HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230075

A Conceptual Model for Implementing Explainable AI by Design: Results of an Empirical Study

Martin VAN DEN BERG[a,1], Ouren KUIPER[a], Yvette VAN DER HAAS[a], Julie GERLINGS[b], Danielle SENT[a], and Stefan LEIJNEN[a]
[a]*Research Group Artificial Intelligence, HU University of Applied Sciences Utrecht*
[b]*Department of Digitalization, Copenhagen Business School*
ORCiD ID: Martin VAN DEN BERG https://orcid.org/0000-0003-3974-7374, Ouren KUIPER https://orcid.org/0000-0002-5033-6173, Yvette VAN DER HAAS https://orcid.org/0000-0002-0887-3979, Julie GERLINGS https://orcid.org/0000-0003-3776-5341, Danielle SENT https://orcid.org/0000-0002-4703-5345, Stefan LEIJNEN https://orcid.org/0000-0002-4411-649X

Abstract. Artificial Intelligence (AI) offers organizations unprecedented opportunities. However, one of the risks of using AI is that its outcomes and inner workings are not intelligible. In industries where trust is critical, such as healthcare and finance, explainable AI (XAI) is a necessity. However, the implementation of XAI is not straightforward, as it requires addressing both technical and social aspects. Previous studies on XAI primarily focused on either technical or social aspects and lacked a practical perspective. This study aims to empirically examine the XAI related aspects faced by developers, users, and managers of AI systems during the development process of the AI system. To this end, a multiple case study was conducted in two Dutch financial services companies using four use cases. Our findings reveal a wide range of aspects that must be considered during XAI implementation, which we grouped and integrated into a conceptual model. This model helps practitioners to make informed decisions when developing XAI. We argue that the diversity of aspects to consider necessitates an XAI "by design" approach, especially in high-risk use cases in industries where the stakes are high such as finance, public services, and healthcare. As such, the conceptual model offers a taxonomy for method engineering of XAI related methods, techniques, and tools.

Keywords. Explainable AI (XAI), explainability, financial services, conceptual model

1. Introduction

Artificial Intelligence (AI) nowadays has an increasing large impact on individuals and organizations. AI offers organizations unprecedented opportunities to automate, optimize, generate insights, and create human-like interactions [1,2]. Despite its advantages, AI also comes with multiple risks. One such risk is that outcomes and processes of AI systems are not intelligible for humans [1,3]. With the rise of new AI

[1] Corresponding Author: Martin van den Berg, martin.m.vandenberg@hu.nl.

techniques, such as neural networks, AI models have become increasingly complex and opaque, making it hard to determine how they operate. This, along with legal requirements for the right to an explanation [4,5] has led to the emergence of explainable AI (XAI) [3,6,7].

XAI focuses on generating explanations in a way that AI systems are transparent and understandable [3,6]. XAI is intended to increase trust and acceptance of AI systems among stakeholders such as customers, regulators, and users. This can be achieved by examining to what extent stakeholders understand the decisions of an AI system and by explicitly explaining the decisions to stakeholders [8]. The explainability of AI systems is seen as one of the building blocks of the responsible use of AI as it helps explain how an AI system works and, as such, supports the detection of bias, fairness, and discrimination [9,10]. Explainability of the outcomes and functioning of AI systems is particularly important in industries where trust plays a crucial role, such as healthcare and finance [11].

The implementation of XAI is not straightforward [12]. On the one hand, it requires knowledge and understanding of how AI systems work and, on the other hand, insight and understanding of the explanability requirements of the stakeholders. Both technical and social aspects deserve attention [13,14,15]. Technical aspects focus on integrating explainability into an AI system. Social aspects focus on integrating explanations into decision making processes and how to convey explanations to stakeholders. Different studies focus on either the technical or social integration aspects, and predominantly from a theoretical perspective [e.g., 16,17,18,19,20].

This study aims to empirically examine the XAI related aspects that different stakeholders of AI systems encounter during the development of the AI system. Aspects refer to the factors that require a decision to generate and provide meaningful explanations to stakeholders of AI systems. The focus of this study is the Dutch financial industry, leading to the following research question: *What XAI related aspects are considered in the development of financial services AI systems*?

This study presents a conceptual model that encompasses the XAI related categories of aspects that are considered in practice. These categories of aspects offer developers and managers of AI systems an understanding of the decisions that are necessary during the process of making these systems explainable and providing meaningful explanations.

The conceptual model is developed from use cases in the financial industry. This is a highly regulated industry with many high-risk AI use cases for which the explainability of an AI system is an important requirement. We argue that our model is also relevant for industries where the stakes are high, and trust plays a crucial role such as healthcare and public services. The upcoming EU AI Act recognizes different high risk AI systems from these industries where these systems must be able to provide information about the reasoning behind their decisions in a human-understandable form, so that users can understand how the system arrived at its conclusions [21].

This paper is organized as follows. In Section 2 we present related work. Section 3 contains the method. In Section 4 we present the conceptual model and in Section 5 we discuss its implications. Finally, Section 6 contains our conclusions.

