LICENTIATE THESIS
IT FACULTY

Unraveling The Black Box

Building Understandable AI Through Strategic Explanation and User-based Design

Shuren Yu

DEPARTMENT OF APPLIED INFORMATION TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Unraveling The Black Box

THESIS FOR THE DEGREE OF LICENTIATE OF PHILOSOPHY

Unraveling The Black Box

Building Understandable AI Through Strategic Explanation and User-based Design

Shuren Yu

Department of Applied Information Technology
University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2024

Cover illustration: Hyper-threading Algorithm Rubik's Cube by Shuren Yu

Unraveling The Black Box

shuren.yu@ait.gu.se

Printed in Borås, Sweden 2024
Stema Specialtryck AB.

As long as we have correct epistemic ways, there is no black box.

Unraveling The Black Box

Building Understandable AI Through Strategic Explanation and User-based Design

Shuren Yu

Department of Applied Information Technology
University of Gothenburg
Göteborg, Sweden

ABSTRACT

The pervasive integration of Artificial Intelligence (AI) in society presents both opportunities and challenges, with the black-box issue emerging as a significant obstacle in realizing the full potential of AI. The opaque nature of AI decision-making processes impedes user understanding, particularly among non-technical individuals, raising concerns about the reliability of AI recommendations. Therefore, how to help users understand AI decision-making has become an urgent task. This thesis aims to assist developers in contemplating how to construct AI that users can understand. To build understandable AI, researchers have proposed many theories, methods, and frameworks in existing research. However, there are still limitations and challenges in current research. To address these challenges and finish the research aim, starting with a discussion on transparency and interpretability, the thesis elaborates on how to strategically explain to users within three dimensions: simplifying algorithm, appropriate information disclosure, and high-level collaboration. Furthermore, the thesis conducts surveys on users in four high-stakes areas, establishing AI explainability principles based on three stages, conceptualization, construction, and measurement. In addition to these primary contributions, the thesis also covers some supportive work, including challenges faced by explainable AI, user-centered development, and automation trust. These works lay a solid foundation for addressing research questions and achieving research objectives, while also providing room for contemplation in future research.

Keywords: understandable AI, transparency, interpretability, explainability strategy, high-stakes areas, user-based AI, XAI, automation-trust

LIST OF PAPERS

This thesis consists of a published paper and a paper that has been submitted to an international conference.

Paper A: Shuren Yu. Towards Trustworthy and Understandable AI: Unraveling Explainability Strategies on Simplifying Algorithms, Appropriate Information Disclosure, and High-level Collaboration.

In Proceedings of the 26th International Academic Mindtrek Conference, 2023, p.133-143. https://doi.org/10.1145/3616961.3616965

Paper B: Shuren Yu. An Empirical Investigation in High-stakes Areas: User-based AI Explainability Development Principles.

It has been submitted to the 32nd European Conference on Information Systems (ECIS 2024).

ACKNOWLEDGEMENTS

Although life is full of ups and downs and is never peaceful, I always move towards hope because I still have loved ones, family, close friends, and a career to strive for.

I am deeply grateful for the support of my family over the years, whether in the lows of life or moments of glory; they have always been by my side. Thanks to my wife, ***Sihan***, who is not only my best teacher but also my closest friend. She has taught me a lot and always brings the warmest care when I need help the most. I am also very thankful to my fathers and mothers for their support in my education and career. They have provided me with immense help, both financially and emotionally. Without the support of my family, I couldn't have overcome the challenges to reach where I am today. Although I know the road ahead may still be rough, as long as I have my family by my side, any difficulties can be easily overcome.

In my academic journey, I am very grateful to ***Jonas Ivarsson***, my former supervisor, for giving me the starting point for my doctoral research. Similarly, I appreciate my current supervisors, ***Thomas Hillman,*** and ***Alan Said***, for their careful and patient assistance, helping me complete my current research and guiding me toward the next phase. My manager, ***Tomas Lindroth,*** and department head, ***Helena Lindholm,*** provided strength when I needed help the most. My friends and Ph.D. fellows, ***Nadia Ruiz Bravo***, ***Bo Yang***, ***Jabbar Hussain***, and ***Bahram Salamat Ravandi***, inspired me with much academic inspiration. ***Pär Meiling*** and ***Karin G Pettersson*** always offered their assistance with a smile. Many people have given me warmth in the past journey, whether it's a word or an action, I always remember your kindness.

Thank you, grateful for your companionship. With you all by my side, I can find light even in the midst of thorns. Yesterday is gone, and tomorrow will be brighter.

CONTENT

1. INTRODUCTION 1
2. RESEARCH SCOPE AND QUESTIONS 5
3. TRANSPARENCY AND INTERPRETABILITY 7
3.1 Two important characteristics of black-box systems 7
3.2 The essence of transparency and interpretability 8
3.3 Building an understandable AI 9
4. CONSTRUCT STRATEGIC EXPLANATIONS 11
4.1 The goals of XAI 11
4.2 The technologies of XAI 12
4.3 Strategically explain to users 17
5. UNDERSTANDING ON USERS NEEDS AND WORK PRACTICES 21
5.1 The importance of user needs 21
5.2 Three stages for understanding users' needs 22
6. METHODOLOGY 25
6.1 Literature review and a concept-centric approach 25
6.2 Semi-structured interview with voice recording 26
7. CONTRIBUTIONS 27
8. THE FUTURE WORK 33
9. CONCLUSION 35
REFERENCES 39
APPENDED PAPERS 51
Paper A 51
Paper B 87
APPENDIX 121

1. INTRODUCTION

AI (Artificial Intelligence) is everywhere. Indeed, AI has become integrated into all facets of our society. Almost any industry that can utilize information technology has deployed and applied various AI products and models. AI systems provide decision-makers with plenty of predictions to assist their daily decision-making. However, AI's known black-box issue brings critical thinking to stakeholders about how to use these predictions correctly and to what extent to trust them. Blindly using these decisions may lead to potential hazards, especially in high-stakes areas such as healthcare and the judiciary. It can be said that ubiquitous opaque AI systems pose significant risks to privacy, responsibility, and justice to individuals and society at large (Pasquale, 2015). The black-box issue of AI has become a socio-technical challenge. Therefore, AI needs to assume more responsibility for its products and societal impacts and should have the ability to help people build an understanding of decisions.

Why does the black-box issue arise? Technology-centered AI has always been the main driving force for its development (Shneiderman, 2020). The pursuit of algorithm performance has become the main goal of AI developers. Every year, tens of thousands of developers are enthusiastic about designing and developing more advanced algorithms to improve their accuracy. However, the enhancement of algorithm performance will inevitably lead to an increase in its complexity. When complexity increases to a certain extent, even developers find it difficult to explain the reasons why AI makes decisions. All decision-making occurs in an opaque process, and we neither know how nor why. When these AI with opaque processes are integrated into people's work and lives, it will inevitably lead to decision risk. Try to imagine, when doctors diagnose patients based on decisions made by AI systems, if they do not know the reasons for the decision, should they adopt AI decisions and to what extent should they trust diagnostic recommendations? When the credit department of a bank uses AI to evaluate credit ratings, should bank staff determine whether loans should be issued based on the evaluation of AI if they are unaware of the decision-making process? Therefore, when decisions are crucial, people need to understand the process and reasons behind the decision-making, i.e., by understanding how and why AI should be used to what extent and decisions should be adopted.

It is difficult for people to understand the mechanism of black-box AI and the reasons for making decisions. To address this challenge, researchers from various fields have been discussing effective solutions or standardized

methods. Some researchers try to solve the black-box issue by technically enhancing algorithm transparency and interpretability. For example, developers can provide access to code, algorithms, and data to enhance their transparency and troubleshoot problems when they occur (Selbst et al., 2019). Developers can also annotate the code to enhance its interpretability. However, this does not fundamentally solve the black-box issue. These methods may be effective for developers. As users, many of them, such as doctors and judges, may not have the technical background, and therefore cannot rely on publicly available and annotated code to understand decisions. Therefore, considering users and specific scenarios, increasing transparency and interpretability should be endowed with deeper logic and thinking, rather than simply relying on technology solutions. Some researchers advocate solving the black-box issue based on XAI (eXplainable AI). XAI is a notion whose purpose is to help developers and users understand AI model behavior by many of the concepts and techniques (Gunning et al., 2019). XAI-based approaches, such as explainability framework (Sanneman and Shah, 2020), dimensions of explanations (Sperrle et al., 2020), and techniques (Singh et al., 2020), can help people understand the decision-making process of an AI system by constructing self-explanations of algorithms or providing additional explanation modules. However, these approaches have limitations in solving the black-box issue. Many of these approaches still stay at the theoretical stage due to the lack of testing in real environments (Dazeley et al., 2021), so "open the black box" may only be a "pay lip service" from AI researchers. Moreover, XAI-based approaches can achieve satisfactory results in certain specific scenarios while performing mediocrely in other environments. How to use these approaches requires consideration of specific data sources, algorithms, and user needs. Therefore, when designing understandable AI, AI developers should consider the applicability of methods and the combination of different frameworks, dimensions, and technologies to provide strategic explanations for decision-making. Some researchers argue a human-centered perspective is crucial. A human-centered perspective in AI development emphasizes prioritizing the needs and preferences of end-users throughout the design process. This involves understanding user requirements, incorporating user feedback, and creating AI systems that are intuitive and user-friendly. Some scholars also emphasize the participation of users in the AI development process (Kim et al., 2023) and the understanding of AI based on user perspectives (Jin et al., 2021). However, current methods and relevant literature indicate a lack of user-based research (Naiseh et al., 2020), including the collection of data on user perception, preferences, roles, experiences, and other aspects when using AI systems. Therefore, academia and industry should consider conducting more

extensive investigations to help AI developers design AI products that are understandable to users.

Based on the above discussion, AI developers should consider the following questions when designing AI that users can understand: How to understand transparency, interpretability, and their relationship? How to design strategical explanations based on XAI, including the integration of different frameworks, dimensions, technologies, and concepts? How to consider specific scenarios and user needs to provide personalized settings? How to consider using needs, experience, perception, and perspectives to design AI? How to verify these methods and principles in a real environment? These questions help us broaden our perspective on the main goal, how to design AI that users can understand. With reflections on these questions, we will initiate a series of discussions to implement our research.

2. RESEARCH SCOPE AND QUESTIONS

In this research, we aim to help AI developers think about how to design and develop AI that users can understand. The inspiration for this research comes from reflections on a series of issues related to transparency, interpretability, explainability, and user needs. When we think about these issues, we will face the following three challenges.

Challenge 1 (C1): When developers think about building understandable AI, increasing transparency (Kumar et al., 2020; Salahuddin et al., 2022), and improving interpretability (Markus et al., 2021; Kaur et al., 2021; Lyu et al., 2021; Li et al., 2022) can be seen as crucial approaches to help people understand AI decisions. However, there is still insufficient research on how to build understandable and even trustworthy AI for users through transparency and interpretability, as well as the relationship between the two.

Challenge 2 (C2): Based on the previous discussion, building understandable AI for users needs to think about how to design and implement strategic explanations. The concepts and technologies related to XAI have been widely discussed in various fields. However, research on how to conduct strategic explanations is still limited. More importantly, existing explainability approaches focus on AI developers rather than users (Bhatt et al., 2020). Therefore, it is necessary to study how to develop strategies to help users better understand the decisions of AI. This also involves verifying and implementing these strategies in real scenarios.

Challenge 3 (C3): Although there is a vast amount of research on the existing explainability, there is still a lack of extensive user investigations (Naiseh et al., 2020). Methods on explainability lack validation in real-life environments (Bruij et al., 2022; Jin et al., 2021; Khan et al., 2022), evaluation (Sperle et al., 2020; Markus et al., 2021), comparison (Kim et al., 2023; Dazeley et al., 2021), and trust calibration based on long-term observation (Naiseh et al., 2023). "Producing explanations that fully consider user contexts and tasks remains an understudied area" (Sanneman and Shah, 2020, p.107).

Based on the above description of challenges, this licentiate thesis will focus on three research questions:

- *Q1: How do we build AI that users can understand based on transparency and interpretability?*

- *Q2: How do we strategically explain AI decisions and help users better understand them?*
- *Q3: How do we design and develop AI that users can understand based on their perspectives, experiences, preferences, and satisfaction?*

To address Q1, this thesis will conduct a series of discussions. These discussions involve the importance of transparency and interpretability and the essence of them. Furthermore, this thesis will argue how to build understandable AI through transparency and interpretability. To address Q2, this thesis will carry out a comprehensive investigation of existing XAI-based explainability methods, technologies, and frameworks. In this survey (Paper A), we will first elaborate on the challenges that explaining AI decisions faces; Secondly, we will discuss how to build strategic explaination to users in three dimensions: simplifying algorithm, appropriate information disclosure, and high-level collaboration. Finally, our discussion on how to conduct explainability validation in real-world environments and how to consider user-centered AI will guide our future work. To address Q3, in paper B, we will first define three stages of explainability for users. Then, we will conduct interviews with users in four high-stakes areas: banking, education, healthcare, and justice. Compared to other fields, users in high-stakes areas need to know more about the reasons behind AI decisions, as decisions can be fatal. The purpose of the study in Paper B is to elucidate how to build understandable AI for users by analyzing their needs, satisfaction, and perspectives. Finally, we will conceptualize use-based AI design principles, that will navigate developers in designing and developing understandable and even trustworthy AI products for users. Additionally, we will also emphasize the necessity of validation in real environments.

To carry out this research, the rest of the thesis is structured as follows: Firstly, Section 3 argues the essence of transparency and interpretability and how to build an understandable AI based on them. Section 4 elaborates on how to construct strategic explanations, including the goals and technologies of XA, and how to strategically explain them to users. In Section 5, we can understand the importance of users' needs and the three stages of understanding users' needs. Section 6 describes the methodology of the thesis. The contributions, future work, and conclusion are showcased in Sections 7, 8, and 9 separately.

3. TRANSPARENCY AND INTERPRETABILITY

Based on the above discussion and reflection on Q1, the following content will respond to how to build understandable AI for users through discussions on transparency and interpretability. The following content includes arguments on the essence of transparency and interpretability and the construction of understandable AI. These contents will serve as guidance for addressing C1.

3.1 Two important characteristics of black-box systems

The black-box AI has been widely described and defined by academia and industry. In some intuitive descriptions, the black box is associated with "not understanding" (Lipton, 2018) and "not observation" (Bucher, 2016), that is, people cannot know and see the internal logic and mechanism of the black box. The earliest statements about the black box issue probably derived from Zadeh and Ashby.

