Towards Design Principles for User-Centric Explainable AI in Fraud Detection*

Douglas Cirqueira[1,3][0000−0002−1283−0453], Markus Helfert[2,3][0000−0001−6546−6408], and Marija Bezbradica[1,3][0000−0001−9366−5113]

[1] School of Computing, Dublin City University, Dublin, Ireland
douglas.darochacirqueira2@mail.dcu.ie
[2] Innovation Value Institute, Maynooth University, Maynooth, Ireland
[3] Lero - the Science Foundation Ireland Research Centre for Software, Dublin City University, Dublin, Ireland

Abstract. Experts rely on fraud detection and decision support systems to analyze fraud cases, a growing problem in digital retailing and banking. With the advent of Artificial Intelligence (AI) for decision support, those experts face the black-box problem and lack trust in AI predictions for fraud. Such an issue has been tackled by employing Explainable AI (XAI) to provide experts with explained AI predictions through various explanation methods. However, fraud detection studies supported by XAI lack a user-centric perspective and discussion on how principles are deployed, both important requirements for experts to choose an appropriate explanation method. On the other hand, recent research in Information Systems (IS) and Human-Computer Interaction highlights the need for understanding user requirements to develop tailored design principles for decision support systems. In this research, we adopt a design science research methodology and IS theoretical lens to develop and evaluate design principles, which align fraud expert's tasks with explanation methods for Explainable AI decision support. We evaluate the utility of these principles using an information quality framework to interview experts in banking fraud, plus a simulation. The results show that the principles are an useful tool for designing decision support systems for fraud detection with embedded user-centric Explainable AI.

Keywords: Explainable AI · Fraud Detection · Decision Support Systems · Artificial Intelligence · Design Principles · HCI · Human-AI Interaction · Human-Centered AI.

1 Introduction

Digital platforms are convenient for customers in online retail and banking as they allow quick transactions and a choice between multiple E-commerce and

* This research was supported by the European Union Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 765395; and supported, in part, by Science Foundation Ireland grant 13/RC/2094$_P$2

offline channels [8]. However, the convenience is followed by increased fraud cases [30, 33]. Companies increasingly use Artificial Intelligence for decision support and fraud detection systems to automatically classify and alert experts of cases where revision is needed. However, fraud experts are not knowledgeable about AI's inner workings and face the black-box problem [35].

Ideally, fraud experts should be able to trust AI partners in such scenarios, as due to the complexity of their work, if decisions those experts make are wrong, this might cause harm and financial loss for institutions and customers [62]. Indeed, trust in AI is recognized as essential by agencies in Europe and worldwide, which develop guidelines for trustworthy and responsible AI for businesses and society [20]. In the meantime, Explainable AI (XAI) research was developed to optimize a diversity of explanation methods (EM) and to enable user's understanding of AI predictions for decision support [2].

Concerning fraud detection specifically, previous studies supported by XAI provide experts with explanations [12], but a user-centric perspective is lacking. Requirements are not elucidated before the deployment of explanations, and they also lack prescriptive principles for the alignment between EM and fraud experts requirements [3, 61]. Therefore, this can cause the lack of trust from fraud experts in AI predictions [11]. Since XAI research has a major focus on optimizing EM and AI models, and previous XAI studies for fraud lack a user-centric perspective, in this work, we develop design principles (DP) to prescribe the alignment of fraud expert's tasks to EM for enabling Explainable AI decision support (XAIDSS) in fraud detection. We adopt an Information Systems (IS) lens and follow a design science research methodology to develop and evaluate DP through multiple iterations with fraud experts. We evaluate the utility of our principles through an information quality framework with experts in banking fraud via interviews and also via a simulation on a real transaction fraud dataset.

This paper is structured as follows: Section 2 presents the theoretical foundation and related work; Section 3 highlights the research methodology; Section 4 depicts the developed design principles as the core contribution of this study; Section 5 provides the evaluation design and results for design principles; Section 6 discusses results and implications followed by Section 7 with conclusions.

2 Theoretical Foundation and Related Work

Shopping transactions via digital retail platforms are constantly increasing [10], which opens opportunities for fraudsters to act. This work focuses on transaction fraud cases, which occur when a customer card or online account balance is used without a customer's consent to perform a transaction, for instance, in retail or via bank transfers[68, 45]. In order to review and make decisions on fraud cases classified by AI models, experts rely on fraud detection and decision support systems [14].

Explainable AI research applies and develops a diversity of explanation methods to explain AI predictions in particular applications [4]. There is no consensus on how to classify those methods. Currently, well-regarded surveys classify EM

based on their dimensions of scope (local and global explanations), target (to explain the data, model, or features), and explanation type [67, 5, 18, 4, 2, 38, 54, 39, 41]. Therefore, each method provides the users with an explanation type for decision support, and it analyses particular aspects of AI predictions and models. Recent research in XAI advocates for the importance of user-centric XAI, relying on Human-Computer Interaction (HCI) and interdisciplinary social sciences [1]. The researchers focus on the design of EM following user requirements to enhance decision support and trust in AI [58].

Design principles are well researched in Information Systems and HCI disciplines. They enable prescriptive knowledge on establishing and designing decision support systems to aid user practices [52, 15, 53]. Design principles have also been developed to guide the design of user interactions with AI, support explanatory data analysis, and debugging AI models [6, 65]. Indeed, [19] highlights the need for the development of principles for informed predictions and interactions between users and AI predictions, which can then be mitigated with HCI research support.

Previous studies employing XAI and EM for fraud detection explore the effects and performance of explanations in fraud expert's work. In [31], authors provide a service architecture for security experts with explanations, aiming to introduce more context for the outlier score given anomalous records of network flows. In [12], authors provide experts with Shapley Additive Explanations (SHAP) [34] for why particular warranty claims are marked as anomalies by a machine learning (ML) model. In [61], the authors also work with SHAP explanations for fraud cases, and they observe through experiments that explanations positively impact the decision-making for fraud alerts. The same authors in [60] go further and develop case-based explanations with visualizations for similar fraudulent cases in banking. In [7] authors develop an EM to explain the importance of current and past events and features on sequential data, enabling experts with a temporal perspective on explanations for recurrent ML models such as RNNs and LSTMs [25]. They evaluate their model through experiments and simulations regarding the relevance of features, events, and efficiency for providing explanations that can support debugging AI models for fraud detection. In [26], authors evaluate popular EM and tools focusing on feature importance explanations and their impact on user's accuracy and time to make decisions. However, the assessment of user requirements and exploration of different explanation types are left for future studies.