2. Related Work

This section provides an overview of related work regarding the application of XAI in general and the application of XAI in the financial industry.

2.1. Application of XAI in General.

The number of XAI publications has exploded in two years from 186 papers in 2018 to 1505 papers in 2020 [7]. Many of these studies are theoretical in nature. Relatively few papers address the practical application of XAI. To the best of our knowledge, only Dhanorkar et al. offer an empirical understanding of explainability practices [22]. They found that explanations are iterative, interactive, and emergent, rather than a static quality of a model. Our study aims to increase the empirical understanding of how XAI is developed and used by studying the aspects that require a decision.

This study uses the following definition of XAI: "Given a stakeholder, *XAI* is a set of capabilities that produces an explanation (in the form of details, reasons, or underlying causes) to make the functioning and/or results of an AI solution sufficiently clear so that it is understandable to that stakeholder and addresses the stakeholder's concerns" [23]. This definition covers the technical and social aspects of XAI as discussed in Section 1.

To establish a foundation for this study, we derived eight groups of aspects of the application of XAI from the literature. Table 1 contains these groups in which a distinction is made between the use case and the use case transcending level. Most decisions are made on the use case level. This is the level at which the AI system is developed. However, certain decisions can be made on an organization-wide level. Therefore, we identified the "Overall XAI" category.

Table 1. Groups of XAI Related Aspects.

Group	Level	Meaning	Example question	References
Overall XAI	Use case trans-cending	General policies, principles, and ways of working on XAI	What are the principles for how to deal with explainability and XAI?	[18,24,25,26]
Transparency and explainability	Use case	Role and impact of transparency and explainability	What is the trade-off between the explainability and performance of the AI model?	[16,18,20,25, 28,29]
AI	Use case	Role and impact of AI	What is the purpose of AI in the use case?	[17,22,25,29, 30,31]
XAI system	Use case	Goal and approach of XAI system	What is the purpose of XAI in the use case?	[17,18,25,32, 33]
Stakeholder's needs for explanations	Use case	Stakeholders and their needs for explanations	What are possible scenarios to prompt explanations (e.g., understanding inner workings, anticipating user questions, details about data, model mechanics at a high level, and ensuring ethical considerations during model development)?	[16,17,18,22, 24,25,28,29,33]
Explanations	Use case	Why, what, and how to explain	What kind of information to provide as an explanation and to which stakeholders?	[17,18,19,20,24 25,26,27,34,35]
XAI methods and techniques	Use case	Methods and techniques to develop the XAI system	What technical method(s) to use to generate explanations (e.g., SHAP, LIME, Anchors)?	[10,16,17,18, 19,20,24,26,27, 29,36]
Methods and techniques to evaluate XAI	Use case	Methods and techniques to evaluate the XAI system	How to measure stakeholder satisfaction with the explanations provided (e.g., user engagement, Likert scale questionnaires, simulated experiments)?	[16,18,19,20, 24]

The following terms are used in Table 1 and the remainder of this study: An *AI system* is a system that is developed and used to decide or predict based on how it learned from data. An *XAI system* is a system that is developed and used to generate explanations from data and communicate these explanations to relevant stakeholders. Data for an XAI system can be retrieved directly from the AI system or indirectly by using additional post-hoc techniques such as SHAP [37] and LIME [38]. An *AI model* is a model that is developed and used in an AI system such as a decision tree, random forest, or neural network.

2.2. Application of XAI in the Financial Industry

Studies into an approach to how XAI can be implemented in financial services are scarce and often limited to single use cases. Bracke et al. present an explainability approach for predicting mortgage defaults using the quantitive input influence method [39]. Bussman et al. propose an XAI model that can be used in fintech risk management [40]. El Quadi et al. focused on companies' credit scoring and benchmarked different machine learning models supplemented by SHAP [41]. Misheva et al. applied LIME and SHAP to machine learning based credit scoring models [42]. Despite the importance of XAI for the financial industry, no studies discuss how explainability can be integrated into the development of AI systems.

3. Method

To find aspects that play a role in the implementation of the explainability of AI systems, we conducted a multiple case study in two financial services companies where we collected data from four use cases and developed a draft conceptual model of XAI related categories of aspects. The draft conceptual model was validated in two focus groups.

3.1. Development of Draft Conceptual Model

In the multiple case study, we focused on four use cases from two financial service providers. The first company is a Dutch fintech, offering business loans to small and medium-sized enterprises ("Lender" from now on). We researched their use cases on customer acceptance (CA) and customer review (CR). Company number two is a bank, with well over 3 million customers and a focus on payments, savings, and mortgages for retail customers, entrepreneurs, and small and medium-sized enterprises ("Bank" from now on). The two use cases from Bank are arrears management (AM) and personal finance in the mobile banking environment (PF). Bank and Lender provided documents for each of the use cases. The use cases were studied through semi-structured interviews with employees involved in the development of AI in these use cases. A total of 15 interviews were conducted. Table 2 contains an overview of the participants.

