# Explainable Robo-Advisors: Empirical Investigations to Specify and Evaluate a User-Centric Taxonomy of Explanations in the Financial Domain

Sidra Naveed[1], Gunnar Stevens[2] and Dean-Robin Kern[3]

[1,2,3] *University of Siegen*

Abstract

Even though Recommender Systems (RS) have been widely applied in various financial domains such as Robo-advisors (RA), these systems still operate as a black box with no or limited explanations. Even in cases, where explanations are provided, such systems are mostly designed from the developers' perspective where the user needs and perspective of explanations are not taken into account. In this work, we aim to address the challenges of designing eXplainable Robo-Advisors (XRA) – by adopting a user-centric methodology. For this purpose, we applied a mixed-method approach, in which we conducted three qualitative focus group discussions (FGD) and supplemented the results with a quantitative survey insight. More specifically, we made two major contributions: 1) We extended the existing explanation categories to contextualize it for the financial domain – by identifying the user's specific needs for explainability in the context of the financial domain, 2) We quantified the user preferences of specific explanations with regard to the financial domain and explainable RA – by evaluating the user's personal relevance (PRE) and perceived quality (PQE) of explanations.

**Keywords**

Recommender Systems in Financial Domain, Robo-Advisor, Explainable Robo-Advisor (XRA)

## 1. Introduction

With recent advances in artificial intelligence (AI) and machine learning (ML), a growing number of complex decision-making tasks are delegated to software systems and applications. In this context, recommender systems (RS) have been widely used in various financial services domains (such as online banking, loan, stocks, asset allocation, and portfolio management) – to support people in complex financial decisions e.g., financial investment or retirement planning [1].

Compared to other domains, such as movie or song recommendations, the financial domain has several peculiarities: the finance world is complex where the risk involved in wrong decisions is high and the average financial literacy of common users might be quite low [2]. For instance, most people might be familiar with the term *Action Movie*, but technical terms such as *Bonds* or *ETFs* are typically known to domain experts only. Another issue refers to the long-term validation of the recommendation quality. In the case of movies, a user can directly evaluate the quality of the recommended movie by watching it. This is different in the case of financial

*IntRS'22: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 22, 2022, Seattle, US (hybrid event)*

sidra.naveed@uni-siegen.de (S. Naveed); gunnar.stevens@uni-siegen.de (G. Stevens); dean-robin.kern@uni-siegen.de (D. Kern)

recommendations, where the actual recommendation quality might only be evaluated in the long run e.g., years after investing money in buying a house or investing in funds. Furthermore, the consequences of accepting the recommendations are different, where an unsuitable movie recommendation might be bothersome, but an unsuitable financial recommendation can have a dramatic impact on the life of a private investor. Thus, applying a RS in the financial domain is a challenging task.

In recent years, Robo-Advisors (RA) [1] have become a popular financial RS that provides a digital alternative for human financial advisors. Such technologies are now being considered the *"new wealth management interface of the 21st century"*. In general, RAs ask a series of questions about the financial situation, risk tolerance, and investment preferences to create a personalized portfolio recommendation [3]. However, unlike in-person advice, users of RA cannot ask for explanations and clarification as the technology still operates as a black box [3].

To deal with the black box situation, various authorities such as the European Commission (EC) and finance regulators have demanded better transparency and explainability of such system [4]. Moreover, recent studies [5, 3] have also shown that such technologies are less accepted by users due to the lack of transparency, explanation, and balancing of information asymmetries.

Currently, explanations of such financial RS have been commonly designed based on the developer's intuition for a *"good explanation"* [6] – without considering the user's perspective on the issue. In this work, we argue for a user-centric approach by considering the user's need and understanding of explanations in the specific context, to make explanations meaningful and usable for the common user.

To address the challenges of designing eXplainable Robo-Advisors (XRA) [7, 3] from a holistic user perspective, we adopted a user-centric approach to mainly address the following research questions:

***RQ1*:** What are the domain-specific user needs for explanations w.r.t. RA systems?

***RQ2a*:** How important are the explanations for the users when interacting with RA systems?

***RQ2b*:** How the quality of explanations are perceived by the users when interacting with RA systems?

These research questions aim to address the following: (1) the specification of explanation taxonomies [8], which takes the domain-specific needs and peculiarities into account, and (2) evaluate the taxonomy from a user's perspective regarding the two aspects mentioned in the literature: the *Personal Relevance of Explanations* (*PRE*) [9] and the *Perceived Quality of Explanations* (*PQE*) [10, 7].

In this work, we adopted a mixed-method approach: To answer the *"What"* aspect of explanations, we first conducted three qualitative focus groups discussion to explore the domain-specific need for explanations by users (*RQ1*).

To answer the *"How"* aspect of explanations, in the second step, we applied a quantitative online survey to evaluate the personal importance (*RQ2a*) and the perceived quality (*RQ2b*) of explanations identified in the first step.

Overall, this work contributes in the following ways:

- Contextualizing and extending the theoretically-driven general taxonomies of explanations with regard to the financial domain and XRA. This theoretical contribution of a user-centric taxonomy provides in-depth insights into the user's understanding of what constitutes a good explanation and why it is needed.
- Quantifying user preferences of explanations with regard to the financial domain and XRA. This practical contribution aims to help designers to identify and prioritize the user need and relevance of specific explanations in the context of designing explainable financial RS.

In the following, we will discuss relevant related work. Next, we will describe the methodologies for both the qualitative study using focus group discussions and a quantitative online survey. Afterward, we will present the results and insights from both qualitative and quantitative studies. Finally, we will critically discuss the findings of both studies and conclude the work by providing an outlook for future research.

# 2. Literature Review

## 2.1. Explanations in Recommender Systems

In the recommender systems (RS) literature, a number of explanation approaches for RS have been proposed [11]. The existing research in explainable RS has shown that explanations can be beneficial for the success of RS in different ways e.g., providing the reasoning behind a recommendation, enhancing the system acceptance by providing users with negative and positive consequences of recommendations, helping users in making well-informed decisions, or enabling communication between the service provider and the user [12, 13]. Despite the considerable amount of research in RS for generating and presenting explanations, providing adequate explanations from the user's perspective, and addressing the user's needs for explanations in a specific context or domain, remain under-explored. Providing different types and levels of explanations could impact the user's perception of the system in various forms e.g., lack of explanations could result in difficulty in understanding recommendations, which could negatively affect the overall system acceptance [14].

Different types and classifications of explanations have been proposed in the RS literature. A work presented in [15] provided a detailed overview of different explanation types along with the question that can be addressed by each explanation type. These explanation types are case-based, contextual, contrastive, counterfactual, everyday, scientific, simulation-based, statistical, and trace-based explanations. Others have classified explanations in terms of the underlying algorithm (e.g., collaborative filtering, content-based, or hybrid) or different sources of information that influence the style of explanations generated [12].