Existing literature employing EM in fraud detection has not tackled fraud expert's requirements, making it challenging to align explanations and these requirements to establish trustworthy XAIDSS. Furthermore, as used by IS, AI, and HCI studies, design principles are lacking in fraud scenarios for providing prescriptive knowledge on how to deploy explanations for fraud expert's decision support with XAI. This study aims to address this gap by developing and evaluating principles for user-centric XAI and enable XAIDSS in fraud detection.

3 Research Methodology

We follow design science research (DSR) [42], an IS research methodology focused on interactive developing and evaluating artefacts for solving a practitioner's problem and bringing research contributions. We start by identifying the research problem through a literature review presented in the theoretical background section and discussions with experts in fraud detection within a European bank. We identify the problem of fraud experts' lack of trust towards AI predictions due to insufficient alignment between their tasks and explanations to review fraud cases. The next step is to establish the research objective, which is to align expert's tasks to explanations according to their needs for XAIDSS. Adopting HCI and IS theoretical lenses, this study develops an artefact, a set of design principles to guide such alignment.

For the design and development phase of design science research, we first need to establish the kernel theory governing the artefact development process for solving the identified problem [22, 56]. Given the poor alignment between fraud expert's tasks and explanations for XAIDSS, our problem relates to an expert's decision-making process when reviewing fraud cases. We then establish the kernel theory to develop the artefact based on two main sets of constructs: 1) fraud expert tasks and 2) design features revealing meta-design principles of EM, which facilitate experts to perform their tasks. Those constructs compose the design knowledge to develop our principles.

Regarding the first set of constructs for artefact development, we rely on our previous study results to elicit 13 fraud expert's tasks when analyzing suspicious fraud cases [51]. In the referred study, we adopt expert interviews with a scenario-based method, and a systematic literature review [11]. Scenario-based elicitation facilitates an HCI and problem-centered perspective to identify stakeholder requirements, goals, tasks and knowledge to develop decision support systems [64]. In the current study, and guided by [51], we extend the previous work by grouping the tasks into requirements. Those requirements should reflect experts' actions and goals when analyzing fraud cases supported by a decision support system. For instance, experts compare, cluster and contrast cases, so those tasks are grouped within a requirement established as similarity and previous pattern matching. We employ the tasks and requirements in our design principles.

In relation to the second set of constructs for artefact development, we perform a systematic literature review following [59] to identify design features of EM, which enable fraud experts to perform their tasks and understand AI predictions. Those features help identify meta-design principles, which are post-instantiated principles found on a class of artefacts [57, 13]. In our case, the class of artefacts is an EM within the XAI literature. Then, we start defining the main research question as "What are the design features and meta-design principles of explanation methods for their user's decision support?". To answer the research question through relevant literature, we define a search query as "("explainable ai" OR "explainable artificial intelligence" OR "interpretable machine learning")" and query the databases of Scopus, ACM Digital Library, IEEEXplore,

Science Direct, Springer Link, and arXiv. The database search is performed until July 2020. We obtain 2507 studies and read their abstract, introduction and conclusion to select papers that discuss explanation methods. That process gives 372 papers, which are thoroughly read following inclusion and exclusion criteria [4]. We then obtain 140 papers, for which backward and forward search gives us additional 51 papers. We also include papers from scientific events focused on XAI research following our inclusion criteria and extending our coverage, which adds 43 papers to our pool. Therefore, the total of selected papers was 177. Those are analyzed to extracting design features of explanation methods. The complete data for our systematic literature review is available externally[5].

To analyze the systematic literature review results and elicit design features of EM, we adopt a classification for explanations following our theoretical foundation [67, 5, 18, 4, 2, 38, 54, 39, 41]. Given that our DP should align EM to fraud experts tasks to establish XAIDSS, this foundation guides our elicitation of EM design features. Therefore, we focus our analysis on every paper selected based on how the EM employed enables decision support based on the scope, target, and explanation type they provide. We adopt a Concept Centric Matrix [28] to structure the findings and design features of EM from every selected paper. Finally, we establish the DP reflecting the instantiated meta-design principles and design features aligned to fraud expert's tasks.

In the first iteration of our design science research, the design principles are discussed with fraud experts in one major bank and our project partner. We ask the experts about their perceptions regarding the principle's correctness, understandability, and comprehensibility according to their tasks and requirements for analyzing fraud cases. From that iteration, we obtain issues with the terminology adopted to describe each principle, which is deemed ambiguous to describe requirements and their grouping towards design features of EM. After analyzing our kernel knowledge and references discussing templates for DP development, we incorporate the guidelines with the anatomy of DP by [23] bringing principle's aim, implementer, context, mechanism, and rationale based on expert's requirements. The authors developed a systematic template for clear delivery of DP based on rigorous analysis of IS literature, including the development of intelligent decision support systems, which all relate to our context. The next round of iterations was focused on evaluating the tasks, requirements, design features, and DP presented in Section 4.

[4] Inclusion criteria (IC) for papers: IC1-"Paper focuses on providing explanation methods for supporting understanding of AI predictions or user decision-making in user experiments"; IC2-"Paper presents explanation methods that possess an interface for providing explanations as outputs"; IC3-"Paper focuses on discussing classifications or types of explanation methods". The exclusion criteria (EC) is elaborated as: EC1-"Paper is not written in English"; EC2-"Paper is not a journal, conference, workshop article, or Ph.D. thesis"; EC3-"The paper is not fully available".

[5] https://github.com/dougcirqueira/hcii-design-principles-user-centric-explainable-ai-fraud-detection/tree/main/resources/systematic_literature_review

4 Design Principles

Figure 1 presents developed design principles from our study. It starts with the 13 fraud expert's tasks when analyzing fraud cases (T1 - T13). Those tasks are grouped, and 7 requirements are established (R1 - R7), as described in our methodology in Section 3. Table 1 presents the descriptions for requirements.

Table 1. Grouped fraud expert tasks and requirements with descriptions

Tasks	Requirements	Description
T1, T2	**R1: System Confidence and Limitations**	To provide predicted fraud cases based on a probability ranking, and the limitations for classifying cases based on the current AI model performance on training and validation datasets
T3, T5, T6	**R2: Similarity and Previous Pattern Matching**	To provide similar and dissimilar classified fraud cases to enable comparative analysis of AI predictions
T3, T12	**R3: System Interactivity**	To provide interactivity and enable experts with a dynamic view on data and detail when comparing fraud cases or investigating the impact of attributes on predictions
T4, T7, T13	**R4: Relationships**	To provide relationships between attributes in single and multiple classified fraud cases
T8, T9	**R5: Importance of Attributes**	To provide the importance of attributes used by the AI model for classifying fraud cases
T10, T11	**R6: Inference Path**	To provide the reasoning process of the AI model for classifying fraud cases, based on rules and a friendly language for fraud experts
T12, T13	**R7: Impact on Specific Decisions**	To provide the impact of attributes used by the AI model on specific classifications given to suspicious fraud cases

The fraud expert's tasks and requirements are aligned with the 8 design features extracted from existing explanation methods (DF1 - DF8). Table 2 presents the descriptions for each design feature. According to the explanation scope, target, and type, the dimensions of EM are also highlighted within the description of each design feature. Those design features reveal meta-design principles of EM, from which we derive the five design principles (DP1 - DP5) developed for user-centric XAIDSS in fraud detection.