The main purpose of the interviews was to gain insight into the aspects for which decisions have been made or foreseen regarding XAI during the development of the AI system in the use case. Due to COVID-19 restrictions at the time, the interviews were conducted online in March-April 2022. Informed consent applied for all interviewees. Interviews were conducted by one researcher at a time and prepared with an interview guide. Each interview was structured around different topics: after an introduction, the interviewee was asked to explain the use case and the purpose and function of AI. As a

next step, the interviewee was asked to recall past and future decisions that were made regarding XAI. Then, to gain a more complete image of decisions related to XAI, further questions were asked about the use case, the AI system, the stakeholders of the AI system, the importance of explainability for the stakeholders, and the processes that the AI system supports. Participants were asked about the obstacles and learning points encountered during the implementation of XAI.

Table 2. Participants in interviews.

Nr	Company	Function	# of years function	# of years of work experience	Use cases
P1	Lender	Chief Technology Officer	6	17	CA, CR
P2	Lender	Senior Risk Manager	2	10	CR
P3	Lender	Head of Compliance	1	8	CA, CR
P4	Lender	Machine Learning Engineer	2	3	CA, CR
P5	Lender	Risk Manager	4	5	CA
P6	Lender	Head of Risk	4	17	CA, CR
P7	Bank	Senior Data Scientist	2,5	5	PF
P8	Bank	Senior Risk Manager	0,5	29	AM, PF
P9	Bank	Head of Risk and Compliance	0	25	AM, PF
P10	Bank	Data Engineer	6	26	PF
P11	Bank	Business Developer	1	17	AM
P12	Bank	Product Owner	0	5	PF
P13	Bank	CRM Marketeer	2	20	AM
P14	Bank	Data Scientist	2,5	2,5	AM
P15	Bank	Lead Modeller	2,5	14	AM, PF

The interviews were transcribed and hereafter coded in ATLAS.ti[2]. In the first round of analysis, we conducted a form of template analysis. We used a set of codes based on the eight groups in Table 1. As a result, we found 254 quotes in which explicitly or implicitly a certain decision was implied. In a second round, we gave these quotes more concrete codes to identify aspects. We found 46 XAI related aspects, i.e., 46 distinct types of decisions about XAI. We used these aspects to draft a conceptual model. The model was created in an iterative process in which multiple researchers participated. In the final draft model, the 46 aspects were clustered into 17 categories of aspects.

[2] Access to data can be requested from the first author.

3.2. Validation of Draft Conceptual Model

Two confirmatory focus group meetings were held to validate the draft conceptual model. Due to COVID-19 restrictions, the first focus group meeting was held online using MS Teams on June 7, 2022. The second focus group meeting was a physical meeting on July 5, 2022. Ten and nine people respectively participated in the two consecutive focus groups. On average, the participants of the first focus group had 14.2 years of relevant work experience and 4.0 years of knowledge and/or experience with XAI. The second focus group had on average 15.9 years of work experience and 3.9 years of knowledge and/or experience with XAI. The participants were representatives of financial services companies, regulators, government, academia, consultancy services companies, and data science services companies. The focus group meetings lasted two hours each. Each focus group meeting was audio recorded and transcribed.

The draft conceptual model was presented at the start of the meeting. The participants were then asked the following question: "To what extent does the model provide insight into what you need to do when you want to apply explainable AI?". To stimulate independent thinking, we asked the participants to individually brainstorm and write down their answers on paper before discussing these in the group. Then, the answers were discussed one by one. The focus group meetings led to the addition of one category of aspects to the conceptual model. The validated conceptual model will be presented and discussed in Section 4.

4. Results

4.1. Conceptual Model: An Overview

In this section, we present the conceptual model. This model, as shown in Figure 1, contains 18 categories of aspects (white boxes) that need consideration in the development of XAI.

The model consists of two levels: the organizational level and the use case level. The organizational level contains categories of aspects that require decisions at a use case transcending level. To govern and streamline the development of AI and XAI systems, guidance from the organizational level may be useful, especially in larger organizations.

The use case level contains categories of aspects that require decisions per use case in which AI is used. An example of a use case in this study is "customer acceptance". As depicted in Figure 1, most decisions occur at the use case level. This is the level at which AI systems and XAI systems are developed. Feedback from experiences at the use case level can lead to adjustments of policies, principles, and guidelines at the organizational level.

Importantly, the AI system and the XAI system are intertwined in that achieving requirements for the two systems cannot be considered independently. Decisions on the development of the AI system can have consequences for the XAI system and vice versa. The goal, stakeholder's requirements, and risks of the AI system can affect the goal and risks of the XAI system. On the other hand, the stakeholder's needs for explanations can impact the AI system, such as what type of AI model to choose.

We will now discuss the categories of aspects one by one starting at the organizational level.