Zadeh (1954) elaborated on the black box issue in his *System Theory*. He argued that any social organization, group, or complex computational structure is a system. Some elements without a specific physical identity in these systems form a black box. A black box in a system is similar to subordination processes in circuit theory. Although there is a functional relationship between inputs and outputs, such relationships are agnostic or not observable in a system with the black box. In *An Introduction to Cybernetics*, Ashby (1956) systematically introduced the black-box issue in engineering. An electrical worker is only able to deduce the internal structure of the operation box by observing its input and output voltages. In one experiment, the values of inputs (α, β) and outputs (x, y), as well as the parameters of the system, are determined. However, how α, β affect x, y within a black box, and how x and y are interrelated can involve an infinite number of possibilities. For people, it is impossible to see the mechanism inside the black box.

Therefore, a black box can be such a system, "an opaque technical device of which only the inputs and outputs are known" (Bucher, 2016, p.83). As a black box user, people do not see what happens (no transparency) inside the black box to cause a specific output, nor does he/she understand how the

input affects the output (no interpretation). The two important characteristics of a black box system are no transparency or/and no interpretation.

3.2 The essence of transparency and interpretability

Although transparency and interpretability are two key factors in opening the black box, improvements of them without aims cannot be directly used to build understandable and trustworthy AI.

Transparency refers to the possibility that an AI system can be investigated (Durán & Jongsma, 2021). According to Durán & Jongsma, if an AI model "A" is transparent, it can show users its structure and the relationships among variables and outputs, which gives users reason to think that "A" will provide reliable output. Transparency should, therefore, be linked to an objective representation of what is inside model "A", because "transparency is meant to refer to an inherent property of a model" (Mencar & Fanelli, 2008, p.4586). For any AI model, higher transparency means that the inner logic and the relationships between variables and outputs are more easily seen by users and vice versa. Although Durán and Jongsma define transparency as an "epistemic manoeuvre" (p.330), such "manoeuvre" should be embodied by the AI models themselves, not based on a user's cognition and comprehension.

Interpretation involves describing the interior of a system in ways that the user understands (Gilpin et al., 2018). Some simple structural AI models can achieve ante-hoc interpretation, such as Naive Bayes, linear regression, and decision trees, which are easy to understand because of their high transparency (Antoniadi et al., 2021; Lisboa et al., 2021). Whilst other models with complex structures, such as deep neural networks, need to use post-hoc interpretation techniques to achieve users' comprehensibility (Lisboa et al., 2021). The purpose of these techniques is to generate an approximation of the original model (Antoniadi et al., 2021) and to make an understandable explanation for users about how an AI decision was made (Moradi & Samwald, 2021). To an AI model, ante-hoc interpretation can be regarded as interpretability and post-hoc as explainability (Lisboa et al., 2021). Whether the interpretation is ante-hoc or post-hoc, however, how much a user can understand an AI decision depends on the user's cognition, expertise, knowledge, and comprehension, not the AI model’s complexity. Therefore, interpretation is "an inherently subjective matter" (Nguyen & Martínez, 2020, p.1) and is based on "contextual" (Miller, 2019, p.3). Interpretation should be different for those who are not at the same

knowledge level (Arya et al., 2019) because explainees have different understandings on an AI model's interpretation.

Transparency is an objective concept and closely related to the complexity of the model. Complex models have lower transparency, and vice versa. And interpretability is based on the user's cognition, experience, and background, and is influenced by the user's subjective factors (Verhagen et al., 2021). A completely transparent model means that it presents all the complex links and relationships between parameters within the model to the user. However, if the user lacks the ability (knowledge, experience, and background) to interpret these links and relationships, this completely transparent model still has very low interpretability. Compared to ordinary users, AI developers or researchers can make models highly interpretable. If we want the model to have high interpretability for ordinary users or even those who do not have professional AI knowledge, we can either improve their understanding of AI through certain strategies or design AI decision-making mechanisms that conform to their cognition and usage behavior. That's why researchers should focus on strategically explaining (when, how, which, etc.) and conduct user investigations in real-world environments when developing AI that users can understand. Therefore, if AI developers can improve the transparency of AI while also ensuring its design aligns with users' experiences, knowledge, and cognition, the black-box issue of AI is likely to be resolved to some extent.

3.3 Building an understandable AI

"Transparency can make incomprehensible systems interpretable, and explainability can make interpretable systems understandable" (Verhagen et al., 2021, p.9). Transparency can contribute to explanation (Bellucci et al., 2021) because people can see the internal connection and the relationship of parameters. This can help a black-box system become an interpretable one. At this time, the system cannot be understood by non-professional AI users, so it is still a black-box system for these users. Similarly, the system can become a system with interpretability if the user is a person with AI professional knowledge such as an AI developer or researcher. At this point, such a system with interpretability is an understandable one for these specific users. As we discussed in the previous section, understanding depends on subjective factors such as users' knowledge, experience, and cognition. If we want to make an interpretable system understandable to non-professional AI users, AI developers can consider two options: using explainability approaches strategically to explain the system. By telling users specific information such as when, how, and which indicators affect the decision,

users may understand the reasons for the decisions. Currently, users have reason to judge whether to use the decisions. At this point, such a system has explainability; Improving the system by analyzing the actual needs, satisfaction, experience, and even work practices of users. If the decision-making mechanism of the system tends to be consistent with the behavior and thinking logic of users, the system also has interpretability. Regardless of the options adopted by AI developers, the system will become an understandable one for these non-professional AI users. In addition, AI systems are more trustworthy if they have "understandable operations and outputs" (Emaminejad & Akhavian, 2022, p.11). Figure 1 depicts the process discussed above.

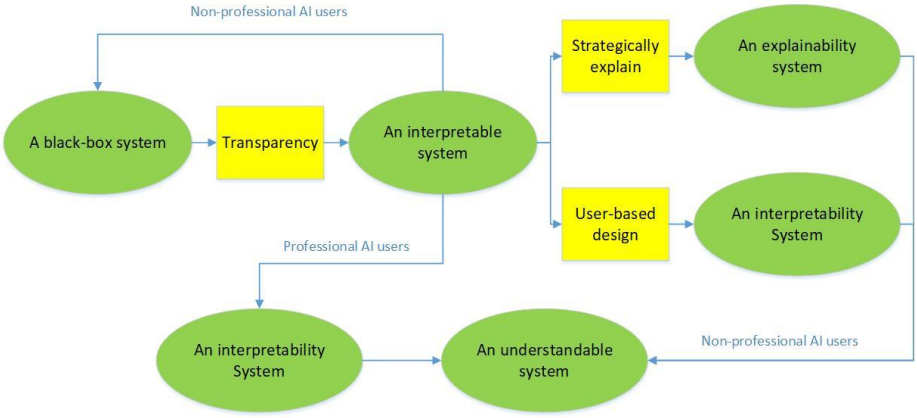


Figure 1. Building an understandable system. The flowchart in the figure shows that when a black box system has transparency, it becomes an interpretability one to professional users and still has a black box to non-professional users. When we provide strategic explanations to non-professional users, this interpretable system will be explainability to them; Also, when we improve this interpretable system based on users, it will also have interpretability. Finally, the interpretability and explainability system will become an understandable one.

4. CONSTRUCT STRATEGIC EXPLANATIONS

In the previous section, we discussed the essence of transparency and interpretability and the relationship between some related concepts. These discussions emphasize that one effective way for AI developers to establish understandable AI for ordinary users is to strategically explain AI systems. In order to face C2 and answer Q2, we first introduce the relevant goals and technologies of XAI in this section; Then we will introduce how to strategically explain to users based on XAI.

The XAI is put forward based on some significant concepts, such as explanations, interpretability, and intelligibility (Gunning et al., 2019). XAI refers to many goals and technologies. They can help users understand the behavior of AI models (Gunning et al., 2019). XAI aims to ensure users can understand, appropriately trust, and effectively manage intelligent systems (Gunning, 2017; Gunning et al., 2021; Adadi & Berrada, 2018; Lim et al., 2019; Meske & Bunde, 2020).

4.1 The goals of XAI

Based on the existing literature, Arrieta et al. (2020) synthesized and enumerated that the goal of XAI includes the following elements: trustworthiness, causality, transferability, informativeness, confidence, fairness, accessibility, interactivity, and privacy awareness.

- Trustworthiness refers to that when the AI makes a decision, it has good robustness (Floridi, 2019) and can produce “the confidence of whether a model will act as intended” (Arrieta et al., 2020, p.86).
- Causality describes that AI models reveal the correlation between data rather than causality (Marcus, 2018). XAI could "validate the results provided by causality inference techniques or provide a first intuition of possible causal relationships within the available data" (Arrieta et al., 2020, p.86).
- Transferability defines the ability to help users understand how to reuse the knowledge of an AI model for other scenarios (Lötsch et al., 2021).
- Informativeness shows that XAI should provide information about the problems being solved by the model, and its purpose is to "be

able to relate the user's decision to the solution given by the model, and to avoid falling into misconception pitfalls" (Arrieta et al., 2020, p.86).

- Confidence is about "an explainable model should contain information about the confidence of its working regime" (Arrieta et al., 2020, p.86).
- Fairness is the idea that, from a societal perspective, models can provide explanations that can guarantee fair decision-making, and models enable ethical analysis that identifies and reduces bias (Arrieta et al., 2020).
- Accessibility involves when constructing a model, users can be more involved in the process of model construction “without going through AI engineer’s interface” (Chander et al., 2018).
- Interactivity means that the end user can exert their influence on the model, and AI developers can use users' feedback to correct errors and make AI decisions consistent with users'. (Guo et al., 2022).
- Privacy awareness refers to XAI should avoid disclosure of private information inside the model by unauthorized explanations (Arrieta et al., 2020).

4.2 The technologies of XAI

The existing XAI technologies have many taxonomic definitions. Generally speaking, the classification of these technologies is not fixed and absolute. It can be divided into many overlapping or non-overlapping classes according to the characteristics of these technologies (Singh et al., 2020). Singh et al. summarized different types of classifications in Figure 2. According to Singh et al., model-agnostic, and model-specific refer to whether an explanatory technology is related to a specific model architecture. Global and local means whether an explanation is for the overall behavior of a model or a single result. Pre-model, in-model, and post-model are when an explanation occurs. The pre-model methods are independent and do not depend on the specific model architecture; The explanatory methods integrated into the model are in-model methods; The explanations implemented after the model establishment are post-model methods. The surrogate methods use the approximation of the explained model to achieve an explanation, while visualization methods explain some parts of the model through visual understanding such as activation maps.

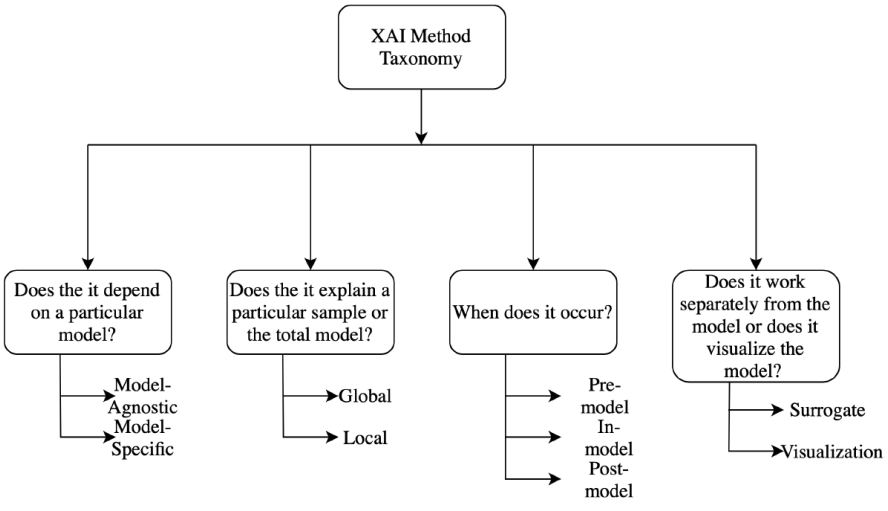


Figure 2. Taxonomy of XAI technologies (Singh et al., 2020, p.3)

According to the complexity of AI models and "when does it occur?", the technologies of XAI can be generally divided into two classes: ante-hoc and post-hoc. Ante-hoc methods refer to an AI model with a simple training structure and good interpretability, while post-hoc methods explain an AI model through explanatory methods.

Ante-hoc

The ante-hoc refers to the interpretability built into a model itself. To a trained AI model, the model's decision-making process or causality between input and output can be understood without additional information. The interpretability of the model occurred prior to model training. The ante-hoc has three categories: self-explanatory models, generalized additive models, and attention mechanisms.

- *Self-explanatory models* should be transparent. The decomposability of self-explanatory models requires that every part of the model, including model structure, model parameters, every input, and every feature in different dimensions, allow intuitive understanding without additional tools (Arrieta et al., 2020). However, the simple structure of self-explanatory models has no advantage in performing overly complex tasks. If the complexity is increased to improve the accuracy

and fitting ability, self-explanatory models might lose their interpretability. For example, a deep decision tree would make it unintelligible to users. Therefore, self-explanatory models require a trade-off between performance and interpretability (Rudin, 2019).

- *Generalized additive models*, as a compromise (the trade-off between performance and interpretability), can not only improve the accuracy of simple linear models but also retain their built-in interpretability of them.
- The *attention mechanism* is a way to help neural networks achieve self-explanation. The attention mechanism comes from cognitive neuroscience (Corbetta & Shulman, 2002). The human brain's attention focuses on salient features and ignores unimportant noises (Xu et al., 2015). An example is introduced in Xu et al.'s research on the attention mechanism. They used a convolutional neural network to extract the features of pictures and a recurrent neural network with the attention mechanism to generate descriptions of these pictures. In this process, attention realized the matching between words in these descriptions and pictures. People can see the corresponding word according to the interest area in these pictures.

Post-hoc

Post hoc takes place after model training. It aims to explain the working mechanism and decision-making behavior of the model by using explanatory methods. According to the summary from Arrieta et al., three types of post-hoc will be introduced.