Therefore, our design principles aim to provide utility and information quality to fraud experts and researchers in XAI to clearly understand how to set up explanation methods for decision support in fraud detection. Each principle highlights the users, aim, mechanism, and rationale for supporting experts with explanations to understand AI predictions for fraud. We focus on the alignment between explanations and expert's tasks and requirements. Section 5 presents the evaluation of our principles regarding the achievement of the goal established in this study.

5 Evaluation and Results

5.1 Evaluation and Experiment Design

We conduct a naturalistic ex-ante evaluation to assess the extent to which our design principles attend the utility and quality requirements established at the

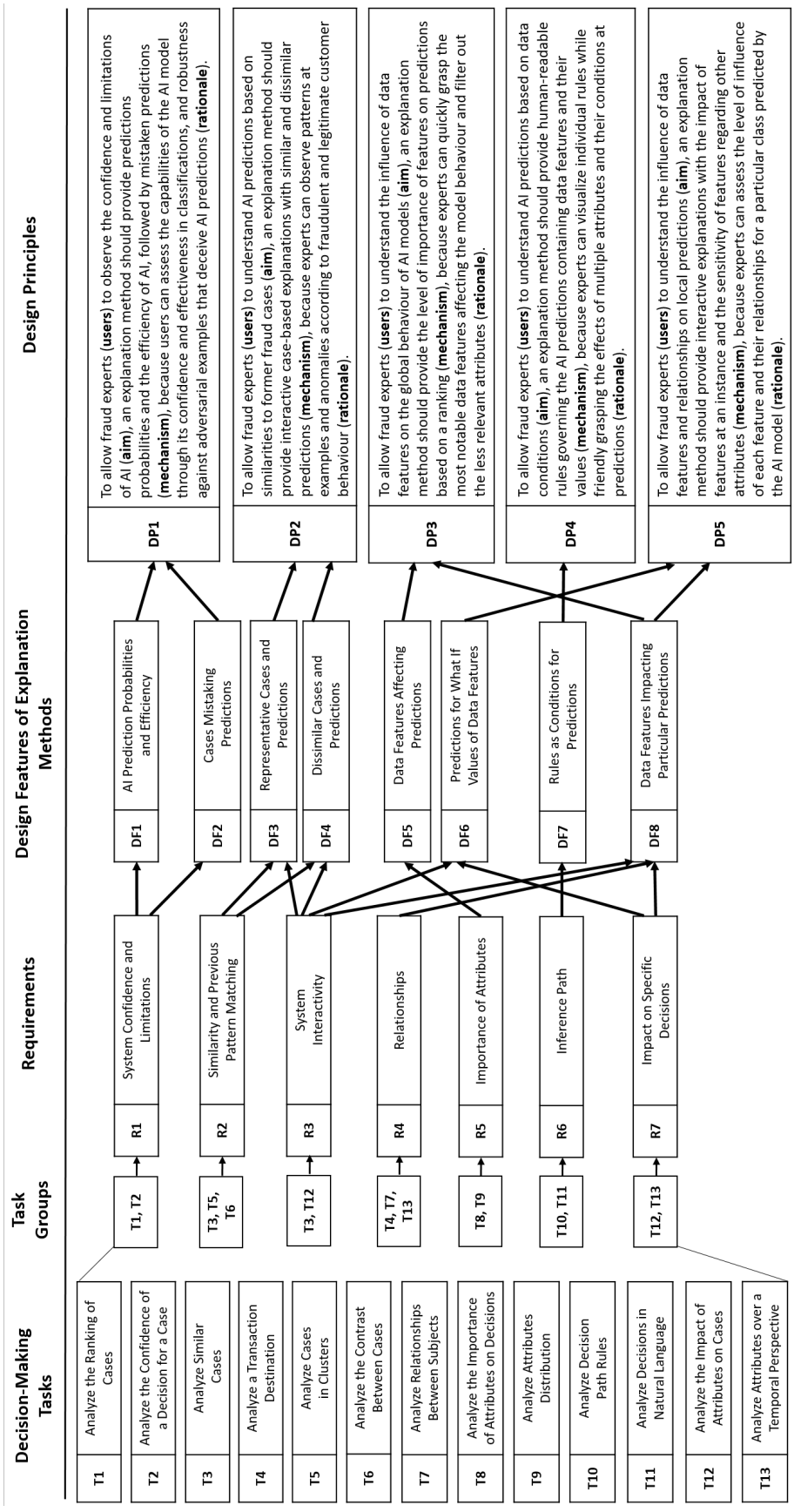

Fig. 1. Design Principles for user-centric XAI in fraud detection

Table 2. Descriptions for design features, which reveal meta-design principles of explanation methods

Design Features	Description
DF1: AI Prediction Probabilities and Efficiency	The inclusion of prediction probabilities and information on the efficiency of AI may give the user with an explicit visualization of the AI model confidence, and attend R1.
DF2: Cases Mistaking Predictions	Explanations containing local (scope) cases (target) which deceive the AI model predictions, such as those provided by adversarial explanation methods (type). These explanations may support the understanding of weaknesses of AI models and attend to R1.
DF3: Representative Cases and Predictions	Explanations containing local (scope) similar cases (target) to a current prediction, such as those provided by prototypes as explanation methods (type). These explanations may help the comparison between cases by an expert and attend R2 and R3.
DF4: Dissimilar Cases and Predictions	Explanations containing local (scope) dissimilar cases (target) to a current prediction, such as those provided by criticisms as explanation methods (type). These explanations may support the contrast between true positives and outliers, which can attend R2 and R3.
DF5: Data Features Affecting Predictions	Explanations containing global (scope) data features (target) considered important by the AI for learning and predicting classes, such as those provided by Feature Importance (type). These explanations may help experts in understanding what data features deserve more attention when deciding on suspicious cases and attend to R5.
DF6: Predictions for What If Values of Data Features	Explanations containing global and local (scope) changes on values of data features (target) which can switch the AI model prediction for a case (type), such as those provided by Counterfactual explanation methods. These explanations help in visualizing needed changes to shift AI predictions and enable fine-tuning against errors, which can attend R3 and R7.
DF7: Rules as Conditions for Predictions	Explanations containing rules in natural language for global predictions (scope) which possess data feature values (target) and their combinations for AI model predictions, such as those provided by rule-based explanations and Decision Trees (type). These explanations provide human-readable relationships from the data and AI and attend to R6.
DF8: Data Features Impacting Particular Predictions	Explanations containing local (scope) data features (target) that impact the AI model predictions for a given class, such as those provided by Feature Impact methods (type). These explanations may enable the user with understanding the important features that impact local cases towards being true positives for the target class and attend to R3, R4 and R7