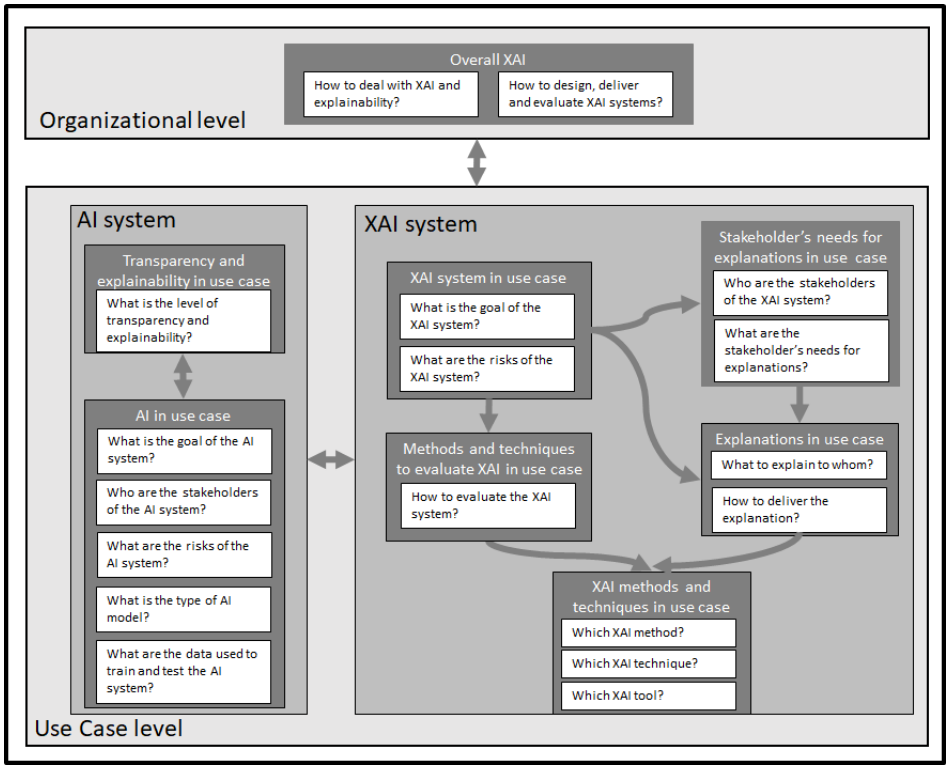

Figure 1. Conceptual model of categories of aspects and relationships relevant to the development of XAI. The white boxes contain the 18 categories of aspects and the dark grey boxes the eight groups from Table 1 (references for follow-up reading are included in Table 1). The arrows indicate the main relationships.

4.1.1. Organizational Level Categories

"How to deal with XAI and explainability?" is the first category to discuss. From the case study, it became clear that organizations may formulate corporate-wide XAI related policies and principles rooted in organizational values. Explainability of algorithms is one of the ethical principles of Bank. According to participant P11: '*One of our guiding principles is banking with a human touch. And if you look at society, we believe that ultimately, we want to create a society in which you can live with confidence and optimism, in which you guide people to prevent long-term financial scarcity. That guides the way we apply AI and XAI'.* Lender has a strong emphasis on simplicity. P2 says: '*One of our guiding principles is to keep it simple, thus explainable'*.

The second category is "How to design, deliver, and evaluate XAI systems". Especially in larger organizations such as Bank, particular decisions occur regularly across different use cases and for efficiency reasons it is beneficial to make these decisions once. Bank aims to embed XAI in its standard AI development lifecycle. P14 emphasizes: *'As a data scientist I prefer the embedding of XAI in our data science life cycle. By default, at least think or reflect on XAI at every step'*. P9 adds: *'Ultimately, we could integrate ethical aspects such as explainability into our product approval process'*. P8 has concrete ideas on how to integrate explainability in the product life cycle: '*What I would find important is that […] it is documented how explainability is fulfilled in that situation […] Who is responsible? Through which media? […] People must be instructed*

on how the explanation should be given'. In Lender the need for standardized ways of working on XAI begins to arise. P1: '*I think that as you grow as a company, you will feel more a need to write things down and record them'*.

4.1.2. Use Case Level Categories of AI System

The first category on the use case level and part of the AI system is "What is the level of transparency and explainability". This category deals with the required level of transparency and explainability of an AI model. The levels of transparency and explainability are important since they can affect the goal, stakeholder's requirements, and risks of the AI system as well as the type of AI model. Part of this category are the trade-offs with other requirements such as performance. Transparency about the use of AI in a business process is a sensitive topic that requires careful consideration as P12 argues: *'The customer expects that the choices we make are right. […] Banking matters are just very sensitive. So, even if you are very transparent, you cannot blame a machine'*. The trade-off between explainability and performance is a topic that requires careful consideration and is necessary in most use cases in the financial industry. When asking P10 if he will use a black box model if it has better performance, he answered: *'Should be substantially better. Otherwise, preference for a more interpretable model. And even then, we must see if we can't use that knowledge [from the black box model] to make a simpler model that is almost as good as the black box model'*. When asking P5 whether he prefers an AI model that performs better over an AI model that is fully interpretable, he replied: '*Yes, I prefer a higher performing model. However, you must be able to understand well which inputs enter the model and that it also makes sense what comes out*'. From a more managerial perspective, P1 argues: '*If the performance is relatively equal, we opt for the explainable model anyway. We would not readily adopt a model that we cannot explain because we know that if we want to [...] explain it to our customers, then we will have to be able to say something about the explainability'*. P7 agrees with that point of view: '*Usually, I prefer a more interpretable model that works a little less well'.*

The following category is "What is the goal of the AI system?". This category allows one to think about the purpose of using AI in a use case, thereby reasoning about how it affects stakeholders. P1 illustrates this as follows: '*Regarding customer acceptance, the business case for using AI is that we can automatically reject most of these customers. Most of all requests are rejected. And you want to spend as little time as possible on that process'*.