- Focus on rules extraction of models. Ribeiro et al. (2016) proposed a model Local Interpretable Model-Agnostic Explanations (LIME). Specifically, for an input, LIME uses a linear regression model to fit the local boundary of this input in a model. Based on the linear model, LIME gives the reason why the decision of this input was made. The different coefficient of the linear model is used to reflect the importance of different features of this input. Figure 3 is LIME's explanation of a classification model for breast cancer, and it clearly shows why the model makes the decision, 'malignant,' for an input case. Moreover, Guidotti et al. (2019) and Ribeiro et al. (2018) respectively proposed Local Rule-based Explanations (LORE) and the 'Anchors' explanation technique, which are based on rules extraction to a model. However, LIME, LORE, and Anchors assume

that the features of input are independent of each other. They cannot accurately explain the model that has dependency between features, such as recurrent neural networks (RNN). Therefore, Guo et al. (2018) introduced Local Explanation Method using Nonlinear Approximation (LEMNA), which can be well applied to the explanation of RNN models.

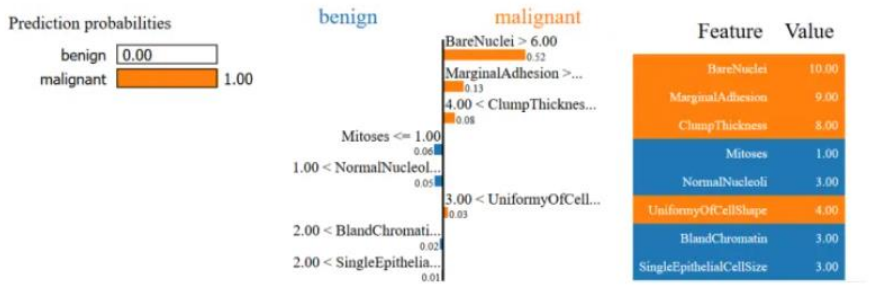


Figure 3. For an input case, the left part shows that the decision made by the model is 'malignant.' The middle part shows the importance of different features of the input to the decision. Compared with other features, 'Bare Nuclei,' 'Marginal Adhesion,' and 'Clump Thickness' make more contributions to decision-making because the weights of their coefficients are 0.52, 0.13, and 0.08 separately. The right part is the prediction value of different features of this input. (https://zhuanlan.zhihu.com/p/273754347)

- Focus on the Shapley value of features in a model. According to Molnar (2020), the Shapley value is a method from coalitional game theory that tells us how to fairly distribute the "total payout" among features. Shapley value assumes that each feature of the instance is a "player" in the game, where the prediction is "total payout". "The concept which is key here is to be able to form 'coalitions' (or subsets) of players in order to measure the performance of each player in every possible team situation" (Heuillet et al., 2022, p.62). The example of Figure 4 shows the basic logic of the Shapley value. Based on a bicycle rental dataset, the model predicted that 2409 bicycles would be leased on the 285th day. The actual prediction value of 2409 of the model is 2108 less than the average prediction value of 4518. The most negative affections are from the weather conditions and humidity, and the temperature of this day has the most positive contribution. The sum of Shapley values of all features yielded the difference between the actual and average prediction, -2108. Because the Shapley value is based on game theory and has

fair distribution effects under different feature coalitions, it might be more in line with legal requirements (Molnar, 2020).

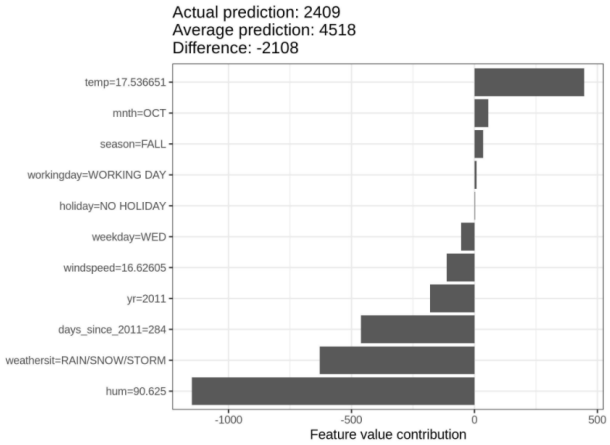


Figure 4. Shapley values of all features for a model based on a bicycle rental dataset (Molnar, 2020, Chapter 9.5.2, https://christophm.github.io/interpretable-ml-book/shapley.html)

- Focus on the visualization and heat map. Visualization is mainly applied to the interpretation of image classification models, such as medical image analysis (Rim et al., 2021; Nazari et al., 2021). The purpose of visualization is to show the importance of different image features in the decision-making of classification models. Many visualization methods are based on the backpropagation of neural networks, such as Integrated Gradients (Sundararajan et al., 2016). Although these methods can locate the important features used for decision-making to an input sample, they cannot quantify the contribution of each feature. Thus, Bach et al. (2015) proposed Layer-wise Relevance Propagation (LRP) that can quantify the contribution of a single pixel to decision-making. According to Singh et al. (2020), generally, pixels that contribute positively to a decision are marked red, while those that contribute negatively are blue. Other similar methods are based on Class Activation Mapping (CAM) (Zhou et al., 2016), such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017), which can well locate the important areas that affect decision-making. However, as Grad-CAM does not show the importance of fine-grained features, Selvaraju et al. (2017) proposed Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM).

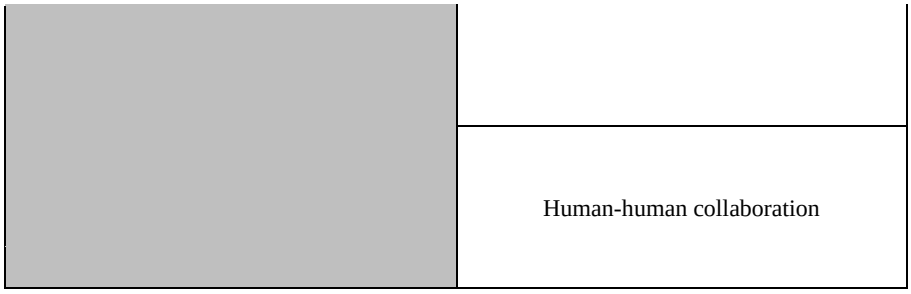
To sum up, XAI provides many concepts and goals that can be used as guidance for building understandable AI systems. At the same time, XAI also establishes a relatively complete technical system to provide explanations for AI models, whether it is ante-hoc or post-hoc.

4.3 Strategically explain to users

Explainable AI (XAI) aims to provide approaches to make AI systems more understandable for users. However, current practices often focus on developers and engineers rather than end-users (Bhatt et al., 2020). To take understandability to users, the thesis proposes three key dimensions and corresponding strategies for constructing effective explainability strategies (Table 1).

Table 1. Explainability strategies on three dimensions

Three dimensions	Explainability strategy
Simplifying algorithm	Semantic explanation
	Architecture explanation
	Global explanation and local explanation
	Causal explanation and interactive explanation
Appropriate information disclosure	Professional explanation
	Legal explanation
	AI supplier explanation
High-level collaboration	Human-machine collaboration



The first dimension advocates for simplifying algorithms without compromising decision criteria. Complex algorithms, while effective, may be nearly impossible for non-technical users to comprehend fully. Simplifying these algorithms without diminishing decision criteria can enhance user understanding. The second dimension involves appropriate information disclosure. Recognizing the need to balance business secrecy and transparency, we suggest disclosing key information such as summary results and benchmarks. This disclosure can address information asymmetry issues between AI companies and users, promoting better algorithm understanding. The third dimension emphasizes high-level collaboration between humans and AI through interactive machine learning. Engaging users in the learning process fosters understanding and collaboration. Explanation should be viewed as a social process, highlighting the importance of effective communication and collaboration among different stakeholders in the AI environment to enhance the overall understanding of AI systems.

Based on our study (Paper A), explainability strategies for simplifying algorithms encompass six approaches: Semantic, Architectural, Global, Local, Causal, and Interactive explanations. Causal, interactive, and semantic explanations aim to make AI mechanisms understandable for users, while local and global explanations are geared toward system developers. The efficacy of architectural explanation for user understanding, especially for non-technical individuals, requires further validation. In terms of information disclosure, three strategies include Professional, Legal, and AI Supplier explanations. Professional explanations evaluate and supervise content from legal and AI supplier explanations. Legal explanations offer judicial supervision schemes, and AI supplier explanations disclose algorithm information. High-level collaboration in explainability strategies involves human-machine and human-human interactions. Human-machine interaction focuses on information exchange within an AI system, while human-human interaction centers on collaboration among stakeholders across diverse domains and disciplines in AI-centered systems. However, it should be noted

that these strategies need to be further validated in real environments, which can be seen as a future research direction.

These three dimensions provide a comprehensive perspective for developers and scholars to construct explainability strategies, ultimately facilitating better user comprehension of AI decisions.

5. UNDERSTANDING ON USERS NEEDS AND WORK PRACTICES

Reviewing the discussion in section 1.1.2 and Figure 1, if AI developers want to design AI that users can understand, it is necessary for them to conduct user surveys. The purpose of doing so is to design the system based on their needs, satisfaction, preferences, and work practices. Moreover, user feedback after using the system will further assist AI developers in improving the system. This process can lead to iterations of multiple improvement systems, with the main goal of enabling users to understand the AI decision-making process and determine whether to use some or all of the AI features.

5.1 The importance of user needs

As discussed repeatedly in previous sections, the development of XAI represents a crucial aspect of advancing AI technologies, with a primary focus on ensuring not only accuracy but also trustworthiness in AI systems. Despite the strides made in XAI development, a prevailing challenge lies in its predominantly algorithmic orientation, often tailored for technical users, neglecting the broader audience of end-users without a technical background (Jin et al., 2021). Recognizing the potential for harmful unintended consequences, it is imperative for XAI to inherently consider its end-users, highlighting a fundamental issue: the substantial gap between the highly technical nature of XAI development and the diverse range of end-users who will engage with AI systems.

A pivotal perspective emphasizes that AI systems should encompass end-user values throughout the entire AI development lifecycle, demanding a nuanced understanding of the needs, preferences, and expectations of end-users (Bond et al., 2019). In particular, those users who lack technical expertise should be paid more attention. Achieving this understanding is essential to foster trust and reliance on AI systems. The presentation of explainability should be intuitive and easily comprehensible, enabling users to comprehend and trust the functioning of AI systems. Understanding user needs goes beyond merely addressing technical proficiency; it involves delving into their cognitive processes and the diverse contexts in which they engage with AI. This multifaceted understanding is essential for tailoring XAI systems to meet the specific requirements of different user groups. Research efforts should explore the intricacies of user needs, considering factors such as decision-

making processes, information assimilation, and contextual variables influencing user interactions with AI systems.

To enable non-technical users to effectively engage with XAI, there is an inherent need for comprehensive education and training initiatives. These initiatives should focus on imparting fundamental concepts of XAI, including model transparency, explainability techniques, and the benefits and limitations of AI systems. By enhancing users' understanding of these concepts, initiatives can empower users to make informed decisions when interacting with AI tools. This, in turn, promotes and ensures that users can navigate the evolving landscape of AI technologies with confidence. Furthermore, it contributes to the democratization of the discourse surrounding AI, allowing a wider audience to participate in discussions about the societal impacts and ethical considerations associated with AI (Garvey, 2018). Even, by fostering a user-centric approach, XAI can become a tool that is not only technically proficient but also aligns with the diverse expectations and preferences of the end-users it serves.

5.2 Three stages for understanding users' needs

Understanding user needs is important and necessary. However, user needs should not be a broad concept but should be described in detail and accurately throughout the entire process of AI operation. By defining the three stages involved in explainability, conceptualization, construction, and measurement (Paper B), we attempt to illustrate the three stages in which user needs occur in AI systems, which can also match the working practices of users.

- Conceptualization: The integration of AI into various workflows necessitates a thorough understanding among users before operational deployment. Users' perceived ease of use and usefulness significantly influence their acceptance of AI technologies (Eze et al., 2021). Ensuring users possess accurate information about AI system development, potential pitfalls, and limitations is crucial. Educating users preemptively helps in reducing overgeneralization and unintended use. Furthermore, transparency in sharing technical knowledge by AI developers fosters trust and reduces ethical risks associated with AI systems （ACM，2018）.

- Construction: Explainability construction in AI systems extends beyond technical methods, emphasizing the importance of stakeholder participation and co-creation for user understanding. Explainability methods must align with specific application domains, recognizing unique operational requirements and interpretability needs. For instance, in healthcare, IF-THEN rules may be suitable for diagnosis explanations (El-Sappagh et al., 2018). The choice of explainability methods should adhere to domain-specific legal and ethical guidelines (Arrieta et al., 2020), considering the diversity of users' technical expertise (Mohseni et al., 2021). Personalized explainability methods, based on users' knowledge structure and level, contribute to better user understanding, surpassing the limitations of unified standards.
- Measurement: User involvement in evaluating AI system explainability in high-stakes areas is crucial for continuous enhancement. Users play a pivotal role in regularly evaluating AI model outputs to ensure comprehensible, relevant, and coherent explanations. Their domain-specific knowledge is valuable in gauging the adequacy of explanations in practical scenarios. User feedback is indispensable for addressing deficiencies in AI explainability, improving the intelligibility and precision of explanations. Active user participation is necessary to document and report observed disparities or biases in AI explanations, facilitating adjustments (Chu & Shen, 2022). The involvement of domain experts and ethicists is recommended to enhance user interpretation of AI explanations, contributing to a thorough assessment of AI explainability. Additionally, users also need to define their role in the measurement stage.

6. METHODOLOGY

In section 1.1 of this thesis, we responded to the Q1 and discussed how AI developers should design AI that users can understand based on transparency and interpretability. For addressing Q2 and Q3, this thesis adopted the following two methods: literature review and semi-structured interview.

6.1 Literature review and a concept-centric approach

Based on Figure 1 and related discussions, it is necessary for researchers to discuss how to carry out strategic explanations to bridge the cognitive gap between professional AI technology knowledge and users, especially those without technical background. Extensive discussions have been conducted on XAI-based technologies and concepts in an attempt to uncover the veil of the black box. However, research on explainability strategies for users' understanding is still limited. Therefore, in Paper A, we classified existing literature by three dimensions and ultimately discussed how to strategically explain and construct AI that users can understand.

Conducting the literature review for Paper A will adhere to the comprehensive eight-step systematic review guidance (Okoli & Schabram, 2010). Furthermore, this study will incorporate the concept-centric methodology proposed by Webster and Watson (2002). The rationale behind adopting the eight-step review guide lies in its efficacy in offering a structured and methodological approach to the compilation of literature, enabling a precise and objective expression of the underlying connotations within a literature review. The utilization of the concept-centric approach from Webster and Watson adds a layer of refinement to the review process. This approach is instrumental in facilitating a more streamlined and coherent presentation of the literature. By emphasizing key concepts and their interrelations, it enables the construction of a concise and logically structured narrative. Moreover, the amalgamation of the eight-step systematic review guidance and the concept-centric approach not only ensures a methodologically rigorous literature review but also enhances the clarity, conciseness, and logical coherence of the presentation, thereby fortifying the groundwork for subsequent analyses. The details of how the literature review was conducted are reported in Paper A.