In line with the type of information sources that generate explanations, a more recent work, presented in [8], has provided a systematic literature review on explanations of RS. The authors categorized explanations into four main categories based on the type of information presented in the explanations. These explanation types are:

1. **User Preferences or Input-Output Explanation** – explanation exploits the content related to the user's provided inputs, which is further measured in terms of decisive input

values, preference match, feature importance analysis, and suitability estimate.

2. **Decision Output or Outcome Explanation** – explanation is provided by analyzing the features of the alternative decisions, which might include a list of features and pros and cons of each alternative, or the decisive features used in the inference process.
3. **Decision Inference Process or Procedural Explanation** – explanation based on the inference process of a specific decision problem, which could be provided in terms of the inference trace, inference and domain knowledge, decision method side-outcomes, and self-reflective statistics.
4. **Background and Complementary Information or Knowledge-Based Explanation** – explanation based on the additional background information related to the specific decision problem, information about the knowledge sources used in the inference process, past suggestions, user choices in similar situations, etc.

## 2.2. Financial Recommender Systems

RS for stock investments have existed since the 1990s. While early work focused on individual stocks [16], there is a trend in the literature towards recommending entire portfolios [1], where the portfolio management shall consider different assets such as stocks, bonds, ETFs, crypto-assets to provide the broadest possible risk diversification, match the risk appetite and financial situation of the investor, and consider social-ethical constraints. Algorithmic portfolio management traditionally uses statistical procedures, but in recent years machine learning, especially deep learning methods have also become popular.

Robo Advisors (RA) have become the most prominent example of portfolio management in the mass market [17]. As they currently work as black boxes, the call for explainable RAs becomes a serious issue in society and as well as in academia [1].

To address the issue of the black box situation, in the last years technical studies have been carried out [3, 18]. For instance, Babaei et al. [18] presented a portfolio optimizer, where the input was a set of cryptocurrencies, and the output is the recommended portfolio. To provide input-output explanations, they used the framework SHAP, which computes the Shapley values of each cryptocurrency in the recommended portfolio. Krishnan et al. [3] used the data from the RA system *Paytm Money*. This study aimed to show the link between the answers provided for the initial questions asked by the RA and the resulting risk classes determined by the RA. For this, they explored different XAI frameworks such as LIME, SHAP, and DeepLift, to provide input-output explanations.

Both works [3, 18] did not run any user study, but only focused from an engineering perspective on the technical feasibility. Addressing the human desirability for explanations from an HCI perspective, studies such as [10, 7, 19] have evaluated the effect of providing explanations in the financial domain.

Ben et al. [10] conducted an Amazon Mechanical Turk study, where the authors compared the effects of providing different types of explanations on user's perception ("no explanation", "human expert", "global", "feature-based", and "performance-based"). In this study, participants made financial decisions in a fictive game scenario, selling lemonade with the help of an advisory system. Based on the results of their experiment, the authors concluded that there was a significant difference in perception of explanations, and that the willingness to accept non-human advice was higher on average than the so-called human advice.

Schemmer et al. [19] implemented an AI advice system, which estimates the price of a property to prevent buying overvalued houses. Their system estimates the price as the output by various input features, such as the year of construction or living area. The system gives input-output explanations by showing the feature importance calculated by the LIME framework [20]. The researchers conducted two focus groups, where the quality feedback given indicates that explanations, in general, will be useful but for instance, the concept of feature importance was difficult to understand by users and some also found it confusing to get multiple explanations.

Deo et al. [7] have conducted a user study regarding XRA. For this reason, the authors used a replica RA system. This system assigns a risk category based on ten questions answered by the user. In the second step, a recommender system algorithm matches the user's risk profile to a fund risk profile to give funds recommendations. Users got additional explanations about local and global feature importance using a SHAP-like visualization. In addition, explanations based on features that affect the user and fund risk were provided. Using an online survey, Deo et al. [7] measured the user preferences and how the user comprehends the provided explanations. The authors concluded that users benefit from explanations that relate the user input to the system outcome. Yet, explanations should be simple, clearly stated, and meaningful so as not to disrupt the user experience.

Overall, even though the application of RS can be found extensively in various financial domains e.g., online banking, loan, insurance, real estate, stocks, and asset allocation, to the best of our knowledge, providing explanations in financial RS is underexplored in current explainable RS literature. This holds especially true for RAs, which have become popular in recent years. There is not just a lack of quantitative studies evaluating XRA, but also qualitative studies understanding users' perceptions and the need for explanations in the financial domain. Regarding this, our work complements the research presented in [7], in which we relied on the above-mentioned theoretical classification of explanations presented in [8] – used as a sensitizing concept to investigate the user's perception, need, or demand for explainability in the financial domain.

## 3. Methodology

In this work, we used a mixed-method approach, supplementing the qualitative focus group discussions with a quantitative online survey. In the following, we will describe both methodologies in detail.

### 3.1. Qualitative Focus Group Discussion

By its very nature, users' needs and personal meanings are subjective and open-ended. For this reason, we used a qualitative research approach using Focus Group Discussions (FGD) [21]. As a qualitative research technique, FGD has the ability to explore topics in-depth and determine the constituting elements of the phenomenon (the "*What*") and their meaning (the "*Why*") from a user perspective [22]. Such an explorative approach is particularly suitable when there is a lack of predetermined hypotheses, theoretical concepts need to be contextualized regarding a specific domain, or different perspectives on the issue need to be explored.

We conducted three focus group discussions with the following stakeholders to consider various perspectives about designing and using an eXplainable Robo-Advisor:

1. **Domain Experts** – participants with background and/or expertise in the financial domain.
2. **HCI Experts** – participants with background and/or expertise in Human-Centered Interaction and eXplainable AI (XAI).
3. **Common Users** – participants with no or limited background and/or expertise in the financial domain and XAI.