The category "Who are the stakeholders of the AI system?" is closely related to the previous category. Stakeholders of the AI system might also be the stakeholders in need of an explanation. So, knowing the stakeholders is important in the context of XAI and was mentioned often in the interviews. In the personal finance use case, the customer was mentioned by everybody as the main stakeholder. Additionally, marketing, the AI team, the business, and society were mentioned. Most often mentioned external stakeholders were the customer and the regulator. Internal stakeholders mentioned were diverse: underwriter, risk manager, risk and compliance manager, sales officer, marketing officer, data scientist, machine learning engineer, model owner, and senior manager. Each of these stakeholders may have different needs for explanations.

"What are the risks of the AI system?" is a category of aspects that allows one to assess the potential harm and ethical concerns associated with the use of the AI system. The higher the risks, e.g., in terms of violating human rights or loss of reputation, the

greater the need for XAI. P12 indicates: *'What is the risk for our customers and also for ourselves, reputation damage for example'*. According to P8: *'The application of AI touches on the bank's duty of care […]. To what extent is a false positive acceptable and did we explain that?'*. P13 sees XAI as a means *'To reduce risks'*.

Next is the category "What is the type of AI model?" which is the choice between types like decision tree, random forest, or a neural network. This category was frequently mentioned in the interviews and is deemed to be an important decision. P4 indicates: *'This is one of the most important things. You just need to be clear right away about what's going on here. […] We prefer to use something linear because context is extremely important. […] So preferably you use a model that is as simple as possible, because the moment you start using boosting or a neural network, for example, these models can become very specific and overfit much faster'*. According to P15, Bank is following developments in *'Advanced machine learning techniques'*. P14 indicates that explainability is an important requirement for the choice of type of AI model: *'The most important thing for retraining our current model will be that the model itself remains explainable'*.

The last category is "What are the data used to train and test the AI system?". One of the most frequently mentioned decisions in the interviews is what variables (features) to include in the AI model. When asked for the three most difficult decisions regarding explainability, P14 answered: *'From a data science perspective, which AI model you use, to what extent you weigh the performance of a model against its explainability, and which features to use'*. P13 finds the decision *'What data to include in the AI model'* the most difficult one. P5 indicates *'We are continuously looking for which variables we can add to the model to improve it'*. P3 adds to this: *'The regulator requires us to take certain risk factors such as geography risks into consideration'*. Although this category was mentioned in the interviews, we initially did not include it in the conceptual model. Based on comments from the first focus group, we added it. Data is the foundation for any AI system, and therefore also for an XAI system.

4.1.3. Use Case Level Categories of XAI System

The first category of the XAI system, as shown in Figure 1, is "What is the goal of the XAI system?". P4 states: '*I think a big part of machine learning adoption has to do with trust. The moment you can explain an outcome very well, and that explanation is correct, users will gain more confidence in it over time. I think that is very valuable'*. P15 indicates: *'Explainability is important. If we make a choice, whether someone gets a mortgage or not, it must be ethically responsible'*. These examples demonstrate that the purpose of XAI can vary, e.g., to increase trust or to justify decisions.

The next category is "What are the risks of the XAI system?". One of the risks mentioned by P5 is that *'If you give too much information, the system can be gamed. We must carefully consider what information we provide to our customers'*.

"Who are the stakeholders of the XAI system?" is a category very closely related to "Who are the stakeholders of the AI system?". The interviews showed no differences between the AI and XAI system in terms of who the stakeholders are.

The following category "What are the stakeholder's needs for explanations?" is the most frequently mentioned category across all interviews. Interviewees are aware that the development of an XAI system starts with the stakeholder's needs. In all four use cases, both internal and external stakeholders were considered as in need of explanations. These needs are diverse: feature importance, the accuracy of the prediction, insight into

what data is used and how, reasons for a specific prediction, insight into how the AI model can be improved, and insight into the inner workings of the AI system and AI model. When asked to what extent customers ask for explanations P1 answered: *'Customers don't ask for features, they don't understand that. I think the question is simply, why are you making this decision? And whether that decision is made by a person or by a model does not matter to customers. They just want to see a substantiation for that decision'*.