6.2 Semi-structured interview with voice recording

Based on Figure 1, designing AI that users can understand should consider user needs. In order to collect user needs, we conducted semi-structured interviews and recorded the interviews. Next, we converted the recording into text. Finally, we proposed nine user-based AI design principles (Table 1) by extracting key information from textual data.

An interview is a data collection method used in research to obtain information from individuals or groups through face-to-face or online interaction. According to Kvale (1996), an interview is "an interactional, communicative event where at least two persons are engaged in a mutual attempt to communicate with each other, for a particular purpose" (p. 2). The purpose of the interview is to discuss the interviewees' personal experiences and opinions or to collect information about specific topics. In this research, the interview is semi-structured because it allows more flexibility and allows the interviewer to follow up on the interesting questions raised by the interviewees. The accuracy of the data can be guaranteed by using the recording in the interview because the researcher can replay the recording to check the details or clarify the answer. The disadvantage is that data storage may cause the interviewees to worry about privacy disclosure. The design of the questions in the interviews was guided by the principles of human-centered explainable system designing by Mueller et al. (2021). The details of how the semi-structured interview was conducted are reported in Paper B.

7. CONTRIBUTIONS

This thesis has made substantial contributions to AI, explainable AI, and human-computer interaction through the discussions about how to answer Q1, Q2, and Q3. Especially, some significant research findings can guide AI developers in designing AI that users can understand. Additionally, discussions on research limitations and shortcomings, as well as prospects for future research, can also navigate further research on explainable and human-centered AI. Both the author of this thesis and other researchers in the field can benefit from it.

Paper A: Towards Trustworthy and Understandable AI: Unraveling Explainability Strategies on Simplifying Algorithms, Appropriate Information Disclosure, and High-level Collaboration

This literature review aims to explore how to overcome the challenges of strategic explainability through simplifying algorithms, appropriate information disclosure, and high-level collaboration, thereby offering future research direction for building AI systems that are trustworthy and understandable to users. The main contributions are the following:

Three dimensions on explainability: The construction of three dimensions is a conceptually thematic interpretation of strategic explanations, guiding AI developers and researchers in thinking about how to build understandable and trustworthy AI. The construction of three dimensions is primarily based on the discussion of three types of opacity by Burrell (2016), Pasquale's (2015) exploration of the reasons for the black box issue from the perspective of protecting business secrets, as well as discussions by Ansgar et al. (2017) and Prahalad and Ramaswamy (2004) on user participation and value co-creation. The work of these predecessors has illuminated the path for this study, and in turn, this study also complements their work, more importantly, providing a theoretical framework that can be drawn upon for future research. The dimension of simplifying algorithm can help AI developers think about the gap between human understanding and the impossibility of understanding complex algorithms such as neural networks. At the same time, simplifying algorithm also effectively addresses Burrell's (2016) concerns about the opacity of algorithms. Pasquale (2015) argued that businesses will prevent information disclosure of algorithms due to the protection of their secrets. The introduction of the dimension of appropriate information disclosure can prompt AI developers to consider how to balance informing the public and protecting business interests from multiple perspectives. High-level

collaboration can lead stakeholders in the AI environment to delve into how to build an understandable AI environment through collaboration, participation, and interaction. Importantly, the combined different explainability strategies based on these three dimensions can be seen as an effective implementation of explainable approaches in future real-world environments.

Face the challenges of explainability: Explanatory approaches in AI face challenges in terms of reliability, comprehensibility, trust calibration, and BDI (Belief, Desire, and Intention). Some approaches, while effective, can yield different and even contradictory explanations across datasets. For instance, visualization methods may be fragile, producing diverse explanations with slight data perturbations (Ghorbani et al., 2019). Unreliable explanations pose a risk of user trust erosion. Comprehensibility challenges stem from cognitive limitations (Durán & Jongsma, 2021) or epistemic absence (Zednik, 2021). Opacity arises when an AI system fails to provide all epistemically relevant elements for understanding the input-output transformation (Humphreys, 2009), leaving users, especially non-technical ones, with a superficial grasp of AI processes. Existing explanatory methods may not adequately address cognitive and epistemic opacity for all stakeholders. Trust calibration, vital for appropriate AI utilization, concerns the alignment of trust with automation capability (Lee & See, 2004). Over-trust occurs when trust exceeds AI capability, while distrust arises when trust falls short. Despite studies indicating that explanations can enhance trust, there's a lack of clarity on how they contribute to trust calibration (Naiseh et al., 2021). Research in this area is scarce, leaving questions about the nuanced relationship between trust, explanations, and AI capability unanswered. BDI forms the foundation of trust construction. Explanations provide information, but their role in shaping beliefs and advancing trust generation is understudied. Most current explanatory methods focus on reactive, non-intentional systems, explaining single decisions based on features and parameters rather than delving into the agent's beliefs and desires (Dazeley et al., 2021). Achieving "social explanation" (p.11), which encompasses higher-level interpretations involving beliefs and consciousness, remains a challenge unaddressed by current methods.

The existing emphasis on debugging parameters and visualizing model outputs, while crucial, neglects socio-technical challenges. Explainability strategies should extend beyond addressing black-box issues to encompass user-centric elements such as reliability, trust, comprehensibility, and belief. The current literature lacks sufficient research on constructing explainability strategies tailored to user understanding. Proposing three dimensions—

simplifying algorithms, appropriate information disclosure, and high-level collaboration (Table 1)—can enhance user comprehension of AI decisions. While the efficacy of simplifying algorithms for reliability is yet to be fully confirmed, a promising research direction involves providing verifiable explanations through semantic, local, and global explanations. This approach empowers users to assess system decisions' reliability. Appropriate information disclosure enhances transparency, aiding users in understanding AI's internal mechanisms and decision criteria, and fostering trust calibration. High-level collaboration facilitates interactions, allowing stakeholders to collectively explore and validate the decision-making process, addressing reliability and comprehensibility challenges. Additionally, collaborative efforts help unveil the system's beliefs, desires, and intentions, promoting shared understanding and calibration of system behavior, particularly in the context of the BDI challenge

Paper B: An Empirical Investigation in High-stakes Areas: User-based AI Explainability Development Principles

This study emphasizes the importance of users comprehending AI decision-making and cultivating appropriate trust, particularly in high-stakes domains. To achieve this, the research adopts a focused approach, conducting semi-structured interviews with AI users across four critical sectors: banking, education, healthcare, and justice. The objective is to gain insights into users' requirements, satisfaction levels, and perspectives regarding the explainability of AI within their workflow. The findings from this investigation are intended to offer valuable guidance to AI developers, aiding them in designing systems that are both trustworthy and understandable for users, especially in high-stakes contexts. The main contributions are the following:

Three stages of explainability: To address C3 and answer Q3, user needs from the above three stages (see section 1.3.2) can help AI developers build user-based explainability principles. According to these principles, AI developers can try to think about how to build AI that users can understand. These principles are shown in Table 2 (Paper B).

Table 2. The explainability of AI design principles based on user needs, satisfaction, and perspectives

Three stages of explainability	Principles	Users' needs, satisfaction, and perspectives

Conceptualization	Pre-explanations	Pre-explanations should be provided by AI developers or suppliers before AI is embedded in the workflow
	Users' practices and behaviors	Before designing AI, developers should have a deep understanding of users' work practices and behavior using digital tools
Construction	Comparison, validation, and auxiliary tools	Users tend to use the repeated verification and comparison of historical data and results to verify the accuracy of AI. Developers should consider AI decision-making as an auxiliary tool rather than a fully automated final decision that instead users.
	Direct or indirect participation	For better understanding, developers should consider allowing users to directly or indirectly participate in the development process of AI
	Explain training data and results in effective ways	Developers should be responsible for explaining the composition of training data and the results of the model. Such explanations should be based on the actual work situation and process of users.
	Provide personalized explainability approaches for different scenarios	Developers should design AI based on user needs, workflows, and best practices for understanding AI decisions
Measurement	Long-term verification in practice	Whether explanations of AI decision-making can increase trust requires long-term verification in practice.
	Meeting users' roles	Allowing users to play various roles in evaluating explainability and continuous improvement, such as feedback provider, beneficiary, bridge, leader, and co-creator, can increase users' trust in the AI systems.
	Factors of trust or distrust	It is crucial to determine the trust and distrust factors of AI systems based on user-specific work practices for designing trustworthy AI.

In the conceptualization stage, AI developers should make the AI system clear and understandable, facilitating its use by non-experts. This can enhance the perceived ease of use and usefulness (Eze et al., 2021). As

shown in Table 2, this involves providing instructions about AI before integrating AI and understanding users' work practices and behaviors in key domains. Informing users about the benefits of AI in their work (Bölen, 2020) and making them feel an enjoyable perception of using the system (Ashfaq et al., 2020) will increase their willingness and positive attitude toward using the system. In the construction stage, based on Table 2, AI should be a tool for users during development rather than a fully auto-decision-maker. Users need functionality to validate AI results based on data, allowing the system to comply with legal and ethical constraints discussed by Arrieta et al. (2020). Transparency in training data is crucial, and explanations should be relevant to real-world situations. Additionally, developers should encourage user participation (Ansgar et al., 2017) and customize AI systems to meet the diverse needs of different users, as the knowledge structures of different users vary (Mohseni et al., 2021). In the measurement stage, evaluation involves testing the interpretability of AI systems in real-world scenarios for non-experts. Users can use their domain-specific knowledge and background to assess the adequacy of these explanations in practical situations (Szymanski et al., 2021). This assessment requires continuous long-term monitoring. Users need to understand their roles so that they can provide feedback at different levels and categories to address unexpected, adverse, and biased effects of decisions (Barocas et al., 2017). AI developers should refine the system based on user feedback and practical experience. Additionally, AI developers need to understand factors influencing trust during the system's operation to ensure its sustainability.

Automation-Trust Calibration, Resolution, and Specificity: Since there is always a gap between humans' understanding of automation and its actual capability, a deficiency in human's lack of objective assessment of automation capability can only be remedied with trust (Blomqvist, 1997). Thus, research on the relationship between user trust and automation capabilities is crucial. Research on the user trust and automation capabilities relationship, termed automation-trust calibration (Muir, 1987). The automation-trust resolution has been described as the correspondence between trust and automation capabilities (Cohen et al., 1998). About functional and temporal specificity (Lee & See, 2004), high functional specificity corresponds to trust in subfunctions, while low specificity extends to the entire system. High temporal specificity aligns trust with immediate fluctuations, while low specificity matches long-term changes. Optimal calibration, resolution, and specificity will mitigate underuse and overuse of automation (Lee & See, 2004). In this study, the explainability based on user needs, satisfaction, and perspectives (Table 2) can connect to automation trust calibration, resolution, and specificity (Table 3). This can provide

researchers with a broad perspective on how to conduct research in user-based trust-automation. Although the applicability of these principles that contribute to improving trust-automation needs to be further validated in practice, at least they guide the direction of future research.

Table 3. Design principles and automation-trust calibration, resolution, and specificity. The √ represents the three elements of automation-trust and their correlation with corresponding principles.

Three stages of explainability	**Principles**	**Calibration**	**Resolution**	**Specificity**
Conceptualization	Pre-explanations	√		
	Users' practices and behaviors	√		
Construction	Comparison, validation, and auxiliary tools	√		
	Direct or indirect participation	√		
	Explain training data and results in effective ways		√	
	Provide personalized explainability approaches for different scenarios			√
Measurement	Long-term verification in practice			√
	Meeting users' roles	√		√
	Factors of trust or distrust	√	√	

8. THE FUTURE WORK

A brief review of this thesis shows that by analyzing the nature of transparency and interpretability, we have built strategic explainability in three dimensions, and investigated user needs in three stages. The purpose is to discuss how to build AI that users can understand. Although we discussed trustworthy AI in Paper B through user-based principles and the relationship between trust automation, this thesis did not delve too much into how to build trustworthy AI after building understandable AI. The main direction of future research should be based on the user-based explainability principles and trust automation discussed in this thesis, to elaborate on how to construct trustworthy AI.

People will have inappropriate trust in and improper use of AI in various environments. The four terms often involve in the interaction between humans and automation, "use, misuse, disuse, and abuse" (Parasuraman & Riley, 1997). Disuse results from distrust (Grigsby, 2018) and under-trust (Parasuraman & Riley, 1997; Okamura & Yamada, 2020). People do not trust machines that they do not understand (Burrell, 2016), and such machine-generated prediction (Azodi, et al., 2020). In other words, if AI users cannot explain AI's results, they will abandon the use of it (Strohm et al., 2020). The risk of algorithms and their systematic bias may also lead to distrust (Green, 2020), which may trigger people to re-evaluate the use of algorithms. Meanwhile, over-trust results in misuse (Madhavan & Wiegmann, 2007; Bahner et al., 2008; Wagner et al., 2018; Okamura & Yamada, 2020). Many users are insensitive to the reliability of automation. They may misuse a system by over-relying on it, which comes from over-trusting the system and causing users to accept the system's recommendations and ignore their correctness (Bussone et al., 2015). In this case, users may trust systems more than trust themselves (Madhavan & Wiegmann, 2007).

Some examples show people's distrust and under-trust in AI. In the medical field, radiologists without an understanding of AI and related techniques become sceptical about the results produced by the algorithms (Strohm et al., 2020). "If the result of an algorithm cannot be explained by the physician to the patient, the algorithm must not be used, not even as a second opinion" (Ursin et al., 2022, p.146), which may eventually lead radiologists to abandon the use of AI (Strohm et al., 2020). In judicial trials, AI's risk score of specific cases cannot be transmitted to the judge in a transparent way, which leads to the judge's not understanding of AI decisions (Stevenson & Slobogin, 2018). For example, why AI shows that young people have a potentially

higher risk than others (Stevenson & Slobogin, 2018), and blacks are at higher risk of becoming recidivists (Barenstein, 2019; Larson, 2016). Judges may, therefore, generate less trust in the decisions of AI. Wagner et al. (2018) used two examples to explain over-trust. One example is that people follow the instructions of robots during an emergency evacuation without any questions because most people think robots know how to evacuate more than people do. Another example is that some people use autonomous driving instead of driving a car manually. These two examples reflect the possible risks caused by over-trust, that is, unsuccessful evacuation and car accidents. Wagner et al. thought that people might accept the risks they cannot usually tolerate (e.g., unsuccessful evacuation and car accidents) because they treat the intentions of robots or autopilots as developers' while ignoring possible mechanical failures.