start of the project [42]. The ex-ante evaluation aims to assess the partial design of artefacts before their deployment in real settings. We interviewed three fraud experts (minimum of 3 years of experience) within a bank partner. To structure the evaluation, we adopt the utility and information quality framework of [24]. Those authors there provided a practical framework with semiotic-based pragmatic, semantic, and syntactic levels to establish evaluation criteria for the information quality of DSR artefacts. The framework is suitable for complex design environments, which matches our study context of aligning diverse explanation methods to user requirements for XAIDSS in fraud detection.

To perform an evaluation following [24] framework, we employ a problem-centered interview method [63], which is an approach ratified for ex-ante evaluation in DSR studies [55]. We interviewed each expert for one hour on the matters on how they perceive the correctness of DP's terminology matching their experience and knowledge based on syntactic quality (adequacy, accessibility, consistency), semantic quality (unambiguity, preciseness, understandability, interpretability, and accuracy), and the principle's instantiation helping in everyday work. We also encourage experts to provide reasons for their views and enrich our qualitative data collection. We aim to understand their stressing points and issues worthy of further investigation. Our evaluation strategy matches current research assessing the utility and impact of DP [36]. To structure the feedback collected, we also allow experts to provide their answers based on objective criteria following a Likert scale from extremely unlikely to likely.

We also perform a simulation to evaluate the quality of design principles based on their instantiation. We instantiate the principles by developing an interface mockup to implement explanation methods that reflect the design features within our principles, illustrated in Figure 2. The mockup instantiates DP1 by showing the AI confidence after the training and testing phases (DP1). It instantiates DP3 and DP5 by presenting the EM of Local Feature Importance (LFI), Global Feature Importance (GFI), and Feature Impact (FI) to provide relevant data features for predictions and their relationships. We instantiate DP2 and DP4 by presenting the EM of Prototypes for providing experts with similar cases to the fraud under analysis, and we provide Anchors for presenting rules governing predictions. The explanation methods described explain the predictions of a Random Forest model trained on a bank partner's dataset with 3269 suspicious cases out of 7653 transactions over three months. The transactions belong to ten customers of the institution. The set of features adopted for training and testing are: amount, device, anonymous receiver ID, receiver location, sender location, and currency. For each customer, the institution has provided 75% of past transactions for training and 25% for testing. Python programming language version 3.8 is adopted for this implementation and the scikit-learn[6] library for training and testing ML models. Python libraries for LIME [48], Anchors [49] and SHAP [34] are adopted for the implementation of LFI, GFI, FI and Anchor rules. Regarding the EM of Prototypes, we follow the guidelines of [38, 60] and train a KNearestNeighbor [44] classifier based on the SHAP values of the training instances. The complete implementation and dataset used for our simulation results are available externally[7].

An expert would use the described mockup and explanations following our principles. Therefore, this simulation focuses on automatically evaluating the goodness of those explanations to estimate the user confidence in them. The methodology is aligned with the functionally-grounded evaluation established by [17], where an author defines proxy tasks for assessing how good an explanation is in achieving its goal without human participation. Given that our interface has multiple EMs, their goodness is computed according to the explanation types. For Local Feature Importance, Anchors, GFI, and FI, we compute their fidelity to the AI model being explained [40, 43, 46]. For that, we retrieve the features highlighted in those explanations and change their values in data instances until the prediction for those instances changes. We report the average prediction switching point (ASP). Lower values for switching prediction point indicate that the EM presents the features that contributed most towards the predicted class, which would foster user confidence in the explanation [40]. We compare this result with a random deletion of features. Ideally, the average prediction switching point should be lower than a random deletion switching point to assure the quality of explanations.

[6] https://scikit-learn.org/

[7] https://github.com/dougcirqueira/hcii-design-principles-user-centric-explainable-ai-fraud-detection/tree/main/resources/simulation

Concerning the EM of Anchors, we compute the percentage of instances that change their predicted class when following the Anchor rules to change feature values. That should be above 0.5 to ensure the quality of those explanations. For Prototypes, we compute the number of neighbors which match the correct label prediction for the current transaction under analysis. Therefore, we obtain the explanations for predictions belonging to the last suspicious transactions in the test set from five customers with suspicious fraud cases. Those customers are selected because the AI confidence was diverse, with levels ranging from 66% to 98%. We report the Anchor percentage of instances and percentage of Prototypes to illustrate the refereed explanations' goodness following our design principles.

