary literature generally demonstrates a bias towards the development of post-hoc rather than intrinsic models. A popular argument suggests this is because black box models are more accurate than their interpretable counterparts. Yet this statement remains controversial for some authors, especially in light of the need for trust and transparency in high-stakes decision-making. Increased attention from scholars will help to progress this debate and may help to challenge the hegemony of incumbent risk scoring systems.

Third, it recognises that authors should consider implementing measures of faithfulness to give an indication of how close an explanation is to the reasoning process of the underlying model, and thereby help to establish trust in the explanation. Previous authors have demonstrated the tendency for an audience to over-trust and mis-use explanation tools without some measure of faithfulness. Its inclusion as an evaluation tool is particularly pertinent in the field of high-stakes decision-making such as fraud detection, where the consequences of an incorrect decision can be damaging to multiple parties.

Finally, it suggests recognising the value of human expert knowledge in this domain and incorporating an ability to provide a “learning loop” which ultimately enhances the performance of both the XAI agent and the audience. The current corpus of literature recommends that explanations should not only be designed with the audience in mind, but also recognise the nuances of human cognition in order to deliver an explanation that is useful to the intended user in a practical setting. The explanation should subsequently evolve into a dialogue, wherein the audience can repeatedly question the explanation agent until a point of understanding is reached, and likewise contribute expert knowledge into the model to enhance mutual understanding.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable feedback.

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

REFERENCES

[1] Europol, “Serious and Organised Crime Threat Assessment,” 2021.

[2] UK Finance, “Fraud - The Facts, 2021,” 2021.

[3] S. Nilson, “The Nilson Report: Card Fraud Losses Reach $27.85 Billion,” 2019.

[4] European Central Bank, “The revised payment services directive (psd2) and the transition to stronger payments security,” https://www.ecb.europa.eu/paym/intro/mip-online/2018/html/1803_revisedpsd.en.html Accessed: July 27, 2022, 2018.

[5] J. Allen, “What is authentication enrichment and why should you do it?” https://www.ravelin.com/blog/what-is-authentication-enrichment-and-why-should-you-do-it Accessed: November 30, 2021, 2020.

[6] Visa, “New and improved 3-d secure,” 2019. [Online]. Available: https://usa.visa.com/content/dam/VCOM/global/visa-everywhere/documents/visa-3d-secure-2-program-infographic.pdf

[7] T. Cray, “How will sca adoption impact chargebacks?” https://www.ukfinance.org.uk/news-and-insight/blogs/how-will-sca-adoption-impact-chargebacks, 2021.

[8] K. Dowd, “The war on cash is about much more than cash,” *Economic Affairs*, vol. 39, no. 3, pp. 391–399, 2019.

[9] UK Finance, “UK Payments Market Summary 2021,” 2021.

[10] Open Banking Limited, “Open Banking Impact Report, 2022,” https://openbanking.foleon.com/live-publications/the-open-banking-impact-report-june-2022/ Accessed 28 July, 2022, 2022.

[11] The International Bank for Reconstruction and Development, “Payment systems worldwide - a snapshot,” https://documents1.worldbank.org/curated/en/115211594375402373/pdf/A-Snapshot.pdf, 2020.

[12] G. Dvorsky, “Hackers have already started to weaponise artificial intelligence,” https://gizmodo.com/hackers-have-already-started-to-weaponize-artificial-in-1797688425, 2017.

[13] CIFAS, “Fraudscape 2020,” https://www.fraudscape.co.uk/, 2020.

[14] N. F. Ryman-Tubb, P. Krause, and W. Garn, “How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark,” *Engineering Applications of Artificial Intelligence*, vol. 76, pp. 130–157, 2018.

[15] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[16] F. Doshi-Velez and B. Kim, “A roadmap for a rigorous science of interpretability,” *arXiv preprint arXiv:1702.08608*, vol. 2, p. 1, 2017.

[17] O. Biran and C. Cotton, “Explanation and justification in machine learning: A survey,” in *IJCAI-17 workshop on explainable AI (XAI)*, vol. 8, 2017, pp. 8–13.

[18] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019.

[19] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[20] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[22] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[24] P. Hall, N. Gill, and N. Schmidt, "Proposed guidelines for the responsible use of explainable machine learning," *arXiv preprint arXiv:1906.03533*, 2019.

[25] S. Atakishiyev, H. Babiker, N. Farruque, R. Goebel, M. Kima, M. H. Motallebi, J. Rabelo, T. Syed, and O. R. Zaïane, "A multi-component framework for the analysis and design of explainable artificial intelligence," *arXiv preprint arXiv:2005.01908*, 2020.

[26] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[27] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3-4, pp. 1–45, 2021.