**Table 1**
Summary of participants in terms of background, level of technical knowledge, or experience in the financial domain.

| | ID | Gender | Age | Occupation and Background Knowledge | Expertise with Financial Domain | Experience in Years |
|---|---|---|---|---|---|---|
| **Domain Experts** | P01 | Male | 32 | Research Assistance in Social-Interactive Robots | Self-investor and have knowledge about business models, stocks, Crypto, and momentum trading | 14 years |
| | P02 | Male | 30 | Research Assistant in Machine Learning in the area of production and Explainable AI | Self-investor in stock trading, medium to long-term investments, and value investment | 5 years |
| | P03 | Male | 27 | Research Assistant in Cyber-Physical Systems | Self-investor in stocks and ETFs through stock picking | 4 years |
| | P04 | Male | 46 | CEO of a finance company | Experience in investing and trading, apprenticeship at Deutsche Bank | 25 years |
| | P05 | Male | 25 | Student of Business Administration and Business Informatic | Assistant financial advisor, Self-learner about financial mathematics and stock market | >1 year |
| **HCI Experts** | P06 | Female | - | Research Assistant in Human-Centered AI, Digital Transparency, UX Design and HCI | No experience | - |
| | P07 | Female | 27 | Research Assistant in Digital Tasting, Media Practices, Sustainability, and HCI | No experience | - |
| | P08 | Female | 33 | Research Assistant in HCI and Usable Security | No experience | - |
| | P09 | Female | - | Research Assistant in HCI | No experience | - |
| **Common Users** | P10 | Male | 24 | Research Assistant in Business Informatics | Less than one year of experience in trading | ∼1 year |
| | P11 | Female | 25 | Student of Entrepreneurship and SME Management | No Experience | - |
| | P12 | Male | 25 | Student of Masters of Business Analytics | Basic Knowledge about ETFs and financial market | ∼2 years |
| | P13 | Male | 25 | Student of Masters of Science and Business Informatics | No Experience | - |

We used the convenience sampling method [23] to recruit participants for the different focus groups. In total, we recruited 13 participants (Domain Experts: 5, HCI Experts: 4, Common Users: 4). All of them were well-educated (at least a university degree), but only the domain experts had long experience with the financial domain (see Table 1).

The domain experts and common users focus groups were conducted online using video conferencing software with cameras switched on. The whole session was video recorded with the consent of all participants. To conduct the sessions online for the two groups, for each group we organized a virtual whiteboard using *Miro*[1]. The HCI experts focus group was conducted in presence and audio of the entire session was recorded with the consent of all participants.

[1] https://miro.com

### 3.1.1. Study Procedure

The procedure for all focus groups was similar, but we had to make some adaptations for the special needs of the online FGD. To conduct the focus group discussions, we followed the steps and procedures described below which are also shown in Figure 1.

1. **Welcome:** Participants were welcomed and were asked to introduce themselves in terms of their background knowledge as well as their expertise in the financial domain.
2. **Introduction I:** A short introduction on the purpose of the focus group discussion was provided as well as the expected outcomes of the study were explained to the participants.
3. **Introduction II:** In case of the Onsite-FGD, participants were given a brief explanation of how to use Post-Its to record their thoughts. In the case of the Online-FGD, participants were given a brief introduction to the *Miro Board* to collect thoughts and structure the discussion.
4. **Walk through:** Participants were given a walk-through to a replica of a real (German) Robo-Advisor *bevestor*[2]. Our RA replica was an interactive mock-up that was built using Axure[3], with no real functionality but allowing participants to interact with it.
5. **Evaluate and Discussion:** Participants were then asked to explore and critically analyze the prototype. The following session was designed for brainstorming and reflects a discussion covering three steps:

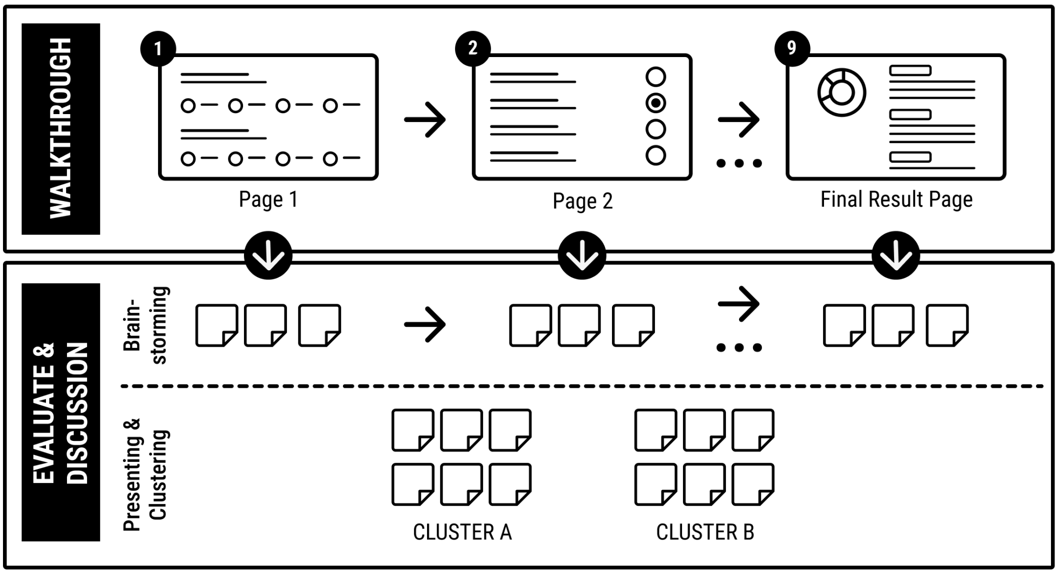

**Figure 1:** Procedure for Focus Group Discussions

[2] https://bevestor.de/

[3] https://ancnjj.axshare.com

- Brainstorming – in which participants were required to write down any questions they want the system to answer or explain to them, related to any aspect of the system that is unclear to them.
- Presenting – in which participants were asked to present and explain the motivation behind each written question.
- Clustering – in which participants were then asked to create clusters of similar questions and label the clusters at the end.

Each focus group was carried out by two researchers. One researcher took on the role of a moderator. Her task was to explain the individual steps and stimulate the discussion. Beyond this, she refrained from interfering. For instance, she took care not to participate in the discussion herself or to evaluate the opinions and views of participants. The second researcher took on the role of an observer by taking supplementary field notes in addition to the audio and video recordings.

### 3.1.2. Data Analytics

All FGDs were audio recorded and transcribed using amberscript[4]. The analysis was done with the help of MAXQDA[5] following the principles of the *Thematic Analysis* [24]. During analysis, each FGD data set was coded by two independent researchers. Subsequently, the two researchers discussed and analyzed each FGD jointly. The codes were consolidated and used to define themes, which capture and summarize the core issue of coherent and meaningful pattern [24]. Themes were discussed in joint interpretation workshops by the group of authors to gain a mutual understanding of the material.

Regarding the coding and defining themes, three principles have been discussed in the literature: First, the principle of emergent, inductive, data-driven, and button-up coding [25], where the researcher starts with a blank sheet and all codes resulted only from reading and interpreting the empirical data. Second, the principle of deductive, theory-driven, and top-down coding [25] starts with a pre-defined category system, where the aim of the analysis is mainly to validate and illustrate the category system based on the empirical data. Third, the principle of abductive, reflexive, and counter-current coding that integrates the elements of both approaches [24]. The approach seems to sensitize researchers in their empirical work but needs to be flexible enough to allow for new experiences.