By reviewing the previous discussion once again, in paper A, we emphasize the construction strategy explanation around three dimensions. These solutions not only involve algorithms and business information but also involve interactions between other stakeholders in the same AI environment. The purpose of these strategies is to build understandable AI; In paper B, we discussed how user-based AI design principles affect trust-automation. Based on these discussions, we can try to imagine how to incorporate interpretive strategies into these user-based AI design principles, thereby advancing understandable AI (Figure 1) towards trustworthy AI. For example, referring to the "Pre-explanations" principle (Table 2), developers should consider how to implement "Professional explanation", "Legal explanation", and "AI supplier explanation" (Table 1) to calibrate trust before embedding AI systems into work practice (Table 3); Similarly, referring to "Meeting users' roles" (Table 2), how developers should strengthen "Human-human collaboration" or "Human-machine collaboration" (Table 1) to achieve regulation of calibration and specificity (Table 3).

The above discussion can serve as the starting point for our next research. This starting point can guide the future work on "trustworthy AI" as an extension of "understandable AI." In addition to verifying our research findings of this thesis in real environments and conducting broader user surveys (see more details in Paper A and B), we will advance our research on trustworthy AI and make continuous contributions to this field.

9. CONCLUSION

The ubiquity of AI in our society brings forth both opportunities and challenges, with the black-box issue standing as a significant hurdle in realizing the full potential of artificial intelligence. As AI becomes increasingly integrated into various facets of our lives, especially in critical domains like healthcare and the judiciary, the lack of transparency and interpretability poses risks to privacy, responsibility, and justice. The opaque nature of AI decision-making processes hinders users, particularly those without technical backgrounds, from understanding the reasons behind the recommendations or decisions made by AI systems. The black-box issue stems from the relentless pursuit of algorithmic performance, which, while enhancing accuracy, results in increased complexity. As algorithms become more sophisticated, explaining the intricacies of their decision-making processes becomes challenging even for developers. The consequences of blindly trusting these opaque decisions are particularly dire in fields where decisions hold high stakes. Researchers have proposed various solutions to address the black-box issue, ranging from technical enhancements in algorithm transparency and interpretability to Explainable AI (XAI) approaches. While these efforts contribute valuable insights, they fall short in providing comprehensive and universally applicable solutions. Technical solutions may be effective for developers but often lack practicality for end-users, such as doctors and judges. XAI approaches, while promising, face limitations in specific scenarios and require careful consideration of diverse factors like data sources, algorithms, and user needs.

This research embarked on a critical exploration of how to design AI systems that users can understand, navigating through three key challenges (C1, C2, and C3). The first challenge emphasizes the need to balance transparency and interpretability, shedding light on the relationship between the two and their collective role in building trustworthy AI. The second challenge delves into strategic explanations, emphasizing the importance of developing methods that cater to users rather than solely focusing on the convenience of developers. The third challenge underscores the paucity of user-based research, calling for a human-centered perspective in AI development that prioritizes understanding user requirements, experiences, and perspectives. To address these challenges, the research formulated three pivotal questions (Q1, Q2, and Q3). Through a series of discussions, investigations, and user interviews, this research seeks to provide insights into these questions and offer principled guidance for AI developers.

This thesis explored transparency and interpretability in-depth, elucidating the objectivity of transparency and the subjectivity of interpretability. Paper A extensively reviewed existing strategies for explainability, revealing a predominant focus on simplifying algorithms to provide explanations, while placing less emphasis on appropriate information disclosure and fostering high-level collaboration. This underscores the urgent necessity to enhance research efforts in these dimensions. While practical validation is crucial for addressing challenges like reliability, comprehensibility, calibration, and BDI, a judicious integration of simplifying algorithms, appropriate information disclosure, and high-level collaboration should guide the development of explainability strategies. Although strategic explanations have enormous potential for improving transparency in AI decision-making, their implementation faces numerous limitations, necessitating further exploration and validation in real-world settings. Nevertheless, these challenges serve as guideposts for researchers, outlining the next steps in advancing AI explainability strategies. Future research on user-centered explainability should prioritize user feedback on system understanding, explore how AI suppliers can offer relevant information, concentrate on interactive AI development involving professionals and users, and tailor explanations to diverse user needs. Paper B, based on interviews with nine non-technical professionals in high-stakes areas, formulated key principles for AI explainability development. These principles guide developers in creating reliable and comprehensible AI systems, especially in critical domains, with the potential to enhance automation-trust facets: calibration, resolution, and specificity. Effective calibration, high resolution, and specificity can address issues of underuse and overuse of AI. However, practical implementation presents challenges for AI developers, including heightened system complexity and resource optimization. Limitations in research methods and data volume in Paper B underscore the need for extensive future work, such as broader user surveys and domain extension. Despite being grounded in empirical evidence, validating these key principles in real-world contexts and assessing their efficacy in achieving understandable and trustworthy AI demand further exploration, marking a crucial direction for future studies.

In essence, the research underscores the necessity of a multidimensional approach that combines technical advancements with human-centered perspectives. It advocates for AI systems that not only meet the demands of algorithmic performance but also prioritize user understanding, satisfaction, and trust. By addressing these challenges and questions, this research contributes to the ongoing dialogue on creating AI systems that align with the values and needs of the users they serve, fostering a future where AI is not just powerful but also transparent, interpretable, and accountable.

REFERENCES

ACM (2018). ACM Code of Ethics and Professional Conduct. URL: https://www.acm.org/code-of-ethics

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

Ansgar, K., Perez, V. E., Helena, W., Menisha, P., Sofia, C., Marina, J., & Derek, M. (2017). Editorial responsibilities arising from personalization algorithms. *The ORBIT Journal*, 1(1), 1-12.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, *11*(11), 5088.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv:1909.03012*

Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, 54, 101473.

Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the black box: interpretable machine learning for geneticists. *Trends in genetics*, *36*(6), 442-455.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.

Bahner, J. E., Elepfandt, M. F., & Manzey, D. (2008, September). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. In *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting (Vol. 52, No. 19, pp. 1330-1334). Sage CA: Los Angeles, CA: SAGE Publications.

Barenstein, M. (2019). ProPublica's COMPAS Data Revisited. *arXiv preprint arXiv:1906.04711*.

Bellucci, M., Delestre, N., Malandain, N., & Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. *Procedia Computer Science*, *192*, 241-250.

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. arXiv preprint arXiv:2007.05408.

Bond, R. R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., ... & Adel, T. (2019, October). Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making. *In RoCHI*, 2-9.

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, 13 (3), 271-286.

Bucher, T. (2016). Neither black nor box: Ways of knowing algorithms. In *Innovative methods in media and communication research*, 81-98. Palgrave Macmillan, Cham.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, 160-169. IEEE.

Bölen, M. C. (2020). Exploring the determinants of users' continuance intention in smartwatches. *Technology in Society*, 60, 101209.

Chander, A., Srinivasan, R., Chelian, S., Wang, J., & Uchino, K. (2018, January). Working with beliefs: AI transparency in the enterprise. In *IUI Workshops*.

Chu, H. Y., & Shen, Y. (2022, June). User Feedback Design in AI-Driven Mood Tracker Mobile Apps. In *International Conference on Human-Computer Interaction* (pp. 346-358). Cham: Springer International Publishing.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. In *in Proc. Command and Control Research and Technology Symp*.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, *3*(3), 201-215.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, *299*, 103525, p.11.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.

Emaminejad, N., & Akhavian, R. (2022). Trustworthy AI and robotics: Implications for the AEC industry. *Automation in Construction*, *139*, 104298.

Eze, N. U., Obichukwu, P. U., & Kesharwani, S. (2021). Perceived usefulness, perceived ease of use in ICT support and use for teachers. *IETE Journal of Education*, 62(1), 12-20.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261-262.

Garvey, C. (2018, December). AI risk mitigation through democratic governance: Introducing the 7-dimensional AI risk horizon. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 366-367.

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, 33(01), 3681-3688.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80-89. IEEE.

Green, B. (2020, January). The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 594-606.

Grigsby, S. S. (2018, July). Artificial intelligence for advanced human-machine symbiosis. In *International Conference on Augmented Cognition*, 255-266. Springer, Cham.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6), 14-23.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, *2*(2), 1

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120.

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), e61.

Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022, March). Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces*, 537-548.

Guo, W., Mu, D., Xu, J., Su, P., Wang, G., & Xing, X. (2018, October). Lemna: Explaining deep learning based security applications. In *proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 364-379.

Heuillet, A., Couthouis, F., & Díaz-Rodríguez, N. (2022). Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values. *IEEE Computational Intelligence Magazine*, *17*(1), 59-71. p.62.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615-626.

Jin, W. E. I. N. A., Fan, J. I. A. N. Y. U., Gromala, D., Pasquier, P., & Hamarneh, G. (2021). EUCA: the End-User-Centered Explainable AI Framework. *arXiv preprint arXiv:2102.02437*.

Kaur, D., Uslu, S., Durresi, A., Badve, S., & Dundar, M. (2021, July). Trustworthy explainability acceptance: A new metric to measure the trustworthiness of interpretable AI medical diagnostic systems.

In *Conference on Complex, Intelligent, and Software Intensive Systems*, 35-46. Springer, Cham.

Kvale, S. (1994). *Interviews: An introduction to qualitative research interviewing*. Sage Publications, Inc.

Khan, M. S., Nayebpour, M., Li, M. H., El-Amine, H., Koizumi, N., & Olds, J. L. (2022). Explainable AI: A Neurally-Inspired Decision Stack Framework. *Biomimetics*, *7*(3), 127.

Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023, April). " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-17.

Kumar, A., Braud, T., Tarkoma, S., & Hui, P. (2020, March). Trustworthy AI in the age of pervasive computing and big data. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 1-6. IEEE.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*(1), 3-3.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80, p.55.

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 1-38.

Lim, B. Y., Yang, Q., Abdul, A. M., & Wang, D. (2019). Why these explanations? Selecting intelligibility types for explanation goals. In *IUI Workshops*.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.

Lisboa, P. J. G., Saralajew, S., Vellido, A., & Villmann, T. (2021, October). The coming of age of interpretable and explainable machine learning models. In *ESANN 2021 proceedings European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. ESANN Ciaco.

Lyu, D., Yang, F., Kwon, H., Dong, W., Yilmaz, L., & Liu, B. (2021). TDM: Trustworthy decision-making via interpretability enhancement. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *6*(3), 450-461.

Lötsch, J., Kringel, D., & Ultsch, A. (2021). Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics*, *2*(1), 1-17.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277-301.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655, p.9.

Mencar, C., & Fanelli, A. M. (2008). Interpretability constraints for fuzzy information granulation. *Information Sciences*, *178*(24), 4585-4618.

Meske, C., & Bunde, E. (2020, July). Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support. In *International Conference on Human-Computer Interaction* (pp. 54-69). Springer, Cham.

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

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4), 1-45.

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

Moradi, M., & Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. *Expert Systems with Applications*, *165*, 113941.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human-AI systems. *arXiv preprint arXiv:2102.04972*.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6), 527-539.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, *169*, 102941.

Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendations and calibrated trust: two systematic user errors. *Computer*, 54(10), 28-37.

Naiseh, M., Jiang, N., Ma, J., & Ali, R. (2020). Personalising explainable recommendations: literature and conceptualisation. In *Trends and Innovations in Information Systems and Technologies:2(8)*, 518-533. Springer International Publishing.

Nazari, M., Kluge, A., Apostolova, I., Klutmann, S., Kimiaei, S., Schroeder, M., & Buchert, R. (2021). Data-driven identification of diagnostically useful extrastriatal signal in dopamine transporter SPECT using explainable AI. *Scientific Reports*, *11*(1), 1-13.

Nguyen, A. P., & Martínez, M. R. (2020). On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *Plos one*, *15*(2), e0229132.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Prahalad, C. K., & Ramaswamy, V. (2004). Co-creation experiences: The next practice in value creation. *Journal of interactive marketing*, 18(3), 5-14.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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*, 1135-1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence 32(1).

Rim, T. H., Lee, A. Y., Ting, D. S., Teo, K., Betzler, B. K., Teo, Z. L., ... & Cheung, C. M. G. (2021). Detection of features associated with neovascular age-related macular degeneration in ethnically distinct data sets by an optical coherence tomography: trained deep learning algorithm. *British Journal of Ophthalmology*, *105*(8), 1133-1139.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, *140*, 105111.

Sanneman, L., & Shah, J. A. (2020, May). A situation awareness-based framework for design and evaluation of explainable AI. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 94-110). Springer, Cham, p.107.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59-68.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE international conference on computer vision*, 618-626.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495-504.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, *6*(6), 52.

Sundararajan, M., Taly, A., & Yan, Q. (2016). Gradients of counterfactuals. arXiv preprint arXiv:1611.02639.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proceedings of the IEEE international conference on computer vision*, 618-626.

Sperrle, F., El-Assady, M., Guo, G., Chau, D. H., Endert, A., & Keim, D. (2020). Should we trust (x) AI? Design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433*.

Stevenson, M. T., & Slobogin, C. (2018). Algorithmic risk assessments and the double-edged sword of youth. *Behavioral sciences & the law*, *36*(5), 638-656.

Strohm, L., Hehakaya, C., Ranschaert, E. R., Boon, W. P., & Moors, E. H. (2020). Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology*, 30(10), 5525-5532.

Ursin, F., Timmermann, C., & Steger, F. (2022). Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary?. *Bioethics*, *36*(2), 143-153.

Verhagen, R. S., Neerincx, M. A., & Tielman, M. L. (2021, May). A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, 119-138. Cham: Springer International Publishing.

Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM*, *61*(9), 22-24.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii-xxiii.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption

generation with visual attention. In *International conference on machine learning*, 2048-2057. PMLR.

Zadeh LA (1954). System theory. *Columbia Engineering Quarterly*, p. 16-34.

Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & technology*, 34(2), 265-288.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921-2929.