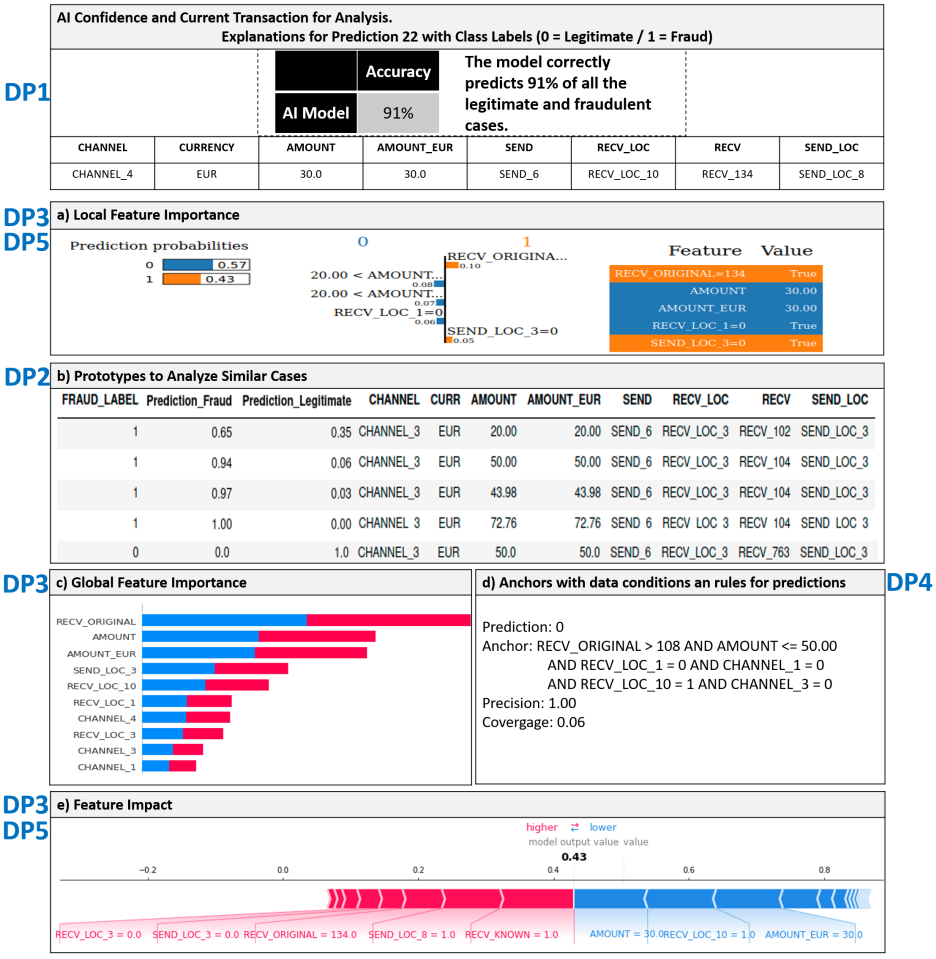

Fig. 2. Mockup with Explanation Methods Following Design Principles

5.2 Evaluation Results

5.3 Information Quality of Design Principles with Syntactic and Semantic Criteria

Table 3 presents the questions and qualitative feedback from experts during our evaluation and the average rating selected by them regarding the information quality criteria assessed. Concerning the syntactic quality, our interviewed experts report that the terminology to describe design principles is extremely likely to be aligned with their internal practices (adequacy). However, they would like to have included examples of cases and fraudulent customer journeys. We suggested the addition of scenarios, such as developed in [11], and the experts agree those are good examples of additions to the DP for fulfilling that need. They also agree that the DP give a sense of better structuring practices and can serve as documentation to rely on for using explanation methods. Furthermore, experts consider the tasks, requirements, design features, and principles to bring the primary information they need, which enable them to quickly grasp the content and discuss with co-workers and interdisciplinary teams (accessibility). Therefore, experts feel confident in following the principles for setting up EM, discussing them with co-workers, and feedback that no contradictions were observed when reading the principles descriptions (consistency).

Concerning the DP's semantic quality, we ask experts regarding DP's unambiguity, preciseness, understandability, and interpretability. Experts agree our principles are extremely likely to be not ambiguous, as they can visualize the differences and how each principle guides the implementation of explanations for AI fraud predictions (unambiguity). They consider the terminology quite precise compared to their internal reviewing process of fraud cases and their connection to design features of EM (preciseness). Regarding understandability, experts highlight the structure of their tasks for analyzing fraud cases is clear and that the alignment of requirements to design features is comprehensible. Therefore, they would be able to tackle why to use or follow particular tasks and use specific EM to deal with the analysis of fraud cases. Therefore, the experts agree they are extremely likely to reflect on their needs when reading the DP for fraud detection, what they can do to support their work, and how to use them (understandability). They appreciate the levels of detail which match their practices, and they agree it is understandable that they would analyze AI predictions with different perspectives through a combination of explanations.

Regarding semantic interpretability specifically, experts acknowledge the DP make it extremely likely to interpret their guidance through the established user, aim, mechanism, and rationale in each principle description (interpretability). Experts conclude the DP structure and layout help understand the DP as it depicts each principle's main goals and what they can do to support their work. Experts are stimulated to give examples of such understanding. They give an example that DP could also help set up separate defense lines for fraud teams, where one team deals with the preliminary filtering of suspicious transactions through the first EM, and more complex cases or new schemes of fraud

go through further analysis. For instance, one expert mentions he could filter out cases based on Local Feature Importance, and discuss more critical cases using further methods provided by our principles. They stress this aspect would also reduce the workload to fraud teams. Finally, experts agree the principles are extremely likely to be accurate regarding their tasks and requirements for fraud detection and how AI predictions can be explained based on their context and dataset that is daily analyzed (accuracy).

Overall, experts regarded the design principles as quite likely to be supportive in their context, as it can be observed by the average score of 2.33 in Table 3, as well as qualitative data and feedback provided by experts. The results highlight our DP's information quality, which would enable fraud experts to reflect on their fraud detection practices and how to employ XAIDSS.

5.4 Quality of Design Principles Instantiation through a Simulation

Regarding the simulation results, Table 4 shows the average prediction switching point for the explanations of LFI, GFI, and FI given different AI confidence levels. The explanations are provided with 50 random seeds. For these methods, the lower the ASP reported, the better the estimated user confidence. Columns 3 to 8 show the ASP for the referred methods. Those are compared to a random deletion of features. The numbers highlighted in bold mean that deleting features by importance reported by EM requires lesser deletions than a random selection, which ensures the quality of explanations. Next, Table 5 shows the percentage of cases with prediction changed following Anchor rules and the cases reported by Prototypes explanations belonging to the correct suspicious case classification. For these methods, the higher the reported value, the better the estimated user confidence. Column 3 shows the results for Anchor rules, which should be at least over 0.5 to be considered good in our scenario. Finally, column 4 shows the simulation results for Prototypes.

6 Discussion and Implications

6.1 Practical Implications of Design Principles

DP1 Each principle would have practical implications when instantiated by researchers and fraud experts. DP 1 states it is essential to enable fraud experts with the capabilities and limitations of AI. Probabilities provide an explicit visualization of AI models' confidence for predictions based on a trained dataset [47, 32]. Moreover, the method of Adversarial explanations enables experts to assess the cases that would affect legitimate predictions. Adversarial transactions are generated in order to deceive the AI, which enables experts to spot weaknesses of the trained models. Adversarial explanations might also be useful for fraud prevention purposes, where simulated fraud schemes can be fed to the AI, which has to determine their legitimacy [66]. Without the instantiation of such a principle, experts might become overconfident in AI predictions and explanations.

Overconfidence is not ideal as users might be misled by blindly believing the AI is correct in every prediction, which can cause damage to AI stakeholders and end-users [12].