In our research, we adopted this third, abductive coding principle, where the theoretical concepts outlined in Section 2.1 served as a sensitizing lens to analyze our data but allowed us to open to new experiences that could not be well-explained by existing explanation categories. As a result, our coding represents a dialog between the theory and the empirical data, by going through the data and themes several times to refine them.

## 3.2. Quantitative Online Survey

The FGD is an established methodology in qualitative research for stimulating group discussion and exploring a topic from different perspectives [26]. However, the method is limited in terms

[4] https://www.amberscript.com/en/

[5] https://www.maxqda.com/

of assigning specific statements to a person and quantitatively measuring personal perceptions and preferences at an individual level. Therefore, we conducted a follow-up online survey, to quantitatively evaluate the user perception of the RA replica in terms of its explainability and to quantify the user relevance for different types of explanations they want in the RA. To address both *RQ2a* and *RQ2b*, we asked our participants to evaluate both, the *Personal Relevance of Explanations (PRE)* as well as *Perceived Quality of Explanations (PQE)*. Both dimensions have been evaluated in terms of the explanation categories identified in the FGD (see Section 4), i.e. *Recommender Explanation*, *Domain-Specific Information*, and *Shared Understanding*.

We adopted items from the *User-Centric Evaluation framework* [27] to get a comprehensive evaluation of the *Recommender Explanation* category. As this framework does not cover the aspects of *Domain-Specific Information* and *Shared Understanding*, for these categories we created items on our own, where we tried to use the wordings from the FGDs as close as possible. For evaluating the *PRE*, the questionnaire items were rated on a five-point scale from *"Not at all important"* to *"Very important"* (see Table 3 and Table 4). For evaluating *PQE*, all questionnaire items were rated on a five-point Likert response scale from *"Strongly Disagree"* to *"Strongly Agree"* (see Table 5).

### 3.2.1. Study Procedure

To conduct the online survey, the following steps and procedures were followed:

1. **Introduction:** A short introduction on the purpose of the online survey and the procedure of the survey was provided to the participants.
2. **Prototype Exploration:** Participants were presented with the same RA replica that they saw before in the FGD.
3. **Evaluation:** Participants were then asked to explore and critically analyze the prototype and were asked to return to the questionnaire after exploring the prototype, to answer a series of questions.

Except for P09, all the other 12 participants from the FGDs completely filled out the survey.

## 4. Qualitative FGD Results and Insights

We identified the following major findings from the focus group studies: 1) The users have multiple perceptions and understanding of explanations, 2) The empirical view of explanations is different from the theoretical view of explanations, and 3) Overall, there is a general structure of the empirical view about explanations for all three focus groups. These findings are reported in the following sections.

### 4.1. A User-Centered Taxonomy of Domain-Specific Explanations

To address the *RQ1*, we followed the labels of clusters given by each focus group during the *Clustering* phase. The labeled clusters were: *"Explanation"* (given by domain experts and common users) and *"Definition and Explanation of Terms"* (given by HCI experts). In addition, common users clustered several cards that explicitly used the term "Explaining", but the cluster

**Table 2**
User-centered taxonomy of domain-specific explanations: Combining theoretic reflection with empirical contextualization.
Some theoretic concepts (marked as blue) had to be adopted from external knowledge sources.

| Overall (N=13) | Domain Experts (N=5) | HCI Experts (N=4) | Common Users (N=4) | Domain Contextualization | Theoretical Concepts |
|---|---|---|---|---|---|
| | | | | **Recommender Explanations** | |
| 16 | 7 | 6 | 3 | Impact of input on risk classification<br>Impact of the input on portfolio generation | User Preference or Input-Output Explanation [8] |
| 12 | 6 | 4 | 2 | System's portfolio generation process<br>Assumptions made to generate the portfolio | Procedural Explanation [8] |
| 10 | 5 | 5 | - | Portfolio information w.r.t. other users<br>Additional information used for portfolio generation | Knowledge-Based Explanation [8] |
| 8 | - | 7 | 1 | Portfolio characteristic w.r.t. risk, changes, future options, etc.<br>Reasoning behind the portfolio optimization | Outcome Explanation [8] |
| | | | | **Domain-Specific Information** | |
| 51 | 17 | 23 | 11 | Definition of domain-specific terms and concepts | Information [28] |
| | | | | **Shared Understanding** | |
| 10 | 2 | 7 | 1 | Mutual understanding between answers given by the user and the system's interpretation of answers | Shared Understanding [29] |

was labeled as *"Information"*. Due to this, we also considered the aspect of *"Information"* in our thematic analysis.

By analyzing the codes in detail, we found that the user's definition of explanation is quite broad, but not arbitrary. From a user's point of view, an explanation should help the user to understand and make sense of the system and the provided recommendations. From such a user perspective, we coded 107 total responses in the context of requesting an explanation from the system (see Table 2).

The further thematic analysis reveals that these responses can be grouped into three main categories: *"Recommendation Explanation"* (43 of 107 codes), *"Domain-Specific Information"* (54 of 107 codes), and *"Shared Understanding"* (10 of 107 codes). In the following, we discuss these categories in more detail.

**Recommendation Explanation.** In our analysis, we used the explanation taxonomy presented in [8] and as described in Section 2.1, but for our coding, we used the term "*Recommender Explanation*" to make this category conceptually distinguishable from the other forms of explanations we discovered from the FGD. We used the term *Recommender* to stress that the requested explanations are directly related to the recommendation made by the system. We further used the sub-categories from the explanation taxonomy presented in [8], to get a more fine-grained coding for this category:

*User Preference Explanation* is used by us, to classify all responses where participants wanted an explanation that shows a link between the recommendation and their preferences or the answers given to the questions. For example, *"What if I give a wrong answer?"* (P07), *"What was the impact of each answer I gave on the result?"* (P02), *"What would be the impact of answering it not honestly?"* (P02).

*Output Explanation* is used by us to classify all responses where participants wanted an explanation about the recommended portfolio, but was not directly linked to the input in terms of their answered questions, such as knowing more about risk, expected revenues, or reasons for the specific portfolio composition. For example, *"How high is the chance of having less than 200 % of my investment after 10 years?"* (P03), *"Why the system is not highlighting the aspects* *that were considered or the reason why it suggested certain shares or funds?"* (P02), *"Why am I not 100 % invested?"* (P03).

*Procedural Explanation* is used by us to classify all responses where participants wanted an explanation about the internal logic and reasoning of the system to recommend a specific portfolio. For example, *"What are the assumptions based on?"* (P10), *"Why are the specific assets selected for me?"* (P04), *"Where do the values (proportion) come from?"* (P10), *"Why should I use this system and not other systems like Trading View?"* (P10).