APPENDED PAPERS

Paper A

Towards Trustworthy and Understandable AI: Unraveling Explainability Strategies on Simplifying Algorithms, Appropriate Information Disclosure, and High-level Collaboration

Shuren Yu

In Proceedings of the 26th International Academic Mindtrek Conference, 2023, p.133-143. https://doi.org/10.1145/3616961.3616965

Abstract

Human-centered artificial intelligence (AI) has garnered significant attention. Explainability strategies based on the concept of explainable AI (XAI) are comprehensive sets of techniques and principles that help users establish understandable and trustworthy AI systems. However, existing explainability strategies still face numerous challenges in enabling users to understand AI system decisions better. This literature review aims to explore how to overcome these challenges through simplified algorithms, appropriate information disclosure, and high-level collaboration, thereby offering future research direction for building AI systems that are trustworthy and understandable to users.

Keywords: Explainability strategy, Simplifying algorithm, Appropriate information disclosure, High-level collaboration, XAI

1. Introduction

AI systems are becoming increasingly opaque due to the complexity of modern machine learning algorithms, the use of large datasets, and the demands of high-stakes applications. One example is the increasing use of deep learning algorithms based on complex neural networks with many layers. These networks can be trained on massive datasets and have millions or even billions of parameters, making it difficult, if not impossible, to understand how the models make decisions or predictions. As a result, these models can become "black boxes" that are opaque to human understanding (Goodman & Flaxman, 2016). Moreover, as machine learning models are trained on more extensive and diverse datasets, it can become more difficult to trace the decisions made by the models back to the underlying data (Burrell, 2016). Finally, some AI systems are designed to learn and evolve, making them difficult to interpret or predict. As these systems continue to learn and adapt, they may become increasingly opaque, making it difficult to understand how they make decisions or why they behave in specific ways (Kleinberg et al., 2018). AI systems are increasingly used in high-stakes applications, such as healthcare and criminal justice, where the decisions made by AI systems can have significant consequences for individuals and society. Because the majority of users of artificial intelligence systems do not have a technical background (Liang et al., 2021), there is an increasing demand for transparency and accountability of artificial intelligence systems, as well as the ability to interpret decision-making methods (Selbst et al., 2019).

XAI (Explainable AI) is a widely discussed set of goals and techniques to establish AI that is easy for humans to understand (Arrieta et al., 2020; Gunning et al., 2019). Researchers have developed many explainability methods, approaches, and frameworks based on XAI (Dazeley et al., 2021; Markus et al., 2021; Naiseh et al., 2023; Sperrle et al., 2020), to help users better understand how systems operate and make decisions. However, in most practices, deploying these explainability strategies focuses on engineers and developers rather than end-users (Bhatt et al., 2020). Therefore, it is necessary to study how to establish strategies that enable users to understand AI system decisions better.

This article examines and synthesizes existing literature on explainability strategies using three key dimensions for XAI as a lens. These three dimensions can provide developers and scholars with a comprehensive and broad perspective on how to construct explainability strategies, thereby helping users better understand AI decisions. The first dimension is

simplifying algorithms without reducing decision criteria (Mittelstadt et al., 2019). "If the machine learning algorithm is based on a complicated neural network or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why" (Bostrom & Yudkowsky, 2018, p.1). The amount of information humans can understand, and the process is limited (Miller, 1956), so transforming complex algorithms into a simpler form may make them easier to understand. The second dimension involves appropriate information disclosure. Considering business secrets, complete disclosure of algorithm code may not be acceptable, but disclosing certain key information, such as summary results and benchmarks, will more effectively communicate algorithm performance to the public (Diakopoulos, 2016). The information asymmetry between AI companies and ordinary users can lead to algorithm opacity and accountability issues (Lepri et al., 2018); Therefore, it is necessary to increase users' understanding of algorithms through appropriate information disclosure. The third dimension concerns high-level collaboration between humans and AI. Participation in interactive machine learning can increase users' understanding of AI algorithms and promote coupling between humans and machines (Amershi et al., 2014). Explanation can be understood as a social process that emphasizes the importance of dialogue (Weld & Bansal, 2019), which means that effective communication and collaboration among different stakeholders in the AI environment can increase their understanding of AI.

By reviewing the relevant literature, this article aims to answer the following question: *How can existing explainability strategies help users better understand artificial intelligence decisions through these three dimensions, simplifying algorithms, appropriate information disclosure, and high-level collaboration?*

In answering this question, this article will have made the following four contributions: *Firstly, a valuable literature review will be provided for the study of explainability strategies. Secondly, by reviewing existing explainability strategies, this article can reveal what contributions existing strategies have made in these three dimensions. Thirdly, this article also discusses some limitations of implementing existing explainability strategies in real-world environments and emphasizes the importance of validating these strategies. Finally, this article provides AI designers and developers with research direction for constructing explainability strategies user-centered in the future.*

2. Background

2.1 The challenges of explanation

Although there are many techniques and methods to establish explainable AI, they may still not be able to provide users with an understandable and trustworthy AI system, as they still face challenges such as reliability, comprehensibility, calibration, and BDI (belief, desire, and intention) (Table 1).

Table 1. The challenges of explanations

Challenge	Explanation
Reliability	Explanations may not be trustworthy.
Comprehensibility	Cognitive limitation or epistemic absence.
Calibration	No calibration of trust and AI's capabilities.
BDI (belief, desire, and intention)	Explanations are based on features and parameters, not beliefs, desires, and intentions (BDI).

Reliability. There are many alternative explanatory methods and techniques, but some of them have defects that may lead to a lack of trust. For example, although LIME (Local Interpretable Model-agnostic Explanations) is widely used in many scenarios, such an explanatory method based on local approximation can only capture the local characteristics of a model but cannot explain its global decision-making behavior (van der Linden et al., 2019). Moreover, LIME is unable to explain the impact of the relationship between the features of specific instances on decision-making. Although Shapley is a good method, especially when it is applied to the development of the SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), in a comparative study of LIME and SHAP, their explanations differed in several defect prediction data sets (Roy et al., 2022). Different and even contradictory explanations might lead to potential risks. Visualization methods can be fragile. A study by Ghorbani et al. (2019) shows that adding perturbation to the original data can produce completely different explanations without changing the prediction of the model. To a user, such unreliable explanations may also lead to a loss of trust.

Comprehensibility. Compared to transparency, opacity is a cognitive limitation (Durán & Jongsma, 2021) or epistemic absence (Zednik, 2021), which may be caused either by features of AI algorithms and the scale required to successfully apply them (Burrell, 2016). Also, opacity may occur because of the inability to determine the reliability of artificial intelligence or the lack of reason to believe in the results of them (Durán & Jongsma, 2021). Cognitive limitation and epistemic absence may be based on epistemic situations (time, status, or process), because:

> *Here a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to know all of the epistemically relevant elements of the process (Humphreys, 2009, p.218).*

Epistemically relevant elements can be understood as "a step in the process of transforming inputs to outputs, or as a momentary state transition within the system's overall evolution over time" (Zednik, 2021, p.269). Thus, the opacity is actually that an AI system does not provide enough epistemically relevant elements to agent X at time t so agent X is not fully aware of how input-output is transformed and how overall transition happens with the AI system. Existing explanatory methods may not provide the full range of epistemically relevant elements for all agents, rather they merely provide some approaches to understanding AI for a few stakeholders, such as AI developers and algorithm engineers. For most agents and stakeholders, such as the non-technical end-users (Jin et al., 2021), the explanations can only make AI less complex, while not fundamentally addressing cognitive and epistemic opacity.

Calibration. Trust calibration is the relationship between trust and automation capability (Lee & See, 2004). Over-trust arises when trust is higher than the capability of AI, and distrust will happen when it is lower than the capability of AI. When trust matches the capability of AI, it appears on the diagonal (Lee & See, 2004, p.55). Therefore, calibrated trust can help users use AI's capabilities properly. Although some studies have provided some evidence about how explanations can help improve trust in AI systems (Jacobs et al., 2021; Naiseh et al., 2021; Yin et al., 2019; Zhang et al., 2020), and other studies show that explanations will, in turn, promote users' over-reliance on AI (Bussone et al., 2015; Jacobs et al., 2021; Naiseh et al., 2023), it is unclear how the explanations support trust calibration (Naiseh et al., 2021). Research on how explanations help to calibrate the corresponding relationship between trust and AI capability is still very scarce.

BDI (belief, desire, and intention). According to the statements of Lee and See (Lee & See, 2004), Ajzen and Fishbein (1980) developed a framework for the definition and construction of trust. In this framework, trust is an attitude, while reliance is a behavior. Available information and personal experience can affect the establishment of beliefs. Beliefs and perceptions also affect attitudes. Attitude determines the generation of intention, and behavior is based on intention. Explanations can provide recipients with information that can be leveraged (e.g., rule extraction by limes, Shapley values for different features, and visualization), and recipients of interpretations can also utilize different levels of personal experience (e.g., knowledge, expertise, and cognition) to process this information. However, how such information and human experience construct beliefs and further advance the generation of trust is seldom addressed by explanatory methods. According to Dazeley et al. (2021), at present, most of the widely used explanatory methods are at a low level - "purely reactive non-intentional systems" (p.8), and their focus is on explaining a single decision based on features and parameters, not "agent's current internal disposition…such as belief and/or desires" (p.9). However, explanations based on decoding beliefs and consciousness need to achieve "social explanation" (p.11) - a higher level of explanation paradigm. The "social explanation" will provide understandable explanations, such as "Why change the meeting time?" and "If I want to raise, how do you plan to play?", which are the scope that the current existing explanatory methods cannot reach out.

In the current research, most techniques and methods in explainability strategies aim to debug parameters, such as a heat map for the results of Convolutional Neural Networks (CNN), rather than consider the requirements of the users (Miller et al., 2017). The focus of explainability strategies should be extended to how to deal with socio-technical challenges. Such challenges should not only involve overcoming the no-transparency and un-interpretability of black-boxes, but also provide cognitive elements for users and other stakeholders, such as reliability, trust, comprehensibility, and belief. Compared to a large amount of literature on the development and discussion of artificial intelligence (AI) technology and explainable AI (XAI), there is still a significant lack of research aimed at building explainability strategies to help users better understand AI decisions. Therefore, we propose three dimensions for constructing explainability strategies to increase user understanding: simplifying algorithms, appropriate information disclosure, and high-level collaboration.

2.2 Three dimensions for better understanding

Simplifying algorithms. Burrell (2016) thought the opacity of AI systems is due to the expertise in writing algorithms, which is generally not available to the public. She emphasized that the language of algorithm writing is very different from human language, so the algorithm needs to be explained before it can be understood by most people. A typical example of this opacity is the complexity of deep learning and the low causal inference. Models based on deep learning cannot be easily interpreted. Deep learning is built on relevance rather than causality because "deep learning learns complex correlations between input and output features, but with no inherent representation of causality." (Marcus, 2018, p.12). This can prevent people from understanding how input produces output. For example, "in the case of DNNs, it may not be possible to understand the determination of output." (Topol, 2019, p.51). This creates opacity. Besides, because deep learning uses a non-linear structure, deep learning is presented in the form of a black box, that is, deep learning does not explain what makes the model conclude (Samek et al., 2017). Likewise, the massive number of parameters of the deep learning system makes it difficult for even developers to annotate a complex neural network in an explainable way (Marcus, 2018). Therefore, AI designers and developers should provide some methods to simplify algorithms, so that users can interpret the mechanism and principle of complex algorithms simply.

Appropriate information disclosure. For one thing, Pasquale (2015) discussed the black box problem in various algorithms. He divided the strategies to keep the black box closed into three categories: real secrecy, legal secrecy, and obfuscation. These strategies describe disclosure insufficiency of information from the perspective that firms protect their business secrets and competitive advantages (Burrell, 2016). Due to insufficient information disclosure, users may not be able to grasp the true information of the system, such as the source of data or the probability of errors. Additionally, Grether et al. (1985) thought that the intention of corporations to increase competitive advantage and public trust through information disclosure often leads to information disclosure overload, namely, a type of behavior by which information is excessively disclosed. Due to disclosure overload, the public will fail to find information that benefits them because they cannot retrieve and extract the needed knowledge from a large amount of information (Nelson, 1994); The public will fall into boredom and anxiety because they spend more time than what is available to them, which can lead to a distance from their goal (Klapp, 1986). Therefore, AI development companies should provide appropriate information disclosure methods and legal regulatory solutions for black-box artificial intelligence systems, so that users have a sufficient and correct understanding of the system's mechanisms.

High-level collaboration. Legal scholars, social scientists, domain experts, and computer scientists should strengthen their partnerships and engage users and the public in discussions with experts on algorithms (Burrell, 2016). Companies must interact with consumers to reduce opacity when developing their products (Prahalad & Ramaswamy, 2004), and achieving responsible technology design, development, and use requires stakeholder involvement throughout the process (Ansgar et al., 2017). Low collaboration increases the opacity of products. This interaction can be understood as value co-creation because consumers trust the products that they create jointly with product developers (Prahalad & Ramaswamy, 2004). An example from the research of Prahalad and Ramaswamy is that patients were more willing to follow the treatment plan they had made with their doctors. In other words, as a product, if an AI system is produced in a low collaboration and co-creation environment, it may be a black-box product for users. Therefore, to generate appropriate understanding and trust in artificial intelligence systems, AI development should emphasize collaboration, including human-human and human-machine, and provide opportunities for stakeholders to participate in the development process, especially between non-professional users and professionals.

3. Methodology

3.1 Literature review and a concept-centric approach

The method of literature review applied in this article draws on eight-step systematic review guidance (Okoli & Schabram, 2010). I have also incorporated the concept-centric approach of Webster and Watson (2002). The eight-step review guide is adopted because it can provide a methodological way to collate the literature and express the connotation of a literature review clearly and objectively. Meanwhile, a concept-centric approach can help me better present the literature in a concise, logical statement and support me with a data basis for subsequent analyses. The description of these eight steps is shown in Table 2.

Table 2. Eight-step for the literature review

Step	Content
Step 1	Describe the purpose of the review
Step 2	Establish screening rules

Step 3	Preliminary literature screening
Step 4	Further literature screening
Step 5	Determine the final literature for review
Step 6	Literature extraction and preservation
Step 7	Data synthesis and topics determination
Step 8	Write the review report

3.2 Eight-step for the literature review

Step 1: Review Purpose. The purpose of this literature review is to sort out explainability strategies for AI systems, and then discuss how existing explainability strategies can help users better understand artificial intelligence decisions through these three dimensions.