DP2 The instantiation of DP 2 provides experts with similarities and dissimilarities to a current case under analysis. According to experts during our evaluation, it is essential to analyze customer behaviour from a dynamic perspective, as it changes over time. Such principle supports the analysis of typical patterns within a dataset [21], as experts need to get insights on typical behaviour of legitimate users. Notably, the EM of Prototypes and Criticisms would play a role in the instantiation of DP 2 [27]. Without this principle, experts would rely only on important features and lack the analogy perspective for understanding predictions, which is inherently part of how humans aim to digest explanations [37].

DP3 DP 3 instantiation enables experts to look closely into what data features the AI is considering as the most important when learning legitimate and fraudulent patterns from the whole dataset of customer transactions. EM methods fulfilling this DP are among the most used in extant Explainable AI studies, as the understanding of important features enables not only explanations for the AI behaviour but also to clarify if the model is working correctly, which is valuable for AI engineers aiming to optimize their models [29]. In the scenario of fraud detection, the analysis of important features is among the first tasks performed by experts when reviewing cases [11]. Therefore, without DP 3, experts would lack explanations for understanding whether the AI is focusing on the correct parameters, according to their domain knowledge in fraud cases. The wide adoption of such a principle is reflected by 66% of studies retrieved in our systematic literature review to establish EM design features, which adopt Global Feature Importance as part of explanations.

DP4 Fraud experts are constantly under pressure for protecting customers and being efficient in reviewing fraud cases. DP 4 instantiation enables a friendly and quick overview of fraud cases, as it advocates for the provision of rules and human-friendly explanations. With rules, experts have an overview of multiple data features and values at once. Indeed, experts have highlighted the need to observe illustrations of the effects of multiple features in AI predictions. When discussing DP 4, experts emphasize they get the sense that its instantiation helps them in deciding faster what an AI prediction means, and if it is right, or if the AI is "thinking wrongly". Without DP 4, experts would not rely on a language-friendly explanation to digest AI predictions, which might affect their performance when working with colleagues and communicating to customers the reasons for suspicious fraud [50].

DP5 Transaction fraud and customer datasets have a temporal nature, and experts need to analyze the local impact of data features on single transactions,

as well as the influence of past behaviour on current transactions. DP 5 instantiation enables experts to observe the local feature impact and the relationships between features in the dataset. Furthermore, the relationship between multiple instances, as transactions or customers, is also embedded into this principle. This principle attends fraud expert tasks when analyzing complex fraud cases involving multiple actors and comparing past patterns, which is constantly performed during their analyses whether through Feature Impact EM or network and graph visualizations of multiple customer transactions over time [16, 9, 26]. Without DP 5, experts would lack such features for understanding AI with temporal and multiple feature and instances perspectives. Therefore, we perceive our design principles are aligned with the user-centric XAI community and can be useful for instantiating XAIDSS to conducting empirical studies and assess the impact of explanations on fraud experts' work and confidence in AI.

6.2 Simulation Findings

From a quantitative perspective, our simulation results illustrate the quality of DP's instantiations for fraud experts. For the EM of LFI and GFI, the ASP is lower for features deleted based on explanations in 6 out of 8 customer cases, respectively. When changing features based on Anchor rules, the prediction changes for more than 50% of transactions for 5 out of 8 customer cases. Regarding the EM of Prototypes, the method returns at least 93% of similar transactions to a suspicious case that belongs to the same class. Therefore, the reported results obtained based on the ASP computation highlight the data features deemed as important by the AI model implemented, impacting predictions if compared to random features, which is required by expert tasks when analyzing suspicious transactions. Finally, experts compare and contrast fraud cases when analyzing suspicious transactions, and the results are satisfactory to attend those needs as illustrated by Anchor and Prototype reported simulation results.

7 Conclusions

Experts recognize the value of Artificial Intelligence decision support for fraud analysis and detection, despite the lack of transparency of black-box models and, consequently, trust in AI predictions. We develop design principles to align expert requirements and explanation methods for decision support and understanding AI predictions. We adopt an Information Systems perspective and a design science research methodology in the study. Based on the results, we argue that IS theoretical lens is valuable towards user-centric Explainable AI development and worth further investigation. The principles may contribute to user-centric XAI design knowledge and Explainable AI decision support in fraud detection, given their foundation based on industry practices and literature. The developed design principles could impact fraud experts' working processes and guide designing fraud operations based on the information provided by explanation

methods through workload splitting. The principles could also be considered by companies developing fraud detection solutions taking into account explainability requirements for the users of such tools.

As limitations, experts highlighted the interest in a prototype with interactivity. However, no interface was deployed at this time, which will be addressed at another iteration of the project. Therefore, as future work, we will conduct an ex-post naturalistic evaluation of the design principles and expand the dataset for experiments with a large bank partner. We can evaluate our artifact's instantiation by assessing an expert's efficiency when using explanation methods instantiated by our design principles. Researchers could also assess the design principles in different fraud detection contexts, including phishing cases. It can be further considered implementing a process perspective for collaboration between experts and explanations for trustworthy and explainable AI decision support, enhancing the efficiency of teams splitting their workload.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: 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. pp. 1–18 (2018)
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
3. Antwarg, L., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv preprint arXiv:1903.02407 (2019)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020)
5. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 (2019)
6. Arzate Cruz, C., Igarashi, T.: A survey on interactive reinforcement learning: Design principles and open challenges. In: Proceedings of the 2020 ACM Designing Interactive Systems Conference. pp. 1195–1209 (2020)
7. Bento, J., Saleiro, P., Cruz, A.F., Figueiredo, M.A., Bizarro, P.: Timeshap: Explaining recurrent models through sequence perturbations. arXiv preprint arXiv:2012.00073 (2020)
8. Cakir, G., Iftikhar, R., Bielozorov, A., Pourzolfaghar, Z., Helfert, M.: Omnichannel retailing: Digital transformation of a medium-sized retailer. Journal of Information Technology Teaching Cases p. 2043886920959803
9. Cheng, D., Wang, X., Zhang, Y., Zhang, L.: Graph neural network for fraud detection via spatial-temporal attention. IEEE Transactions on Knowledge and Data Engineering (2020)
10. Cirqueira, D., Hofer, M., Nedbal, D., Helfert, M., Bezbradica, M.: Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda. In: International Workshop on New Frontiers in Mining Complex Patterns. pp. 119–136. Springer (2019)