[28] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[30] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *arXiv preprint arXiv:2102.13076*, 2021.

[31] D. Sinanc, U. Demirezen, Ş. Sağıroğlu *et al.*, *Explainable Credit Card Fraud Detection with Image Conversion*. Ediciones Universidad de Salamanca (España), 2021.

[32] N. F. Ryman-Tubb and A. d. Garcez, "Soar—sparse oracle-based adaptive rule extraction: knowledge extraction from large-scale datasets to detect credit card fraud," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2010, pp. 1–9.

[33] Gartner, "Market guide for online fraud detection," 2021. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-27FWBFD0&ct=210915&st=sbf

[34] Fisglobal, "What is credit card processing?" https://www.fisglobal.com/en-gb/insights/merchant-solutions-worldpay/article/what-is-credit-card-processing Accessed: March 23, 2022, 2019.

[35] Baymard Institute, "Main reasons why consumers in the united states abandoned their orders during the checkout process in 2021." https://www-statista-com.surrey.idm.oclc.org/statistics/1228452/reasons-for-abandonments-during-checkout-united-states/ Statista Inc.. Accessed: November 30, 2021., 2021.

[36] 3dSecure2, "Why was 3-d secure 1.0 not successful in some countries?" https://3dsecure2.com/blog/why-was-3-d-secure-1.0-not-successful-in-some-countries/ Accessed: November 30, 2021, 2019.

[37] A. Shen, R. Tong, and Y. Deng, "Application of classification models on credit card fraud detection," in *2007 International conference on service systems and service management*. IEEE, 2007, pp. 1–4.

[38] A. Rodriguez, A. Tabassum, J. Cui, J. Xie, J. Ho, P. Agarwal, B. Adhikari, and B. A. Prakash, "Deepcovid: An operational deep learning-driven framework for explainable real-time covid-19 forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 15 393–15 400.

[39] Ethoca, "Solving the cnp false decline puzzle: Collaboration is key," https://hs.ethoca.com/solving-the-cnp-false-decline-puzzle-collaboration-is-key Accessed: July 28, 2022, 2017.

[40] M. Nelson, "Outsmarting fraudsters with advanced analytics," https://usa.visa.com/visa-everywhere/security/outsmarting-fraudsters-with-advanced-analytics.html Accessed: March 23, 2022, no date.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[42] Ravelin, "Retail eCommerce Fraud and Payments Survey," https://pages.ravelin.com/retail-fraud-payments-report Accessed 28 July, 2022, 2021.

[43] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[44] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data mining and knowledge discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[45] A. Cufoglu, "User profiling-a short review," *International Journal of Computer Applications*, vol. 108, no. 3, 2014.

[46] I. Sadgali, N. Sael, and F. Benabbou, "Adaptive model for credit card fraud detection," 2020.

[47] R. Van Belle, B. Baesens, and J. De Weerdt, "Catchm: A novel network-based credit card fraud detection method using node representation learning," *Decision Support Systems*, p. 113866, 2022.

[48] V. Van Vlasselaer, C. Bravo, O. Caelen, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions," *Decision Support Systems*, vol. 75, pp. 38–48, 2015.

[49] M. Arya and H. Sastry G, "Deal–'deep ensemble algorithm'framework for credit card fraud detection in real-time data stream with google tensorflow," *Smart Science*, vol. 8, no. 2, pp. 71–83, 2020.

[50] A. F. Ghahfarokhi, T. Mansouri, M. R. S. Moghaddam, N. Bahrambeik, R. Yavari, and M. F. Sani, "Credit card fraud detection using asexual reproduction optimization," *Kybernetes*, 2021.

[51] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: a realistic modeling and a novel learning strategy," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3784–3797, 2017.

[52] S. M. Darwish, "A bio-inspired credit card fraud detection model based on user behavior analysis suitable for business management in electronic banking," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 11, pp. 4873–4887, 2020.

[53] J. Forough and S. Momtazi, "Sequential credit card fraud detection: A joint deep neural network and probabilistic graphical model approach," *Expert Systems*, vol. 39, no. 1, p. e12795, 2022.

[54] V. Plakandaras, P. Gogas, T. Papadimitriou, and I. Tsamardinos, "Credit card fraud detection with automated machine learning systems," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2086354, 2022.

[55] J. N. Dharwa and A. R. Patel, "A data mining with hybrid approach based transaction risk score generation model (trsgm) for fraud detection of online financial transaction," *International Journal of Computer Applications*, vol. 16, no. 1, pp. 18–25, 2011.

[56] F. Carcillo, A. Dal Pozzolo, Y.-A. Le Borgne, O. Caelen, Y. Mazzer, and G. Bontempi, "Scarff: a scalable framework for streaming credit card fraud detection with spark," *Information fusion*, vol. 41, pp. 182–194, 2018.