"What to explain and to whom?" is the category of aspects that addresses decisions regarding what information to provide and to which stakeholders. What to explain is the second most frequently mentioned aspect in the interviews. It emerged from the interviews that this is a tricky question, especially when it comes to the information that must be provided to customers. According to P13 *'We don't start the conversation [with customers] by providing information from the AI system. After all, the AI system might be wrong'*. P11 indicates: *'If you are going to explain to customers that outcomes come from a model and how it works, that is very difficult to explain. Neither should I tell the customer that he is a false positive or false negative'.* What to explain to internal stakeholders is a learning process. P4 explains: *'SHAP seems more suitable for the data scientist or machine learning engineer than for someone less familiar with it. To me the most important thing is that when you want to make your model explainable, it is also understandable for your end users. That's the most important choice you must make. We are now studying to make it even better. SHAP can be useful to explain model predictions or feature contribution. Much more is possible; we are now investigating that'*.

The next category is "How to deliver the explanation?". In all four use cases we studied, an internal person (human in the loop) was involved in providing customers with explanations. The automation of conveying explanations is still in an embryonic stage. P1 illustrates this: '*I think explanation is very much in the presentation. We should visualize the SHAP plot better. Because with SHAP you see features and they are horizontal. [...] That is too difficult. I think if we were to display the SHAP values on a scale of zero to hundred, as if it were a scorecard, it would be taken more seriously'*. Regarding the kind of language of an explanation, P13 argues: *'The why of a prediction must be clear cut in a language my mother-in-law also understands'*. P12 says: *'We try to avoid technical and banking terms, because it creates distance and people don't understand it any better'*.

The following categories deal with methods and techniques to develop XAI. We have divided this into three categories: methods, techniques, and tools. The interviews showed that decisions on these categories are still scarce but are foreseen for the near future. In terms of methods, decisions have been made for post-hoc local feature importance methods. Regarding techniques, SHAP is mostly mentioned, and regarding tools, the SHAP package is mentioned. P4 mentions the use of Shapley values and the SHAP package while P14 envisions standard XAI tooling to increase the number of different AI models available for selection: *'I prefer standard tooling. The moment we can use a model such as the XGBoost model, to which an explanation is immediately linked with that tooling, then it no longer stops us or at least less, to choose such a model'*.

The final category "How to evaluate the XAI system" covers the evaluation measures for the XAI system and the methods and techniques that are used. This category was not mentioned in the interviews. However, Bank and Lender remarked that this is one of the points of attention for the near future.

4.2. Validation of Conceptual Model

Two confirmatory focus groups were conducted to validate the conceptual model. As previously discussed in Section 4.1, the category "What are the data used to train and test the AI model" was added to the conceptual model based on the results of these focus groups. Overall, the participants expressed positive feedback and were enthusiastic about the model, with comments such as, *'I find the model quite complete', 'I appreciate the distinction between organizational and use case level',* and *'The model provides a good overview of aspects relevant in the implementation of XAI'.*

5. Discussion

This study highlights the need to consider a wide range of distinct aspects in the development of XAI. These aspects show that multiple decisions are needed to arrive at a situation where stakeholders of the AI system receive a proper and meaningful explanation. Furthermore, these aspects can interact with each other. E.g., the type of AI model impacts the way explanations are delivered. Moreover, the stakeholder's needs for explanations can affect the decision of what type of AI model to choose. The number of as well as the dependencies between aspects require XAI to be approached "by design". If an AI system is developed whose outcomes have a major impact on stakeholders such as customers, professionals, and regulators, it is crucial to consider what and how to explain to those stakeholders from the outset [43]. Our conceptual model can serve as a starting point to develop a methodology for XAI by design and as a taxonomy for method engineering of XAI related methods, techniques, and tools.

The conceptual model is derived from use cases in the financial industry, a highly regulated field where explainability of AI systems is an important requirement. We argue that our model is relevant for industries where decisions have a direct impact on individuals and trust is critical, such as healthcare and public services. The categories of aspects in our model will equally support developers of AI systems in these industries to make their systems explainable and provide stakeholders with meaningful explanations.

Dhanorkar et al. found that explanations are iterative, interactive, and emergent [22]. Translated to our model, this would mean that the categories of aspects that we identified do not have a fixed place in an AI development process, may occur multiple times, and can occur suddenly without developers of AI systems being aware of it. Our research also points in that direction. Although we asked about it in the interviews, it was challenging for interviewees to indicate when a certain decision was made during their AI development process. Further research is needed to determine when certain categories of aspects require a decision, such as how to integrate XAI related aspects into a standard AI lifecycle model like CRISP-ML(Q) [44]. Nonetheless, our conceptual model supports developers in being aware of what is needed to handle the explainability of their AI systems and create meaningful explanations.

The case studies illustrate preferences among interviewees. While some preferred interpretability and simplicity of the AI model over a more complex model with slightly better performance, another interviewee found that explainable outputs from SHAP helped stakeholders understand the outcomes and inner workings of the AI model. The case studies also showed that different XAI methods were more suitable for certain stakeholders than others, highlighting the need for further investigating which methods or frameworks are suitable for different stakeholders and under what circumstances.