*Knowledge-Based Explanation* is used by us, to classify all responses where participants wanted an explanation about the data used by the system to generate the portfolio or additional information about other users or situations to answer their questions. For example, *"What is the standard duration for investment"?* (P09), *"What is the rate of success for these investment portfolios?"* (P06), *"What is the database used to derive the results?"* (P08), *"What is the average amount of trades of users within my peers?"* (P03).

Only 46 of the 107 coded requests for explanations are classified in the category of *"Recommendation Explanation"*. This indicates that in most cases, explanations requested by participants do not fall into the categories provided by the theoretical taxonomy of explanations [8].

**Domain-Specific Information.** According to [28], the information relates to a specific context that the user is not familiar with. For example, some domain-specific terminologies or concepts, provide knowledge about something comprising either facts or details about a subject, event, or situation, or provide knowledge that adds a value to a situation in a particular context to make it understandable.

We adopt this understanding to define the category of *Domain-Specific Information*, to classify all responses accordingly. For example, *"What is the risk-return profile?"* (P04), *"What exactly is meant by the number of transactions?"* (P12), *"What does security assets mean?"* (P09).

According to the FGD responses, this category was the most prominent one, as 51 of the 107 coded requests for explanations are classified into this category.

**Shared Understanding.** According to [29], *Shared Understanding* can be defined as elaborating the mutual knowledge, beliefs, or assumptions or it can be an elaboration of how the system interpreted the user's goals and objectives. We adopt this understanding to define the category of *Shared Understanding*, to classify all responses where participants wanted an explanation from the system in order to know, if there is mutual knowledge or how the system interpreted their answers to the questions. For example, *"What if I understand small loss differently from you?"* (P07), *"How can I know, how you interpreted my risk assessment answers?"* (P08), *"What about ethical aspects?"* (P02).

This category was the least prominent, as only 10 out of the 107 coded requests for explanations were classified in this category.

## 4.2. Inter-Group Differences

By comparing the focus groups (see Table 2), we further observed that *HCI Experts* group was the most responsive one (52 codes), followed by the *Domain Experts* group (37 codes), where the *Common Users* group was least responsive (18 codes).

In most cases, the responses from the three groups were related to all three aspects, namely *Recommender Explanation*, *Domain-Specific Information*, and *Shared Understanding*. The only differences between the groups were that there were no responses related to *Knowledge-based Explanation* from the *Common Users* group and *Outcome Explanation* from the *Domain Experts.* Overall, the results depict the need for integrating all three aspects for users in the system design to have an explainable financial RS.

# 5. Quantitative Results and Insights

We identified the following major findings from the online survey: 1) The domain-specific aspects of explainability, i.e. *Recommendation Explanation*, *Domain-Specific Information*, and *Shared Understanding*, are equally relevant for participants without any significant differences, 2) The *domain-specific taxonomy* is significantly more important for participants as compared to the *domain-general taxonomy* adopted from the literature, 3) Evaluation of the RA replica is not well-assessed by the participants w.r.t. its *PQE.* In the following, we describe these findings in detail.

## 5.1. Personal Relevance of Explanations (PRE)

To address our *RQ2a*, we identified the users' importance for explanations in the context of the RA system in two steps: 1) Personal relevance of explanations w.r.t. the empirically-driven, domain-specific categories, and 2) Personal relevance of explanations w.r.t. theory-driven, domain-general categories.

### 5.1.1. PRE w.r.t. Domain-Specific Categories

We first computed the overall mean score of the combined three domain-specific categories and we saw that the personal importance of the combined categories for all participants is quite high on average (Avg. Mean $= 4.08$). This is also true for individual categories, especially regarding *Recommendation Explanation* (Mean $= 4.33$) and *Domain-Specific Information* (Mean $= 4.0$) (as shown in Table 3). The results are also in line with the focus group discussion from Section 4.2, where most of the responses were related to these two categories (see Table 2). Moreover, the category *Shared Understanding* (Mean $= 3.91$) was also rated relatively high, indicating the user importance to have explanations about how the system interpreted the user's input.

To statistically check the inter-category differences, we utilized an ANOVA ($\alpha = 0.05$) on the mean scores of all participants. The result shows that there is no significant difference among the three domain-specific categories regarding personal importance ($F(2, 33) = 1.21, p = .310$, $\eta_p^2 = .07$). Despite the insignificant difference, the high mean scores for all three categories still indicate that these categories are important to participants.

We further analyzed the inter-group differences. We observed that *Recommendation Explanation* was rated higher (Mean $= 4.40$) by the *Domain Experts* group. In the case of *Domain-Specific Information* (Mean $= 4.55$) and *Shared Understanding* (Mean $= 4.33$), the *HCI Experts* had higher ratings as compared to other groups.

We checked the observed inter-group differences statistically by applying one-way MANOVA ($\alpha = 0.05$) on aggregated categories. The result showed that the categories are not significantly rated differently among all three groups ($F(6, 14) = 0.69$, $p = 0.66$, $\eta_p^2 = 0.22$). To further

**Table 3**
Personal Relevance of Explanation w.r.t domain-specific categories ($N = 12$)
Note: 5-point Likert Scale coding (1 = "Not at all important", 3 = "Moderately Important", 5: "Very Important"). Significant differences measured with ($\alpha = 0.05$), are marked with *. Higher values (highlighted in bold) indicate better results.

| **Explanation Category and Items** | **Overall Results (N=12)** | | **Domain Experts (N=5)** | | **HCI Experts (N=3)** | | **Common Users (N=4)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *F* | *p* | $\eta_p^2$ |
| **Recommendation Explanation**<br>– For me, it is important that the system explains me why the specific portfolio is recommended to me. | 4.33 | 0.65 | **4.40** | 0.54 | 4.33 | 0.57 | 4.25 | 0.95 | 0.04 | 0.95 | .011 |
| **Domain-Specific Information**<br>– For me, it is important that the system provides me information about specific terms or concepts e.g., what is mixed bond?<br>– For me, it is important that the system provides me with the difference between the terms or concepts e.g., how are bonds different from stocks? | 4.0 | 0.69 | 3.66 | 0.84 | **4.55** | 0.19 | 4.0 | 0.54 | 1.73 | 0.23 | .278 |
| **Shared Understanding**<br>– For me, it is important that the system and I have a shared understanding of my preferences.<br>– For me, it is important to know how the system interpreted my answers. | 3.91 | 0.72 | 3.80 | 0.64 | **4.33** | 0.33 | 3.75 | 1.03 | 0.61 | 0.56 | .121 |

determine individual effects regarding each category, we ran univariate tests. The results shown in Table 3 indicate that there is no significant difference among the groups in terms of individual category i.e. *Recommendation Explanation* ($p = 0.95$), *Domain-Specific Information* ($p = 0.23$), and *Shared Understanding* ($p = 0.56$). As the results were not significant, therefore the observed inter-group differences should not be overrated.