11. Cirqueira, D., Nedbal, D., Helfert, M., Bezbradica, M.: Scenario-based requirements elicitation for user-centric explainable ai. In: International Cross-Domain Conference for Machine Learning and Knowledge Extraction. pp. 321–341. Springer (2020)
12. Collaris, D., Vink, L.M., van Wijk, J.J.: Instance-level explanations for fraud detection: A case study. arXiv preprint arXiv:1806.07129 (2018)
13. Creedon, F., O'Kane, T., O'Donoghue, J., Adam, F., Woodworth, S., O'Connor, S.: Evaluating the utility of the irish hse's paper based early warning score chart: A reflective data gathering phase for the design of the reviews framework. In: DSS. pp. 165–176 (2014)
14. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection: a realistic modeling and a novel learning strategy. IEEE transactions on neural networks and learning systems **29**(8), 3784–3797 (2017)
15. Dellermann, D., Lipusch, N., Ebel, P., Leimeister, J.M.: Design principles for a hybrid intelligence decision support system for business model validation. Electronic markets **29**(3), 423–441 (2019)
16. Didimo, W., Liotta, G., Montecchiani, F., Palladino, P.: An advanced network visualization system for financial crime detection. In: 2011 IEEE Pacific visualization symposium. pp. 203–210. IEEE (2011)
17. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
18. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Communications of the ACM **63**(1), 68–77 (2019)
19. Dudley, J.J., Kristensson, P.O.: A review of user interface design for interactive machine learning. ACM Transactions on Interactive Intelligent Systems (TiiS) **8**(2), 1–37 (2018)
20. Floridi, L.: Establishing the rules for building trustworthy ai. Nature Machine Intelligence **1**(6), 261–262 (2019)
21. Gee, A.H., Garcia-Olano, D., Ghosh, J., Paydarfar, D.: Explaining deep classification of time-series data with learned prototypes. arXiv preprint arXiv:1904.08935 (2019)
22. Gregor, S., Hevner, A.R.: Positioning and presenting design science research for maximum impact. MIS quarterly pp. 337–355 (2013)
23. Gregor, S., Kruse, L.C., Seidel, S.: The anatomy of a design principle. Journal of the Association for Information Systems (2020)
24. Helfert, M., Donnellan, B., Ostrowski, L.: The case for design science utility and quality-evaluation of design science artifact within the sustainable ict capability maturity framework. Systems, Signs and Actions: An International Journal on Information Technology, Action, Communication and Workpractices **6**(1), 46–66 (2012)
25. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
26. Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., Gama, J.: How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. arXiv preprint arXiv:2101.08758 (2021)
27. Kim, B., Koyejo, O., Khanna, R., et al.: Examples are not enough, learn to criticize! criticism for interpretability. In: NIPS. pp. 2280–2288 (2016)
28. Klopper, R., Lubbe, S., Rugbeer, H.: The matrix method of literature review. Alternation **14**(1), 262–276 (2007)
29. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International Conference on Machine Learning. pp. 1885–1894. PMLR (2017)

30. Kumari, P., Mishra, S.P.: Analysis of credit card fraud detection using fusion classifiers. In: Computational Intelligence in Data Mining, pp. 111–122. Springer (2019)
31. Laughlin, B., Sankaranarayanan, K., El-Khatib, K.: A service architecture using machine learning to contextualize anomaly detection. Journal of Database Management (JDM) **31**(1), 64–84 (2020)
32. Le, T., Wang, S., Lee, D.: Why x rather than y? explaining neural model'predictions by generating intervention counterfactual samples (2018)
33. Li, Z., Liu, G., Jiang, C.: Deep representation learning with full center loss for credit card fraud detection. IEEE Transactions on Computational Social Systems **7**(2), 569–579 (2020)
34. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in neural information processing systems. pp. 4765–4774 (2017)
35. Marino, D.L., Wickramasinghe, C.S., Manic, M.: An adversarial approach for explainable ai in intrusion detection systems. In: IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society. pp. 3237–3243. IEEE (2018)
36. Meier, P., Beinke, J.H., Fitte, C., Teuteberg, F., et al.: Generating design knowledge for blockchain-based access control to personal health records. Information Systems and e-Business Management pp. 1–29 (2020)
37. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
38. Molnar, C.: Interpretable Machine Learning. Lulu. com (2020)
39. Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein, G.: 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 (2019)
40. Nguyen, D.: Comparing automatic and human evaluation of local explanations for text classification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1069–1078 (2018)
41. Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. User Modeling and User-Adapted Interaction **27**(3-5), 393–444 (2017)
42. Ostrowski, Ł., Helfert, M., Hossain, F.: A conceptual framework for design science research. In: International Conference on Business Informatics Research. pp. 345–354. Springer (2011)
43. Papenmeier, A., Englebienne, G., Seifert, C.: How model accuracy and explanation fidelity influence user trust. arXiv preprint arXiv:1907.12652 (2019)
44. Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
45. Raj, S.B.E., Portia, A.A.: Analysis on credit card fraud detection methods. In: 2011 International Conference on Computer, Communication and Electrical Technology (ICCCET). pp. 152–156. IEEE (2011)
46. Ramon, Y., Martens, D., Provost, F., Evgeniou, T.: Counterfactual explanation algorithms for behavioral and textual data. arXiv preprint arXiv:1912.01819 (2019)
47. Renard, X., Laugel, T., Lesot, M.J., Marsala, C., Detyniecki, M.: Detecting potential local adversarial examples for human-interpretable defense. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 41–47. Springer (2018)
48. Ribeiro, M.T., Singh, S., Guestrin, C.: "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, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016)

49. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: AAAI Conference on Artificial Intelligence (AAAI) (2018)
50. Rosenfeld, A., Richardson, A.: Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems **33**(6), 673–705 (2019)
51. Rosson, M.B., Carroll, J.M.: Scenario based design. Human-computer interaction. boca raton, FL pp. 145–162 (2009)
52. Seidel, S., Chandra Kruse, L., Székely, N., Gau, M., Stieger, D.: Design principles for sensemaking support systems in environmental sustainability transformations. European Journal of Information Systems **27**(2), 221–247 (2018)
53. Seidel, S., Watson, R.T.: Integrating explanatory/predictive and prescriptive science in information systems research. Communications of the Association for Information Systems **47**(1), 12 (2020)
54. Sokol, K., Flach, P.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 56–67 (2020)
55. Sonnenberg, C., Vom Brocke, J.: Evaluation patterns for design science research artefacts. In: European Design Science Symposium. pp. 71–83. Springer (2011)
56. Venable, J.: The role of theory and theorising in design science research. In: Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006). pp. 1–18. Citeseer (2006)
57. Walls, J.G., Widmeyer, G.R., El Sawy, O.A.: Building an information system design theory for vigilant eis. Information systems research **3**(1), 36–59 (1992)
58. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable ai. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–15 (2019)
59. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. MIS quarterly pp. xiii–xxiii (2002)
60. Weerts, H.J., van Ipenburg, W., Pechenizkiy, M.: Case-based reasoning for assisting domain experts in processing fraud alerts of black-box machine learning models. arXiv preprint arXiv:1907.03334 (2019)
61. Weerts, H.J., van Ipenburg, W., Pechenizkiy, M.: A human-grounded evaluation of shap for alert processing. arXiv preprint arXiv:1907.03324 (2019)
62. West, J., Bhattacharya, M.: Intelligent financial fraud detection: a comprehensive review. Computers & security **57**, 47–66 (2016)
63. Witzel, A., Reiter, H.: The problem-centred interview. Sage (2012)
64. Wolf, C.T.: Explainability scenarios: towards scenario-based xai design. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. pp. 252–257 (2019)
65. Yang, Q., Suh, J., Chen, N.C., Ramos, G.: Grounding interactive machine learning tool design in how non-experts actually build models. In: Proceedings of the 2018 Designing Interactive Systems Conference. pp. 573–584 (2018)
66. Yuan, X., He, P., Zhu, Q., Li, X.: Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems **30**(9), 2805–2824 (2019)
67. Zerilli, J., Knott, A., Maclaurin, J., Gavaghan, C.: Transparency in algorithmic and human decision-making: is there a double standard? Philosophy & Technology **32**(4), 661–683 (2019)
68. Zheng, L., Liu, G., Yan, C., Jiang, C.: Transaction fraud detection based on total order relation and behavior diversity. IEEE Transactions on Computational Social Systems **5**(3), 796–806 (2018)

Table 3. Information quality evaluation results for design principles (Expert rating goes from a -3 to 3 scale, where -3 and 3 represent extremely unlikely and likely, respectively. The letter E stands for Expert)

Theme	Questions and Qualitative Feedback (A for Answer)	Quantitative Average Expert Rating			
		E1	E2	E3	Average
Syntactic Information Quality	1) Do you consider the design principles have an adequate representation in accordance with existing instructional material you have seen and used in your organization to understand software tools? (Adequacy) A1: DP terminology is aligned with internal practices. The principles could be accompanied by scenarios to illustrate the fraud detection context they are applicable.	3	2	2	2.33
	2) Do you consider that the design principles have a good accessibility to yourself and for your discussion with colleagues regarding your needs for fraud detection, and how Explainable AI and methods can support you? (Accessibility) A1: DP can structure their practices and primary information for interdisciplinary discussions with colleagues that do not have deep knowledge of AI. There is no need for lengthy material given the principles A2: DP can serve as documentation to rely on for using EM to analyze and understand fraud cases	3	3	1	2.33
	3) Do you consider that the descriptions for design principles are consistent and do not bring contradictions? (Consistency) A1: No contradictions were observed when reading the DP descriptions. Each DP description is unique to understand their guidance and capabilities when instantiated A2: DP bring confidence for interdisciplinary discussion with colleagues from various departments, including technical and management levels	3	2	2	2.33
Semantic Information Quality	4) Do you consider the design principles description is not ambiguous and that there are no principles that could be viewed as the same? (Unambiguity) A1: DP description and template enable the visualization of the differences and how each principle guides the implementation of explanations for AI fraud predictions	3	3	1	2.33
	5) Do you consider the terminology in the design principles is precise to describe your needs for fraud detection tasks and analysis? (Preciseness) A1: DP terminology is precise compared to their internal discussion themes and analysis tasks for fraud cases, as well as design features of EM	3	2	2	2.33
	6) Can you interpret what the design principles can do to support your work based on their description for understanding fraud cases classified by Artificial Intelligence? (Understandability) A1: The flow from tasks to requirements and design principles is comprehensible and reasonable A2: Experts can understand why to use or follow particular tasks and using specific EM, as well as what EM can do to support their work and how to use them A3: DP highlight the different perspectives through a combination of explanations that can be leveraged to understand AI and review fraud cases	3	1	2	2
	7) Do you consider it is easy to understand the description of design principles related to your current work? (Interpretability) A1: Experts can understand the DP as it depicts each principle's main goals, what is behind them, and what they can do to support their work A2: DP can be considered for guiding organizational changes, such as for setting up separate defense lines for fraud teams	3	3	2	2.66
	8) Do you consider the description of design principles is accurate and free of error when you relate it to your current work and tools for fraud detection? (Accuracy) A1: DP description is accurate to reflect experts tasks and requirements for fraud detection, and how AI predictions can be explained A2: DP give the confidence to refer to internal practices and select appropriate tools for fraud detection while interacting with AI predictions	3	2	2	2.33

Table 4. Estimated user confidence based on average prediction switching point for instantiated explanation methods of Local Feature Importance (LFI), Global Feature Importance (GFI) and Feature Impact (FI)

Average AI Model Confidence for Transaction	Customer ID and Transaction Number	The lower the value, the better the estimated user confidence					
		LFI		GFI		Feature Impact (FI)	
		ASP by Deleting Explanation Features	ASP by Random Deletion	ASP by Deleting Explanation Features	ASP by Random Deletion	ASP by Deleting Explanation Features	ASP by Random Deletion
0.98	SEND_1 - 846	1	1	1	1	**0.63**	1
0.95	SEND_2 - 115	**0.2**	0.8	**0.94**	0.95	**0.61**	0.95
0.89	SEND_4 - 37	**0.2**	0.56	**0.17**	0.22	0.24	0.18
0.82	SEND_1 - 845	1	0.88	1	1	**0.45**	1
0.79	SEND_3 - 89	**0.22**	0.56	**0.1**	0.12	0.28	0.1
0.75	SEND_4 - 36	**0.17**	0.42	**0.24**	0.28	**0.16**	0.27
0.69	SEND_10 - 22	**0.24**	0.49	**0.17**	0.26	**0.23**	0.24
0.66	SEND_4 - 35	**0.11**	0.36	**0.15**	0.17	0.15	0.15

Table 5. Estimated user confidence for instantiated explanation methods of Anchors and Prototypes

Average AI Model Confidence for Transaction	Customer ID and Transaction Number	The higher the value, the better the estimated user confidence	
		Average Percentage of Instances Affected by Anchors Rules	Average Percentage of Prototypes with the Same Label as the Transaction being Analyzed
0.98	SEND_1 - 846	0	**1**
0.95	SEND_2 - 115	0.06	**1**
0.89	SEND_4 - 37	**0.78**	**1**
0.82	SEND_1 - 845	0	**1**
0.79	SEND_3 - 89	**0.98**	**1**
0.75	SEND_4 - 36	**0.6**	**1**
0.69	SEND_10 - 22	**0.64**	**0.93**
0.66	SEND_4 - 35	**0.86**	**1**