[57] I. Mekterović, M. Karan, D. Pintar, and L. Brkić, "Credit card fraud detection in card-not-present transactions: Where to invest?" *Applied Sciences*, vol. 11, no. 15, p. 6766, 2021.

[58] I. Sadgali, N. Sael, and F. Benabbou, "Human behavior scoring in credit card fraud detection," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 3, p. 698, 2021.

[59] S. Ounacer, H. A. El Bour, Y. Oubrahim, M. Y. Ghoumari, and M. Azzouazi, "Using isolation forest in anomaly detection: the case of credit card transactions," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 6, no. 2, pp. 394–400, 2018.

[60] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5916–5923, 2013.

[61] E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 057–13 063, 2011.

[62] S. Jha, M. Guillen, and J. C. Westland, "Employing transaction aggregation strategy to detect credit card fraud," *Expert systems with applications*, vol. 39, no. 16, pp. 12 650–12 657, 2012.

[63] A. G. de Sá, A. C. Pereira, and G. L. Pappa, "A customized classification algorithm for credit card fraud detection," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 21–29, 2018.

[64] J. A. Gómez, J. Arévalo, R. Paredes, and J. Nin, "End-to-end neural network architecture for fraud scoring in card payments," *Pattern Recognition Letters*, vol. 105, pp. 175–181, 2018.

[65] E. Kim, J. Lee, H. Shin, H. Yang, S. Cho, S.-k. Nam, Y. Song, J.-a. Yoon, and J.-i. Kim, "Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning," *Expert Systems with Applications*, vol. 128, pp. 214–224, 2019.

[66] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on svm-recursive feature elimination and hyper-parameters optimization," *Journal of Information Security and Applications*, vol. 55, p. 102596, 2020.

[67] S. Akila and U. S. Reddy, "Cost-sensitive risk induced bayesian inference bagging (ribib) for credit card fraud detection," *Journal of computational science*, vol. 27, pp. 247–254, 2018.

[68] N. K. Trivedi, S. Simaiya, U. K. Lilhore, and S. K. Sharma, "An efficient credit card fraud detection model based on machine learning methods," *International Journal of Advanced Science and Technology*, vol. 29, no. 5, pp. 3414–3424, 2020.

[69] C. Wang and D. Han, "Credit card fraud forecasting model based on clustering analysis and integrated support vector machine," *Cluster Computing*, vol. 22, no. 6, pp. 13 861–13 866, 2019.

[70] M. Rezapour, "Anomaly detection using unsupervised methods: credit card fraud case study," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 11, 2019.

[71] E. Esenogho, I. D. Mienye, T. G. Swart, K. Aruleba, and G. Obaido, "A neural network ensemble with feature engineering for improved credit card fraud detection," *IEEE Access*, vol. 10, pp. 16 400–16 407, 2022.

[72] J. F. Roseline, G. Naidu, V. S. Pandi, S. A. alias Rajasree, and N. Mageswari, "Autonomous credit card fraud detection using machine learning approach," *Computers and Electrical Engineering*, vol. 102, p. 108132, 2022.

[73] A. Pumsirirat and Y. Liu, "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine," *International Journal of advanced computer science and applications*, vol. 9, no. 1, 2018.

[74] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Information sciences*, vol. 557, pp. 317–331, 2021.

[75] L. Zheng, G. Liu, C. Yan, and C. Jiang, "Transaction fraud detection based on total order relation and behavior diversity," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 796–806, 2018.

[76] Y. Xie, G. Liu, C. Yan, C. Jiang, and M. Zhou, "Time-aware attention-based gated network for credit card fraud detection by extracting transactional behaviors," *IEEE Transactions on Computational Social Systems*, 2022.

[77] M. Krivko, "A hybrid model for plastic card fraud detection systems," *Expert Systems with Applications*, vol. 37, no. 8, pp. 6070–6076, 2010.

[78] F. Carcillo, Y.-A. Le Borgne, O. Caelen, and G. Bontempi, "Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization," *International Journal of Data Science and Analytics*, vol. 5, no. 4, pp. 285–300, 2018.

[79] I. Sadgali, N. Sael, and F. Benabbou, "Bidirectional gated recurrent unit for improving classification in credit card fraud detection," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 21, no. 3, pp. 1704–1712, 2021.

[80] H. Z. Alenzi and N. O. Aljehane, "Fraud detection in credit cards using logistic regression," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020.

[81] A. Q. Zainab K., Dhanda N., "Adoca: A novel technique to defraud credit card using an optimized catboost algorithm," *Journal of Theoretical and Applied Information Technology*, 2022.