### 5.1.2. PRE w.r.t. Generally Defined Categories

In addition to the empirically grounded categories, to obtain a comprehensive evaluation, we also adopted the well-established, generally defined categories from the explanation taxonomy presented in [8]. This taxonomy covers additional areas, which have been proven to be relevant in other domains. We used the same five-point scale as in Section 5.1.1, to evaluate the *PRE* w.r.t. the categories theoretically defined by [8], namely: *Input-Output Explanation*, *Outcome Explanation*, *Procedural Explanation*, and *Knowledge-Based Explanation.*

We found that the personal importance of the combined categories for all participants is between *"moderately important"* to *"important"* (Avg. Mean $=$ $3.56$). This is also true for individual categories, where the *Input-Output Explanation* seems to be more important for the participants (Mean $= 3.89$), followed by *Outcome Explanation* (Mean $= 3.58$), *Procedural Explanation* (Mean $= 3.54$), and *Knowledge-based Explanation* (Mean $= 3.21$). To statistically check the inter-category differences, we utilized an ANOVA ($\alpha = 0.05$) on the mean scores of all participants. The result shows that there is a significant difference among the four categories regarding personal importance ($F(2, 45) = 5.44, p = .008, \eta_p^2 = .19$). We further ran individual T-tests to check which of these categories are significantly different from others and found that *Input-Output Explanation* received a significantly higher rating ($p = 0.01$) than *Knowledge-Based Explanation.* We further found that *Procedural Explanation* had a significantly higher rating than *Knowledge-Based Explanation* ($p = 0.009$).

Moreover, we analyzed the inter-group differences. We observed that *HCI Experts* have higher ratings for all four categories as compared to other groups i.e. *Input-Output Explanation* (Mean $= 4.25$), *Outcome Explanation* (Mean $= 4.0$), *Procedural Explanation* (Mean $= 4.16$),

**Table 4**
Personal Relevance of Explanation (PRE) w.r.t. domain-independent categories adopted from [8].
Note: 5-point Likert Scale coding (1 = "Not at all important", 3 = "Moderately Important", 5: "Very Important"). Significant differences measured with ($\alpha = 0.05$), are marked with *. Higher values (highlighted in bold) indicate better results.

| **Explanation Category and Items**<br>*For me, it is important to know ...* | **Overall results (N=12)** | | **Domain Experts (N=5)** | | **HCI Experts (N=3)** | | **Common Users (N=4)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | SD | *F* | *p* | $\eta_p^2$ |
| **Input-Output Explanation**<br>... which inputs are used to determine the recommender portfolio.<br>... which of my preferences and constraints are fulfilled by the system.<br>... how changing my answers to the questions would impact the resulting portfolio.<br>... how the suggested portfolio is suitable for me. | 3.89 | 0.47 | 4.0 | 0.17 | **4.25** | 0.25 | 3.50 | 0.61 | 3.46 | 0.07 | .435 |
| **Outcome Explanation**<br>... which of my input values were the decisive factors for the system to recommend me the portfolio.<br>... how my chosen input value(s) would positively impact my investment goals and constraints.<br>... how my chosen input value(s) would negatively impact my investment goals and constraints.<br>... why option A is recommended to me instead of Option B.<br>... which of my input values were not taken into account by the system to generate a portfolio for me. | 3.58 | 0.81 | 3.68 | 0.65 | **4.0** | 0.40 | 3.15 | 1.13 | 1.0 | 0.40 | .182 |
| **Procedural Explanation**<br>... which decision steps were taken by the system to reach to the resulting portfolio.<br>... how the system recommended me the portfolio.<br>... how much the system is confident about the portfolio it suggested to me.<br>... how many times in the past have other users invested money in the portfolio suggested by the system. | 3.54 | 0.83 | 3.50 | 0.72 | **4.16** | 0.38 | 3.12 | 1.05 | 1.47 | 0.28 | .247 |
| **Knowledge-Based Explanation**<br>... which portfolios are recommended to all other users.<br>... which portfolios are most preferred by all other users.<br>... how the assets recommended to me matched my answers to the questions.<br>... which additional information sources have been used by the system to generate the portfolio for me.<br>... how others have answered the questions asked by the system. | 3.21 | 0.86 | 2.96 | 0.82 | **3.93** | 0.61 | 3.0 | 0.93 | 1.50 | 0.27 | .251 |

and *Knowledge-Based Explanation* (Mean $= 3.93$). Another rating pattern that can be observed is that the *Common Users* have worse ratings for all categories except the *Knowledge-Based Explanation*, as compared to other groups.

To statistically check the inter-group differences, we applied one-way MANOVA ($\alpha = 0.05$) on aggregated categories. The result showed that all four categories were not rated significantly differently among all three focus groups ($F(8, 12) = 1.26$, $p = 0.34$, $\eta_p^2 = 0.45$). To determine individual effects regarding each category, we ran univariate tests. The results, shown in Table 4, indicate that there was no significant difference among the groups in terms of individual category, i.e. *Input-Output Explanation* ($p = 0.07$), *Outcome Explanation* ($p = 0.40$), *Procedural Explanation* ($p = 0.28$), and *Knowledge-Based Explanation* ($p = 0.27$). As the results were not statistically significant, the observed inter-group differences should not be overrated.

### 5.1.3. Empirically-driven Vs. Theory-driven Taxonomies

In our study, we combined two types of taxonomy: the domain-specific (as shown in Table 3) and the domain-general (as shown in Table 4).

With our analysis, we further wanted to identify if there is any inter-taxonomy difference regarding personal importance. For this, we first computed the overall mean score of combined categories for each taxonomy and found that on average the personal importance of the general categories taken from theory [8] was rated lower (Mean $= 3.56$) as compared to the domain-specific categories grounded in the empirical data (Mean $= 4.08$). This is also true for individual categories in both taxonomies, where each domain-general category (See Table 4) was rated less important than each domain-specific category (See in Table 3). Analyzing both taxonomies together, a general pattern can also be observed that for most categories, the *HCI Experts* have higher personal importance as compared to other groups.

To statistically check the inter-taxonomy, we conducted T-test and observed that inter-taxonomy difference is significant ($t(82) = 3.20$, $p = .002$, $d = .70$). This significant inter-taxonomy difference indicates that the empirically-grounded categories seem to be more relevant for the participants than the theoretically adopted ones, in the context of the financial domain.