Step 2: Protocol. I established a keyword-based search protocol to retrieve relevant literature (Appendix A.1). These keywords are searched in titles, abstracts, and keywords to expand coverage. These keywords include ‘explainable strategy’, ‘interpretable strategy’, ‘explainability’, ‘interpretability’, and so on. At the same time, to increase accuracy in the search results, I limited results to journals and conference papers. The purpose of screening journals and conference papers separately is to maximize the sample size and inclusion of retrieval. The literature search period is five years (2019-2023), and the language of the literature is English. After the preliminary search results, I will further determine the literature that needs to be reviewed based on the correlation between the literature and the topic through reading.

Step 3: Search for literature. I queried the Scopus database. Scopus includes web tools for keyword retrieval. Through the initial search process, a total of 1057 papers in journals and 1390 in conferences were retrieved. According to Rowe (2014), comprehensive coverage in the review is not reasonable. "Comprehensiveness can also mean sensemaking, which is also important, especially when a review aims at understanding and viewing a landscape of the accumulated knowledge more cohesively but without exploring all its details and thus does not require completeness in the paper's collection." (p.246). Therefore, I identified the top 100 most cited papers in journals and

50 in conferences as preliminary search results (see Appendix A.2). This selection was made to consider the contributions, breadth of application, and impact of the explainability strategies involved in the highly cited literature within this research field. Moreover, the analysis of these literature findings can provide valuable perspective for a broader range of research areas focusing on explainability strategies.

Step 4: Practical Screening. By reading the title and abstract, some literature on specific topics have been excluded, such as literature reviews on XAI and literature unrelated to research topics. I selected 73 papers in journals and 18 in conferences (Appendix A.3). The screening criteria are 1) Papers with a focus on concepts, ideas, and principles for constructing explainability strategies; 2) Papers focusing on the application of explainability strategies in different scenarios.

Step 5: Quality Screening. After further reading the content of the papers, I ultimately excluded 48 out of 73 journal papers and 6 out of 18 conference papers. These papers were excluded because they either discussed a review of existing methods without strategy construction, the application of existing XAI technology in a specific field, or are unrelated to my focus and the research question. Therefore, I retained 25 journal papers and 12 conference papers (Appendix A.4) for further analysis and discussion.

Step 6: Data Extraction. After carefully reading the paper and evaluating its relevance to my review purpose, I extracted data related to the review purpose and kept them in an Excel table.

Step 7: Data Synthesis. Based on the content of the data I extracted in step 6, I divided these 37 papers into three categories based on three dimensions: simplifying algorithms, appropriate information disclosure, and high-level collaboration. Some papers may appear repeatedly as they involve multiple dimensions. Webster and Watson's (2002) concept-centric approach was used for further data synthesis. I examined the similarities and differences between each paper, resulting in several review topics. The topics I summarized may not fully cover all the content covered in the papers. They were selected and determined based on my purpose of conducting this review and my concern for the parts with the highest correlation to specific dimensions.

Step 8: Write the review. The final step of the review is to write the review report, which mainly includes a ‘report of the review results.’

4. Findings

Because some papers involve multiple dimensions, out of these 37 papers, 26 papers involve simplifying algorithms, 4 papers involve appropriate information disclosure, and 10 papers involve high-level collaboration. Most papers contain explainability strategies related to simplifying algorithms, while relatively few papers cover the other two dimensions (Table 3).

Table 3. Existing explainability strategies on three dimensions and their topics

Three dimensions	Author	Explainability strategy
Simplifying algorithm	Amann et al. (2020)	Semantic explanation
	Zerilli et al. (2019)	
	Angelov & Soares (2020)	
	Gao et al. (2021)	
	Garcez et al. (2019)	
	Vassiliades et al. (2021)	
	Heinrichs & Eickhoff (2020)	
	Ploug & Holm (2020)	
	Clark et al. (2020)	
	Lim et al. (2021)	Architecture explanation
	Guo (2020)	
	Yeom et al. (2021)	
	Wang et al. (2019)	
	Abdul et al. (2020)	
	Kim et al. (2020)	
	Lundberg et al. (2020)	

	Panigutti et al. (2020)	Global explanation and local explanation
	Casalicchio et al. (2019)	
	Giudici & Raffinetti (2021)	
	Shankaranarayana & Runje (2019)	
	Holzinger et al. (2019)	Causal explanation and interactive explanation
	Shin (2021)	
	Holzinger et al. (2021)	
	Holzinger (2021)	
	Frye et al. (2019)	
	Weitz et al. (2019)	
Appropriate information disclosure	Reyes et al. (2020)	Professional explanation
	Amann et al. (2020)	Legal explanation
	Buiten (2019)	
	Arnold et al. (2019)	AI supplier explanation
High-level collaboration	De Bruyn et al. (2020)	Human-machine collaboration
	Sachan et al. (2020)	
	Liao et al. (2020)	
	Feng & Boyd-Graber (2019)	
	Ehsan & Riedl (2020)	
	Reyes et al. (2020)	

	Hong et al. (2020)	Human-human collaboration
	Aizenberg & Van Den Hoven (2020)	
	Ribera & Lapedriza García (2019)	
	Amann et al. (2020)	

4.1 Regarding simplifying algorithm

4.1.1 Semantic explanation.

Semantic explanation is the process of understanding the meaning and context of language or data within a particular domain. It is an important aspect of black box algorithms because it allows us to gain insight into how these algorithms are making decisions and predictions. Semantic explanations provide a way to shed light on these opaque algorithms by analyzing the data inputs and outputs and interpreting results. To facilitate people's understanding of algorithm decision-making, focusing on the overall prediction of AI is more valuable than analyzing the importance of specific features in the algorithm (Angelov & Soares, 2020). This provides us with a paradigm for thinking about semantic explanation, thereby shifting our focus from traditional feature analysis. Semantic explanations can also be implemented by embedding attention maps in specific modules of algorithms (Gao et al., 2021), IF-THEN rules (Angelov & Soares, 2020), a more natural language rule basis (Clark et al., 2020), intentional stance explanations (Zerilli et al., 2019), or constructing logical structures similar to those used in neural symbol systems (Garcez et al., 2019). Regarding semantic explanation, Vassiliades et al. (2021) focused on using the process of argumentation to translate how AI systems make decisions step by step. Although to some extent, this approach is similar to embedding discourse elements into machine learning algorithms (Heinrichs & Eickhoff, 2020), they all require more development examples to answer how they are implemented in practice. The key to semantic explanation also lies in a comprehensive review of data, biases, performance, and decision-making (Ploug & Holm, 2020), which provides people with the opportunity to have a more comprehensive understanding of what algorithms do.

4.1.2 Architecture explanation.

An overall explainability architecture should encompass both the technical aspects of the algorithm and the many factors involved in understanding and

interpreting the results. It is important to define the goals of the explainability architecture. This may include improving the accuracy and fairness of the algorithm, increasing transparency and trust in the decision-making process, and providing insights into the algorithm's internal workings. The Temporary Fusion Transformer (TFT) model based on attention architecture developed by Lim et al. (2021) can analyze the importance of variables, visualize persistent temporal relationships, and define significant regime changes. Similarly, the framework developed by Kim et al. (2020) for text classification also provides a visual approach that is easy for humans to understand. Another explainability architecture involves pruning and compressing neural networks (Guo, 2020; Yeom et al., 2021) to obtain simpler interpretable models. The Cognitive-GAM (COGAM) proposed by Abdul et al. (2020) can provide explanations with the required cognitive load and accuracy by combining expressive nonlinear generalized additive models (GAM) with simpler sparse linear models. AlphaStock based on reinforcement learning can construct an interpretable business investment strategy and logic (Wang et al., 2019).

4.1.3 Local and global explanation.

Local and global explanations are two approaches to interpreting the decisions made by black-box AI algorithms. Local explanations focus on explaining the decisions made by the model for a specific input or instance. These explanations help users understand why a particular decision was made by the model for a specific input. Local explanations can be generated using a variety of techniques, such as LIME (Local Interpretable Model-Agnostic Explanations) or SHAP (SHapley Additive exPlanations), which create surrogate models or perturbations to identify the input features that are most important for the model's decision. Global explanations aim to provide an overview of the model's behavior across the entire dataset or population. These explanations help users understand how the model behaves overall and identify any patterns or biases in its decisions. By using an automatic encoder to modify LIME, stability, and local fidelity can be improved while generating explanations (Shankaranarayana & Runje, 2019). Doctor XAI can provide local explanations to explain the principles behind individual data point classification (Panigutti et al., 2020). The combination of tree and local explanation is more helpful for experts to understand model decisions (Lundberg et al., 2020). Moreover, two visualization tools that represent the importance of local features, Partial Importance (PI) and Individual Conditional Importance (ICI) graphs, can visualize how changes in features evenly affect model performance (Casalicchio et al., 2019). Compared to local explanations, global explanations can be generated using techniques such as Partial Dependence Plots (PDP) or Accumulated Local Effects (ALE)

plots, which visualize the relationship between a feature and the model's output across the entire dataset. On the other hand, regarding global explanations, models based on the application of the Shapley method in Lorenz Zonoid can provide a unified standard for evaluating the predictive accuracy and explainability of the explanatory variables included in machine learning models. Therefore, it is theoretically easier to explain, but it needs to be validated in more environments (Giudici & Raffinetti, 2021).

4.1.4 Causal and interactive explanation.

Causal explanations aim to identify the causal relationships between the input variables and the output of the model. If a model is predicting whether a loan application will be approved, a causal explanation could identify the factors that are causing the model to approve or reject certain applications, such as credit score or income. Interactive explanations provide users with a way to interact with the model and explore its behavior in real-time. This approach is particularly useful when understanding how a model might behave under different scenarios or inputs. Interactive explanations can take many forms, such as interactive visualizations or simulations. They can be designed to allow users to adjust the inputs to the model and observe how the model responds. Holzinger et al. (2019) suggest introducing causal relationships from human models into AI models to explain the reasons for decision-making because causal relationships can reduce the high opacity of algorithms, improve model explanations (Frye et al., 2019), and improve user acceptance (Shin, 2021). Causality can also be improved by adding UX to AI to enhance human-machine interaction, thereby establishing transparent interaction and fair algorithms (Shin, 2021). This can become the foundation for establishing a human-AI interface in the future [38]. Visual output-based interaction can combine XAI methods with language information provided by virtual agents, helping to increase trust and achieve responsible artificial intelligence (Weitz et al., 2019). "We need interactive Human-AI interfaces that enable a domain expert to ask questions to understand why a machine came up with a result, and to ask what-if questions (counterfactuals) to gain insight into the underlying independent explanatory factors of a result" (Holzinger, 2021, p.175).

4.2 Regarding appropriate information disclosure

4.2.1 Professional explanation.

As artificial intelligence (AI) becomes increasingly prevalent in various industries, especially black-box AI, professionals need to promote the correct use of AI to ensure that it is being used ethically, responsibly, and effectively. This could include establishing ethical guidelines and implementing

oversight mechanisms. Professionals in specific fields, such as doctors and judges, should participate in the audit and supervision of AI usage, and before AI application, they should comprehensively evaluate the internal functions and working principles of AI (Reyes et al., 2020), to ensure the comprehensibility and interpretability of AI results.

4.2.2 Legal explanation.

Legal intervention in correct usage of AI is crucial to ensure that AI is developed and used ethically, responsibly, and in compliance with legal principles. This includes laws that ensure that AI is used in compliance with ethical principles, such as fairness, accountability, transparency, and privacy. Legal intervention can also contribute to establishing standards for data protection and cybersecurity that AI systems must comply with. The law needs to be responsible for the regulation of AI, including the acquisition and analysis of relevant data (Amann et al., 2020). However, it must be clarified that legal regulations, such as GDPR, should not only assume responsibility for AI management but also provide clearer explanations for the development of AI systems, such as the extent to which algorithms should be transparent and how much development costs AI suppliers should pay to bear such transparency (Buiten, 2019).

4.2.3 AI supplier explanation.

AI supplier explanation can play a crucial role in ensuring the correct use of artificial intelligence (AI) by providing transparency and accountability to the development and use of AI systems. AI suppliers can provide transparency to AI systems by disclosing how the AI system works, including its data sources, algorithms, and decision-making processes. This information can help stakeholders, including users, regulators, and other interested parties, to understand how the AI system is being used and to identify any potential biases or ethical concerns. AI suppliers can also establish ethical principles and guidelines for the development and use of AI systems and incorporate these principles into the design and operation of the AI system. AI suppliers can collaborate with stakeholders to ensure that AI systems are developed and used in a way that benefits society as a whole. This includes engaging with regulators, policymakers, and other interested parties to ensure that the AI system is compliant with legal and ethical frameworks. Supplier's declarations of conformity (SDoCs) are a typical example. Usually, such documents are not required by law, but the clear statements related to the purpose, performance, safety, and other contents of AI in the documents can help users implement a gradual inspection of AI to strengthen their understanding of AI products (Arnold et al., 2019).

4.3 Regarding high-level collaboration

4.3.1 Human-machine collaboration.

Human-machine collaboration can help explain the black-box of artificial intelligence (AI) by providing transparency and explainability about how the AI system works, even if the system's inner workings are not fully understood. Human-machine collaboration can provide context around the AI system, and human-machine collaboration can break down complex AI concepts into simpler terms that can be easily understood by stakeholders. This can include using analogies and examples to help explain technical concepts in a way that is accessible to non-experts, which can help build trust and understanding of AI systems for stakeholders. The transfer of tacit knowledge between humans and machines is crucial to identifying prejudices and errors, encouraging people to trust AI machines more and accept their decisions with more firm beliefs (De Bruyn et al., 2020). Incorporating knowledge based on belief rules from human experts and users into AI can help build explanatory AI systems (Sachan et al., 2020). XAI-based question banks can help connect users' demand for explainability in artificial intelligence and the technical capabilities provided by XAI (Liao et al., 2020). Similarly, the Q&A task derived from the popular trivia game Quizbowl emphasizes providing explanations to different users corresponding to their skill levels, so as to further improve the cooperation between human beings and AI (Feng & Boyd-Graber, 2019). Therefore, the key to constructing AI explainability strategies through human-machine collaboration may lie in the interaction between stakeholders and the human-machine interface community (Reyes et al., 2020), and integrating HCI strategies such as value-sensitive design and participatory design into the development of artificial intelligence that places people at the center of technology (Ehsan & Riedl, 2020).