The strength of the conceptual model is its completeness, as confirmed by the focus groups. Further research is needed to validate the completeness of the model in practice. Another strength of the model is that it is based on the experiences of practitioners. However, our conceptual model does not guide how to use it in practice. A process behind the model is lacking and requires future research.

This study has limitations. First, the conceptual model is based on four use cases from two organizations in the Netherlands. More research is needed to test the generalizability of the model. Second, two types of biases may occur in case studies: biases from the researcher's effects at the site and the researcher's data collection and analysis. We attempted to minimize the first type of bias by using an interview guide and informing the participants about the purpose of this research. The second type was counteracted by involving three researchers simultaneously in data collection and analysis and by using multiple sources of evidence (triangulation of data). In our case, we used data from documents, interviews, and focus groups.

6. Conclusion

This study presents a conceptual model of XAI related categories of aspects that must be considered in the development of AI systems. The model highlights the importance of approaching XAI "by design", integrating it into the broader AI development process and not treating it as an afterthought. This is particularly crucial in high-risk use cases and highly regulated industries such as healthcare, finance, and public services, where the intelligibility of an AI system can be as critical to its success as its predictive accuracy.

The model provides valuable insights for both practitioners and researchers in the field of XAI. For practitioners, it provides guidance on what aspects to consider in the development of explainable AI systems and how to make informed decisions. For researchers, the model serves as a starting point for further research and the development of XAI methodologies.

It is important to note that the conceptual model is based on data from use cases in the financial industry and more research is needed to validate its generalizability to other industries. Nevertheless, the model provides a comprehensive framework that helps practitioners and researchers alike to understand the complexities of XAI and the importance of approaching it "by design".

Acknowledgement
This research was financially supported by the Dutch Taskforce for Applied Research SIA under reference KIEM.K21.01.046.

Author statement

- Substantial contributions to the conception and design of the work: MvdB, OK, SL.
- Substantial contributions to data collection and analysis: MvdB, OK, YvdH.
- Wrote the first draft of the manuscript: MvdB, DS.
- Revised the manuscript: OK, YvdH, DS, JG, SL.
- Final approval of the version to be submitted for publication and agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: all authors.

References

[1] Benbya H, Davenport TH, Pachidi S. Artificial Intelligence in organizations: Current State and future opportunities. MIS Quarterly Executive. 2020 19(4) article 4, doi: 10.2139/ssrn.3741983

[2] Gartner. What is Artificial Intelligence (AI) Gartner. [cited 2023 Jan 17]. Available from https://www.gartner.com/en/topics/artificial-intelligence

[3] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020 58:82–115, doi: 10.1016/j.inffus.2019.12.012

[4] Kaminski ME. The right to explanation, explained. In: Sandeen SK, Rademacher C, Ohly A, editors. Research Handbook on Information Law and Governance. 2021 (pp. 278-299), doi: 10.4337/9781788119924.00024

[5] Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation". AI magazine 2017 38(3), 50-57 doi: 10.1609/aimag.v38i3.2741

[6] Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018 6:52138–60, doi: 10.1109/ACCESS.2018.2870052

[7] Islam MR, Ahmed MU, Barua S, Begum S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences. 2022 12(3):1353. doi: 10.3390/app12031353

[8] Miller T. Explanation in artificial intelligence: Insights from the Social Sciences. Artificial Intelligence. 2019 267:1–38, doi: 10.1016/j.artint.2018.07.007

[9] HLEG (The High-Level Expert Group on Artificial Intelligence). Ethics Guidelines for Trustworthy AI. EU Document. 2019 [cited 2023 Jan 15]. Available from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[10] Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Philosophical Studies Series. 2021 153–83, doi: 10.1007/978-3-030-81907-1_10

[11] McWaters, RJ. Navigating Uncharted Waters: A Roadmap to Responsible Innovation with AI in Financial Services: Part of the Future of Financial Services Series. 2019 World Economic Forum.

[12] Kuiper O, Van den Berg M, Van der Burgt J, Leijnen S. Exploring explainable AI in the financial sector: Perspectives of Banks and supervisory authorities. In: In: Leiva LA, Pruski C, Markovich R, Najjar A, Schommer C. editors. Artificial Intelligence and Machine Learning. BNAIC/Benelearn 2021. Communications in Computer and Information Science, vol 1530. Springer, Cham. 2021, doi: 10.1007/978-3-030-93842-0_6

[13] Bauer K, Hinz O, Van der Aalst W, Weinhardt C. EXPL(AI)n it to me – explainable AI and Information Systems Research. Business & Information Systems Engineering. 2021;63(2):79–82, doi: 10.1007/s12599-021-00683-2

[14] Liao QV, Varshney, Kush R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. ArXiv. 2021, doi:10.48550/arXiv.2110.10790

[15] Kemper J, Kolkman D. Transparent to whom? no algorithmic accountability without a critical audience. Information, Communication & Society. 2018 22(14):2081–96, doi: 10.1080/1369118x.2018.1477967