## 5.2. Perceived Quality of Explanation (PQE)

The previous section showed that XRA is an important issue for the participants, especially regarding the categories identified in the FDGs. To further address our *RQ2b*, we asked the participants to evaluate the perceived quality of explanations provided by the RA replica w.r.t. the domain-specific categories i.e. *Recommendation Explanation*, *Domain- Specific Information*, and *Shared Understanding* using a five-point scale from *"Strongly Disagree"* to *"Strongly Agree"*.

**Table 5**
Perceived Quality of Explanations ($N = 12$)
Note: 5-point Likert Scale coding (1 = "Strongly Disagree", 3 = "Neutral", 5: "Strongly Agree"). Significant differences measured with ($\alpha = 0.05$), are marked with *. Higher values (highlighted in bold) indicate better results.

| **Construct and Items** | **Overall Results (N=12)** | | **Domain Experts (N=5)** | | **HCI Experts (N=3)** | | **Common Users (N=4)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *F* | *p* | $\eta_p^2$ |
| **Recommendation Explanation**<br>- The system explained why the specific portfolio is recommended to me.<br>- I understood why the portfolio is recommended to me. | 2.08 | 1.10 | 1.40 | 0.65 | 1.66 | 0.57 | **3.25** | 0.95 | 7.30 | **0.013*** | .619 |
| **Domain-Specific Information**<br>- The information provided by the system is sufficient for me to make my investment decision.<br>- The system provided enough information about specific terms of concepts.<br>- The system provided me with information about the difference between terminologies. | 2.19 | 1.08 | 1.53 | 0.76 | 2.33 | 0.88 | **2.91** | 1.25 | 2.25 | 0.161 | .333 |
| **Shared Understanding**<br>- The system showed me how it interpreted my answers.<br>- I felt that the system and I have a shared understanding of my preferences.<br>- I felt that the system and I have a shared understanding of my risk tolerance. | 2.22 | 0.95 | 1.53 | 0.50 | 2.66 | 1.15 | **2.75** | 0.87 | 3.06 | 0.097 | .405 |

We first computed the overall mean score of the combined three domain-specific categories and saw that the perceived quality of the combined categories for all participants, is quite low (Avg. Mean $= 2.16$) which is equal to *"Disagree"*. This is also true for individual categories (as shown in Table 5), where the *Recommendation Explanation* (Mean $= 2.08$) performed the worst, *Domain Specific Information* (Mean $= 2.19$), and *Shared Understanding* (Mean $= 2.22$) are rated higher, but still below 3 (*"Disagree"*). Overall, the results indicate that the explanations provided by the RA replica are not well-perceived by participants w.r.t. all categories.

To statistically check the inter-category differences, we utilized ANOVA ($\alpha = 0.05$) on the mean scores of all participants. The result shows that there is no significant difference among the categories regarding their perceived quality ($F(2, 33) = .06$, $p = .943$, $\eta_p^2 < .01$). This indicates that participants perceived the explanation quality for all categories as similarly worse.

We further analyzed the inter-group differences. The results shown in Table 5 indicate that *Common Users* perceived the quality of explanations in all cases as higher than the other groups.

To check the inter-group differences statistically, we applied one-way MANOVA ($\alpha = 0.05$) on aggregated categories. The result showed that all categories are not significantly rated differently among the three focus groups ($F(6, 14) = 2.09$, $p = 0.11$, $\eta_p^2 = 0.47$). To determine individual effects regarding each category, we ran univariate tests. The results shown in Table 5 indicate that *Common Users* perceived the quality of explanations in all cases as higher than the other groups. However, only in the case of *"Recommendation Explanation"*, the rating is significantly higher for *Common Users* as compared to other groups ($p = .013$).

## 6. Discussion and Outlook

To address the challenges of designing an eXplainable Robo-Advisor (XRA), we applied a mixed-method approach to qualitatively explore the user's need and understanding of explanations in financial RS through FGD, which we then quantitatively verified and supplemented the findings with an online survey.

First, we addressed our *RQ1*, by conducting three qualitative FGD and identified a user-centered taxonomy of domain-specific explanations. With this taxonomy, we demonstrated that general explanation frameworks as presented in [8], need to be adapted to take the domain-specific needs into account. This is highlighted in our FGD insights which revealed that in addition to *Recommendation Explanation*, the aspects of *Domain-Specific Information* and *Shared Understanding* seem to be highly relevant for the users w.r.t. system explainability (See Table 2). The results are also in line with the study presented in [14], where the insights showed the differences in the users' perception of explainable RS between Digital Cameras and Music domains – indicating the effect of the domain on the user's need and perception of explanations.

We further quantitatively verified the FGD results and addressed our *RQ2a* by evaluating the personal relevance of explanations *(PRE)* in two phases: 1) *PRE* w.r.t. our domain-specific categories, and 2) *PRE* w.r.t. the general categories presented in [8]. For the former case, we did not find any significant differences in ratings for the categories. However, the personal relevance was high for all three categories. The results also showed no significant difference in the ratings of the groups (See Section 5.1.1). For the latter case, we also found no significant difference between the rating of categories as well as the ratings made by the different groups (see Section 5.1.2). Even though the results are not statistically significant, the insights from both cases, showed an interesting pattern – where in all cases except for *Recommendation Explanation* category, *HCI Experts* have a higher rating as compared to other groups (see Table 3 and Table 4). We believe that the reason for this could be twofold: 1) It has been shown that in general, *HCI Experts* have higher expectations from the system to be self-descriptive [30]. This might also be the case w.r.t. explainability, but the limited explanations provided in the RA replica might have resulted in a higher need for explanations, 2) the *HCI Experts* in our sample have no experience or knowledge of the finance domain. This might have also affected the results triggering them to have higher needs and relevance for all explanation categories.

Our study further reveals that all general categories have received lower mean scores compared to the domain-specific categories. The further quantitative comparison reveals that this inter-taxonomy difference is significant. This means on average, the domain-specific categories have higher personal relevance compared to the domain-general categories. These enumerative depictions of the results thus verify the importance of domain-specific explanation needs in the context of financial RS to perceive the system as explainable.

We further addressed our *RQ2b* by evaluating the replica of an existing RA in terms of the perceived quality of the explanations *(PQE)* (See Section 5.2). The participants rated the explanation quality as rather low, which highlights the existing need from academia [9, 1] to improve the explainability of financial RS. An interesting pattern we observed, however, is that *Common Users* rated the *PQE* higher for all three categories as compared to other groups. This could be explained under the assumption that all three groups are different in terms of domain knowledge and their ability to perceive and understand system-provided information. In this context, previous works on explainable RS in complex domains have also shown that the complexity of the domain and decision task, also affect the user's need to see explanations at certain levels of detail [31, 32, 33]. It has been previously shown that novice users seem to benefit more from the RS in a complex domain that provides simple or no explanations, to avoid cognitive overload [14]. This might be the reason that, despite limited or no technical background, *Common Users* have a higher perceived quality of explanations when interacting with the RA replica, which provides limited explanations. In addition, in the case of *Recommender Explanation* the initial questions might serve as a *placebo explanation* [34] for the *Common Users*. They might have implicitly assumed that the recommended portfolio and its corresponding explanations were the results of the answers given by users, – which was not the case in our experiment. Compared to this, it seems that *Domain Experts* are more skeptical about such kinds of placebo explanations, thus reflecting in their lower perceived quality.