[82] T. Miller, P. Howe, and L. Sonenberg, "Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

[83] F. Sørmo, J. Cassens, and A. Aamodt, "Explanation in case-based reasoning–perspectives and goals," *Artificial Intelligence Review*, vol. 24, no. 2, pp. 109–143, 2005.

[84] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[85] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović *et al.*, "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[86] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to whom? a role-based model for analyzing interpretable machine learning systems," *arXiv preprint arXiv:1806.07552*, 2018.

[87] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[88] M. Ribera and A. Lapedriza, "Can we do better explanations? a proposal of user-centered explainable ai." in *IUI Workshops*, vol. 2327, 2019, p. 38.

[89] S. Rüping, "Learning interpretable models," 10 2006.

[90] J. W. Vaughan and H. Wallach, "A human-centered agenda for intelligible machine learning," *Machines We Trust: Getting Along with Artificial Intelligence*, 2020.

[91] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, ""Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations," *arXiv preprint arXiv:1904.12991*, 2019.

[92] S. R. Haynes, M. A. Cohen, and F. E. Ritter, "Designs for explaining intelligent agents," *International Journal of Human-Computer Studies*, vol. 67, no. 1, pp. 90–110, 2009.

[93] R. R. Hoffman, G. Klein, and S. T. Mueller, "Explaining explanation for "explainable ai"," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 2018, pp. 197–201.

[94] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–18.

[95] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? the effects of mental model soundness on personalizing an intelligent agent," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1–10.

[96] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 126–137.

[97] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[98] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Machine Learning*, vol. 102, no. 3, pp. 349–391, 2016.

[99] L. J. Scott, A. Tavaré, E. M. Hill, L. Jordan, M. Juniper, S. Srivastava, E. Redfern, H. Little, and A. Pullyblank, "Prognostic value of national early warning scores (news2) and component physiology in hospitalised patients with covid-19: a multicentre study," *Emergency Medicine Journal*, 2022.

[100] Royal College of Physicians, "National early warning score (news) 2: Standardising the assessment of acute-illness severity in the nhs," 2017. [Online]. Available: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2

[101] B. Ustun and C. Rudin, "Learning optimized risk scores." *J. Mach. Learn. Res.*, vol. 20, no. 150, pp. 1–75, 2019.

[102] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[103] C. Rudin and B. Ustun, "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice," *Interfaces*, vol. 48, no. 5, pp. 449–466, 2018.

[104] B. Ustun, L. A. Adler, C. Rudin, S. V. Faraone, T. J. Spencer, P. Berglund, M. J. Gruber, and R. C. Kessler, "The world health organization adult attention-deficit/hyperactivity disorder self-report screening scale for dsm-5," *Jama psychiatry*, vol. 74, no. 5, pp. 520–526, 2017.

[105] A. F. Struck, A. A. Rodriguez-Ruiz, G. Osman, E. J. Gilmore, H. A. Haider, M. B. Dhakar, M. Schrettner, J. W. Lee, N. Gaspard, L. J. Hirsch *et al.*, "Comparison of machine learning models for seizure prediction in hospitalized patients," *Annals of clinical and translational neurology*, vol. 6, no. 7, pp. 1239–1247, 2019.

[106] A. Jacovi and Y. Goldberg, "Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?" *arXiv preprint arXiv:2004.03685*, 2020.

[107] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–14.

[108] S. Kumar and P. Talukdar, "Nile: Natural language inference with faithful natural language explanations," *arXiv preprint arXiv:2005.12116*, 2020.

[109] F. Yin, Z. Shi, C.-J. Hsieh, and K.-W. Chang, "On the faithfulness measurements for model interpretations," *arXiv preprint arXiv:2104.08782*, 2021.

[110] M. R. Wick, P. Dutta, T. Wineinger, and J. Conner, "Reconstructive explanation: A case study in integral calculus," *Expert Systems with Applications*, vol. 8, no. 4, pp. 463–473, 1995.

[111] E. Gianotti and E. D. da Silva, "Strategic management of credit card fraud: stakeholder mapping of a card issuer," *Journal of Financial Crime*, 2021.

[112] K. Sokol and P. Flach, "Explainability fact sheets: a framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 56–67.

[113] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[114] G. Carenini and J. D. Moore, "Generating explanations in context," in *Proceedings of the 1st international conference on Intelligent user interfaces*, 1993, pp. 175–182.

[115] P. Brézillon, "Context in problem solving: A survey," *The Knowledge Engineering Review*, vol. 14, no. 1, pp. 47–80, 1999.

[116] G. Klein, B. Shneiderman, R. R. Hoffman, and K. M. Ford, "Why expertise matters: A response to the challenges," *Ieee Intelligent Systems*, vol. 32, no. 6, pp. 67–73, 2017.