[16] Belle V, Papantonis I. Principles and practice of explainable machine learning. Frontiers in Big Data. 2021 4, doi: 10.3389/fdata.2021.688969

[17] Meske C, Bunde E, Schneider J, Gersch M. Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. Information Systems Management. 2020 39(1):53–63, doi: 10.1080/10580530.2020.1849465

[18] Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of Explainable AI Systems. ACM Transactions on Interactive Intelligent Systems. 2021 11(3-4):1–45, doi: 10.1145/3387166

[19] Schwalbe G, Finzel B. A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts. ArXiv. 2021, doi: 10.1007/s10618-022-00867-8

[20] Vilone G, Longo L. Explainable Artificial Intelligence: A Systematic Review. 2020, doi: 10.48550/arXiv.2006.00093

[21] European Commission. Regulatory framework proposal on artificial intelligence. [cited 2023 Jan 12]. Available from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[22] Dhanorkar S, Wolf CT, Qian K, Xu A, Popa L, Li Y. Who needs to know what, when?: Broadening the explainable ai (XAI) design space by looking at explanations across the AI lifecycle. Designing Interactive Systems Conference 2021. 2021, doi: 10.1145/3461778.3462131

[23] Van den Berg M. Kuiper OX. XAI in the Financial Sector. [cited 2023 Jan 15]. Available from https://www.internationalhu.com/research/projects/explainable-ai-in-the-financial-sector

[24] Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA. Four principles of Explainable Artificial Intelligence. 2020, doi: 10.6028/nist.ir.8312-draft
[25] Leslie D. Understanding artificial intelligence ethics and safety. arXiv preprint arXiv:1906.05684. 2019, doi: 10.5281/zenodo.3240529
[26] Jeyakumar JV, Noor J, Cheng YH, Garcia L, Srivastava M. How can I explain this to you? An empirical study of deep neural network explanation methods. Advances in Neural Information Processing Systems. 2020 33, 4211-4222
[27] Toreini E, Aitken M, Coopamootoo K, Elliott K, Zelaya CG, van Moorsel A. The relationship between trust in AI and Trustworthy Machine Learning Technologies. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020, doi: 10.1145/3351095.3372834
[28] Miller T, Howe P, Sonenberg L. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. 2017, doi: 10.48550/arXiv.1712.00547
[29] Koster O, Kosman R, Visser J. A checklist for Explainable AI in the insurance domain. Communications in Computer and Information Science. 2021 446–56, doi: 0.1007/978-3-030-85347-1_32
[30] Walz A, Firth-Butterfield K. Implementing ethics into artificial intelligence: a contribution, from a legal perspective to the development of an AI governance regime. Duke Law & Technology Review. 2019 18(1), 180-231.
[31] Zhou J, Chen F, Berry A, Reed M, Zhang S, Savage S. A survey on ethical principles of AI and implementations. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) 2020 (pp. 3010-3017). IEEE, doi: 10.1109/SSCI47803.2020.9308437
[32] Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies. 2021 146, 102551, doi: 10.1016/j.ijhcs.2020.102551
[33] Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI conference on human factors in computing systems. 2019 (pp. 1-15).
[34] Liao QV, Gruen D, Miller, S. Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems. 2020 (pp. 1-15), doi: 10.1145/3313831.3376590
[35] Georgieva I, Lazo C, Timan T, van Veenstra AF. From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience. AI and Ethics. 2022 1-15, doi: 10.1145/3290605.3300831
[36] John-Mathews JM. Critical empirical study on black-box explanations in AI. 2021 arXiv preprint arXiv:2109.15067. doi: 10.48550/arXiv.2109.15067
[37] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 2017, p 4768–4777
[38] Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. 2016, doi: 10.18653/v1/n16-3020
[39] Bracke P, Datta A, Jung C, Sen S. Machine learning explainability in finance: an application to default risk analysis. 2019 Staff Working Paper No. 816. London, United Kingdom: Bank of England. 2019, doi: 10.2139/ssrn.3435104
[40] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable AI in fintech risk management. Frontiers in Artificial Intelligence. 2020 3, 26, doi: 10.3389/frai.2020.00026
[41] Qadi AE, Diaz-Rodriguez N, Trocan M, Frossard T. Explaining credit risk scoring through feature contribution alignment with expert risk analysts. 2021 arXiv preprint arXiv:2103.08359, doi: 10.48550/arXiv.2103.08359
[42] Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin SF. Explainable AI in credit risk management. 2021 arXiv preprint arXiv:2103.00949, doi: 10.48550/arXiv.2103.00949
[43] Huynh TD, Tsakalakis N, Helal A, Stalla-Bourdillon S, Moreau L. Explainability-by-Design: A Methodology to Support Explanations in Decision-Making Systems. 2022 arXiv preprint arXiv:2206.06251, doi: 10.48550/arXiv.2206.06251
[44] Studer S, Bui TB, Drescher C, Hanuschkin A, Winkler L, Peters S, Müller KR. Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. Machine Learning and Knowledge Extraction. 2021 3(2), 392-413, doi: 10.3390/make3020020