Overall, the mixed-method approach of our study provides novel insights. However, the approach has its limitations in terms of the small sample size used for both studies. Despite this limitation, the results still shed a positive light on taking the domain-specific user's needs into account, to design the complex financial RS explainable from the user's perspective. Future work will validate the quantitative findings on a large sample size and will further focus on providing the design implications from the user's perspective to make the financial RS explainable for users.

## References

[1] D. Zibriczky, Recommender systems meet finance: a literature review, in: Proc. 2nd Int. Workshop Personalization Recommender Syst, 2016, pp. 1–10.

[2] L. Klapper, A. Lusardi, P. Van Oudheusden, Financial literacy around the world, World Bank. Washington DC: World Bank (2015).

[3] S. Krishnan, S. Deo, N. Sontakke, Operationalizing algorithmic explainability in the context of risk profiling done by robo financial advisory apps (2020).

[4] E. B. Authority, Eba report on big data and advanced analytics, 2020. URL: https://www.eba.europa.eu/sites/default/documents/files/document_library/Final%20Report%20on%20Big%20Data%20and%20Advanced%20Analytics.pdf, accessed = 2022-07-28.

[5] D. Shin, User perceptions of algorithmic decisions in the personalized ai system: perceptual evaluation of fairness, accountability, transparency, and explainability, Journal of Broadcasting & Electronic Media 64 (2020) 541–565.

[6] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[7] S. Deo, N. S. Sontakke, Usability, user comprehension, and perceptions of explanations for complex decision support systems in finance: A robo-advisory use case, Computer 54 (2021) 38–48.

[8] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, User Modeling and User-Adapted Interaction 27 (2017) 393–444.

[9] T. Butler, L. O'Brien, Artificial intelligence for regulatory compliance: Are we there yet?, Journal of Financial Compliance 3 (2019) 44–59.

[10] D. Ben David, Y. S. Resheff, T. Tron, Explainable ai and adoption of financial algorithmic advisors: An experimental study, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 390–400.

[11] F. Gedikli, D. Jannach, M. Ge, How should i explain? a comparison of different explanation types for recommender systems, International Journal of Human-Computer Studies 72 (2014) 367–382.

[12] N. Tintarev, J. Masthoff, Explaining recommendations: Design and evaluation, in: Recommender systems handbook, Springer, 2015, pp. 353–382.

[13] N. Tintarev, J. Masthoff, Designing and evaluating explanations for recommender systems, in: Recommender systems handbook, Springer, 2011, pp. 479–510.

[14] M. Millecamp, S. Naveed, K. Verbert, J. Ziegler, To explain or not to explain: the effects of personal characteristics when explaining feature-based recommendations in different domains, in: Proceedings of the 6th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, volume 2450, CEUR; http://ceur-ws. org/Vol-2450/paper2. pdf, 2019, pp. 10–18.

[15] S. Chari, O. Seneviratne, D. M. Gruen, M. A. Foreman, A. K. Das, D. L. McGuinness, Explanation ontology: a model of explanations for user-centered ai, in: International Semantic Web Conference, Springer, 2020, pp. 228–243.

[16] L. Dymova, P. Sevastianov, K. Kaczmarek, A stock trading expert system based on the

rule-base evidential reasoning using level 2 quotes, Expert Systems with Applications 39 (2012) 7150–7157.

[17] F. Abraham, S. L. Schmukler, J. Tessada, Robo-advisors: Investing through machines, World Bank Research and Policy Briefs (2019).

[18] G. Babaei, P. Giudici, E. Raffinetti, Explainable artificial intelligence for crypto asset allocation, Finance Research Letters (2022) 102941.

[19] M. Schemmer, P. Hemmer, N. Kühl, S. Schäfer, Designing resilient ai-based robo-advisors: A prototype for real estate appraisal, in: 17th International Conference on Design Science Research in Information Systems and Technology, 1st-3rd June 2022, St. Petersburg, FL, USA, 2022.

[20] M. T. Ribeiro, S. Singh, 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.

[21] D. L. Morgan, Focus groups as qualitative research, volume 16, Sage publications, 1996.

[22] M. Bloor, Focus groups in social research, Sage, 2001.

[23] A. S. Acharya, A. Prakash, P. Saxena, A. Nigam, Sampling: Why and how of it, Indian Journal of Medical Specialties 4 (2013) 330–333.

[24] V. Braun, V. Clarke, Thematic analysis., American Psychological Association, 2012.

[25] J. Fereday, E. Muir-Cochrane, Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development, International journal of qualitative methods 5 (2006) 80–92.

[26] T. O. Nyumba, K. Wilson, C. J. Derrick, N. Mukherjee, The use of focus group discussion methodology: Insights from two decades of application in conservation, Methods in Ecology and evolution 9 (2018) 20–32.

[27] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, 2011, pp. 157–164.

[28] A. D. Madden, A definition of information, in: Aslib Proceedings, MCB UP Ltd, 2000.

[29] E. A. C. Bittner, J. M. Leimeister, Why shared understanding matters–engineering a collaboration process for shared understanding to improve collaboration effectiveness in heterogeneous teams, in: 2013 46th Hawaii International Conference on System Sciences, IEEE, 2013, pp. 106–114.

[30] J. Prümper, Software-evaluation based upon iso 9241 part 10, in: Vienna Conference on Human Computer Interaction, Springer, 1993, pp. 255–265.

[31] S. Naveed, An Interactive Hybrid Approach to Generate Explainable and Controllable Recommendations, Ph.D. thesis, University of Duisburg-Essen, 2021.

[32] S. Naveed, B. Loepp, J. Ziegler, On the use of feature-based collaborative explanations: An empirical comparison of explanation styles, in: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 226–232.

[33] S. Naveed, J. Ziegler, Featuristic: An interactive hybrid system for generating explainable recommendations–beyond system accuracy, system 18 (2020) 33.

[34] M. Eiband, D. Buschek, A. Kremer, H. Hussmann, The impact of placebic explanations on trust in intelligent systems, in: Extended abstracts of the 2019 CHI conference on human factors in computing systems, 2019, pp. 1–6.