4.3.2 Human-human collaboration.

Human-human collaboration can promote open communication and dialogue among different stakeholders, including developers, end-users, regulators, and the public. Human-human collaboration can develop clear explanations of how the black box AI works, including its purpose, inputs, outputs, and decision-making process. This can be done through various means, such as visualizations, user manuals, and technical reports, in which end-users are provided with opportunities to interact with the system. To build an explainable AI system, it is necessary for human-human collaboration to distinguish roles, processes, goals, and strategies in different organizations and AI environments (Hong et al., 2020). All stakeholders should achieve interdisciplinary and multi-perspective collaboration (Amann et al., 2020),

and effective and quality communication (Ribera et al., 2019). The advantages and disadvantages of systems should be discussed to determine their compliance with social norms (Aizenberg & Van Den Hoven, 2020), which can lead all parties to better understand and trust the AI system.

5. Discussion

5.1 Response to the research question

Q: How can existing explainability strategies help users better understand artificial intelligence decisions through these three dimensions, simplifying algorithms, appropriate information disclosure, and high-level collaboration?

Regarding simplifying algorithm, the existing explainability strategies offer six ways to explain AI: 1) Semantic explanation; 2) Architectural explanation; 3) Global explanation; 4) Local explanation; 5) Causal explanation; 6) Interactive explanation. Among them, causal, interactive, and semantic explanations prefer to explain the mechanisms of AI in a user-comprehensible manner, and local and global explanations are more oriented toward system developers and professionals. Although architecture explanation can simplify algorithms in different ways, whether it can provide an understanding of AI decision-making for users, especially those without a technical background, needs more validation.

Regarding appropriate information disclosure, the existing explainability strategies provide three ways to understand AI: 1) Professional explanation; 2) Legal explanation; and 3) AI supplier explanation. Professional explanations can evaluate, audit, and supervise the content provided by the other two explanations. The legal explanation can provide judicial supervision schemes, while AI supplier interpretation can provide information disclosure of algorithms.

Regarding high-level collaboration, existing explainability strategies provide explanations based on human-machine and human-human interactions. Human-machine interaction discusses the information transmission between humans, AI, and other components within an AI system. Human-human interaction focuses on the collaboration among all stakeholders within an AI-centered system, across multiple domains, disciplines, and departments.

5.2 Facing the challenges of explanations

Considering the challenges described in section 2.1, simplifying algorithms can make it easier for users to understand the workings and decision-making process of AI systems, thus improving comprehensibility. Although it

remains to be further validated whether the strategies involved in simplifying algorithms can fundamentally address the challenge of reliability, providing verifiable explanations through a combination of semantic explanations, local explanations, and global explanations can be considered a promising research direction for future reliable explanation strategies. At the very least, it can provide users with the opportunity to assess the reliability of system decisions. Appropriate information disclosure can increase the transparency of the system, helping users understand the internal mechanisms and decision criteria of AI systems. Improved transparency and explainability can foster trust calibration and aid users in understanding the basis of system decisions. High-level collaboration can facilitate human-machine and human-human interactions, enabling stakeholders to collaborate, share knowledge, and understand AI systems collectively. This collaboration can contribute to improved reliability and comprehensibility, as professionals and users can jointly explore and validate the decision-making process of the system while learning and understanding from each other. High-level collaboration also helps address the BDI challenge, as interactions among different stakeholders can reveal the system's beliefs, desires, and intentions, promoting shared understanding and calibration of system behavior.

5.3 Potential limitations in implementation

When it comes to simplifying algorithms, the research examined here suggests that researchers need to consider computational complexity, scalability, and performance. Some explainability strategies may introduce significant computational burdens, especially for complex models and large-scale datasets. For instance, semantic and architecture explanations might require analyzing the internals of the model, leading to increased computational costs. Global explanations and causal explanations might necessitate a comprehensive understanding of the behavior of the entire dataset or model, which could become challenging in large-scale applications and result in lower scalability. The effectiveness of interactive explanations partially depends on the feedback and guidance provided by users. However, misunderstandings and biases introduced by human factors during the interaction process could negatively impact the model's performance, which might be unacceptable in certain sensitive applications.

Regarding appropriate information disclosure, professional explanations may require domain-specific knowledge that could still be difficult for the average user to comprehend, thereby limiting the conveyance and understanding of information. Legal explanations involve the jurisdictional aspects of legal frameworks and regulations, which could reduce the general applicability of legal explanations due to regional differences. Relying on AI suppliers for

explanations is influenced by the vendors' commercial interests, making it challenging to provide comprehensive and objective explanations.

Concerning high-level collaboration, collaboration among individuals might face challenges in communication and coordination, especially when multiple domain experts are involved, and challenging each other's authority can undermine the effectiveness of communication. In addition, terminologies in different professional fields may set barriers to effective communication. Human-machine collaboration requires the design of effective interfaces and interaction methods, allowing users to comprehend the model's decision-making process. Designing user-friendly and efficient interfaces remains a challenge, however, particularly when targeting different user groups.

5.4 Validation of strategies

Ensuring the effectiveness of explainability in AI systems is crucial for promoting their sustainability of use. Despite the rich variety of approaches existing in current literature across three dimensions, empirical validation of their effectiveness is lacking. The validation of these strategies can reveal their potential benefits and limitations in practical applications and shed light on their impact on generating appropriate user trust in real-world scenarios. On one hand, the methods for validating interpretability strategies should involve experiments and case studies in various scenarios. By applying different strategies to specific decision contexts through experimental designs and case studies, and collecting user feedback, researchers can assess the impact of these strategies on user comprehension. Furthermore, validating strategies for different user groups with diverse backgrounds allows researchers to understand the pros and cons of different strategy deployments and personalized settings. On the other hand, strategy validation helps researchers evaluate their actual impact on user trust in AI systems. The validation outcomes will guide researchers in selecting suitable strategies in different contexts to enhance user trust and promote selective utilization of AI systems. Additionally, validation processes can contribute to improving strategies in turn, thereby better accommodating distinct user needs and application domains. Therefore, the significance of strategy validation lies not only in strengthening the relevance of explainability theory and practice, but also in advancing the realization of trustworthy and responsible AI.

5.5 User-centered explainability strategies

The long-term lack of research on user needs (Antoniadi et al., 2021) and much research focus on stakeholders within the AI system rather than external stakeholders (Confalonieri et al., 2021) has led to the necessity to "start from a user-centered perspective" (p.15). In terms of future research

directions for user-centered explainability strategies, several potential areas of focus relate to my findings.

One possible direction is to explore ways to combine different types of explanations to provide users with a more comprehensive understanding of AI systems. For example, a global explanation that provides an overview of how an AI system works could be combined with local explanations that explain specific predictions or decisions made by the system. Causal explanations could also be used to help users understand the reasoning behind the system's outputs. Another potential direction for research is to focus on developing more interactive and engaging explanations that use visualizations and other interactive tools. This could help to make explanations more accessible to users who are not familiar with technical jargon or complex mathematical models. Interactive explanations could also be designed to provide users with feedback and opportunities to test their understanding of the system.

Additionally, future research could focus on developing legal frameworks and guidelines for ensuring that AI systems are transparent and explainable, particularly in high-stakes applications such as healthcare or finance. Professional explanations could also be developed to help practitioners in fields such as medicine or law to understand how AI systems are being used and to make informed decisions based on their outputs. AI supplier explanations could focus on providing information about how different AI systems work and what types of explanations are available to users.

Finally, research could also explore ways to improve human-machine interaction and user experience concerning AI explainability. This could involve developing interfaces that are intuitive and easy to use, as well as providing clear and concise explanations that are tailored to the user's level of understanding. By improving the overall user experience of AI systems, researchers could help to increase user trust and adoption of these technologies.

Explainability strategies are making great strides but "there is still some way to go to meet the expectations of end-users, regulators, and the general public" (Singh et al., 2020, p.15). Explainable solutions still have limitations in increasing user trust and understanding of AI, as they may only be used as an analytical tool (Ghassemi et al., 2021). It should also be noted that designing a system that can meet the needs of both experts and ordinary users is a challenging task (Ras et al., 2022), and there is no interpretable method

that can automatically customize explanations for end users in specific fields (p.377). Future work should involve more user surveys.

5.6 Research limitation

The literature review in this article investigates literature related to explainability strategies in the AI field, to identify key themes, debates, and research gaps related to the research question. While this approach has some advantages in terms of providing a comprehensive overview of the most influential literature in the field, it also has some limitations.

One of the limitations of this approach is that paying attention to the number of citations may exclude valuable but under-cited literature that could potentially contribute to the research question such as (Dazeley et al., 2021; Markus et al., 2021; Naiseh et al., 2023; Sperrle et al., 2020). This could be due to a range of factors, such as publication bias, the relative newness of the research, or differences in citation practices across different disciplines or subfields. As a result, the literature review may not provide a fully representative or nuanced picture of the current state of research on the topic. To address this limitation, future research could conduct a more comprehensive search that includes newer or under-cited literature that may contribute to the research question. This could involve using a wider range of databases, search terms, or citation metrics to identify relevant literature. Additionally, interviews with experts in the field could provide additional insights and perspectives on the topic and help to identify emerging research trends or areas of debate that may not be fully captured by the existing literature.

Another limitation is that of a rapidly developing field such as AI and explainability, the sample size of 37 papers may reflect limited responses to the research question. Future research should consider expanding the number and scope of papers reviewed to ensure a more comprehensive literature inclusion of explainability strategies. In addition to journal and conference papers, other resources should also be reviewed, such as AI companies' technical reports, the latest releases from AI developers, and government policy updates on AI development. This may help achieve a more balanced coverage of literature on the three dimensions discussed in this article.

6. Conclusion

I conducted a literature review of existing explainability strategies. My examination has found that existing literature focuses on providing explanations through simplifying algorithms, while there is less emphasis on

providing appropriate information disclosure and encouraging high-level collaboration. Therefore, there is a need to strengthen research in these latter two aspects. While more practical validation is required to address the challenges of explanation (reliability, comprehensibility, calibration, and BDI), it can be considered to develop explainability strategies by appropriately integrating these three dimensions. The explainability strategies still harbor numerous potential limitations in their implementation, and there remains a substantial amount of work for validating these strategies in real-world environments. Nonetheless, these challenges also serve as a compass for researchers, indicating the next work of AI explainability strategies. Furthermore, future research on user-centered explainability strategies should consider the following aspects: first, paying attention to user feedback on system understanding; second, exploring how AI suppliers can better provide relevant information about the system from their perspective; third, focusing on interactive AI development between professionals and users; and fourth, customizing explanations for different users.

References

Abdul, A., von der Weth, C., Kankanhalli, M., & Lim, B. Y. (2020). COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-14.

Aizenberg, E., & Van Den Hoven, J. (2020). Designing for human rights in AI. *Big Data & Society*, 7(2), 2053951720949566.

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Upper Saddle River, NJ: Prentice Hall.

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4), 105-120.

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1), 1-9.

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, 130, 185-194.

Ansgar, K., Perez, V. E., Helena, W., Menisha, P., Sofia, C., Marina, J., & Derek, M. (2017). Editorial responsibilities arising from personalization algorithms. *The ORBIT Journal*, 1(1), 1-12.

Antoniadi, A. M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B. A., & Mooney, C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11), 5088.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5), 6-1.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.

Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. arXiv preprint arXiv:2007.05408.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security*, 57-69, p: 1. Chapman and Hall/CRC.

Buiten, M. C. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation*, 10(1), 41-59.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 2053951715622512.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, 160-169. IEEE.

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, 655-670. Springer International Publishing.

Clark, P., Tafjord, O., & Richardson, K. (2020). Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

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

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299, 103525.

De Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K. U., & Von Wangenheim, F. (2020). Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing*, 51(1), 91-105.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56-62.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, 47(5), 329-335.

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable ai: Towards a reflective sociotechnical approach. In *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, 449-466. Springer International Publishing.

Feng, S., & Boyd-Graber, J. (2019). What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 229-239.

Frye, C., Rowat, C., & Feige, I. (2019). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*.

Gao, K., Su, J., Jiang, Z., Zeng, L. L., Feng, Z., Shen, H., ... & Hu, D. (2021). Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images. *Medical image analysis*, 67, 101836.

Garcez, A. D. A., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11), e745-e750.

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, 33(01), 3681-3688.

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz eXplainable artificial intelligence. *Expert systems with applications*, 167, 114104.

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50-57.

Grether, D. M., Schwartz, A., & Wilde, L. L. (1985). The irrelevance of information overload: An analysis of search and disclosure. *S. Cal. L. Rev.*, 59, 277. Retrieve from

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120. Retrieve from https://openaccess.city.ac.uk/id/eprint/23405/8/

Guo, W. (2020). Explainable artificial intelligence for 6G: Improving trust between human and machine. *IEEE Communications Magazine*, 58(6), 39-45.

Heinrichs, B., & Eickhoff, S. B. (2020). Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Human brain mapping*, 41(6), 1435-1444.

Holzinger, A. (2021). Explainable AI and multi-modal causability in medicine. *i-com*, 19(3), 171-179.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.

Holzinger, A., Malle, B., Saranti, A., & Pfeifer, B. (2021). Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. *Information Fusion*, 71, 28-37.

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1), 1-26.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615-626.

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1), 1-9.

Jin, W. E. I. N. A., Fan, J. I. A. N. Y. U., Gromala, D., Pasquier, P., & Hamarneh, G. (2021). EUCA: the End-User-Centered Explainable AI Framework. *arXiv preprint arXiv:2102.02437*.

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems*, 134, 113302.

Klapp, O. E. (1986). Overload and boredom: Essays on the quality of life in the information society. Greenwood Publishing Group Inc.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. *American Economic Review*, 108(6), 1648-81. Retrieve from https://www.aeaweb.org/articles?id=10.1257/pandp.20181018

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80, p: 55.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, 31, 611-627.

Liang, T. P., Robert, L., Sarker, S., Cheung, C. M., Matt, C., Trenz, M., & Turel, O. (2021). Artificial intelligence and robots in individuals' lives: how to align technological possibilities and ethical issues. *Internet Research*, 31(1), 1-10.

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748-1764.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1-15.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1), 56-67.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30. Retrieve from https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113, 103655, p: 9.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2), 81.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*, 279-288.

Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N., & Ali, R. (2021). Nudging through Friction: an Approach for Calibrating Trust in Explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, 1-5. IEEE.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169, 102941.

Naiseh, M., Cemiloglu, D., Al Thani, D., Jiang, N., & Ali, R. (2021). Explainable recommendations and calibrated trust: two systematic user errors. *Computer*, 54(10), 28-37.

Nelson, M. R. (1994). We have the information you want, but getting it will cost you! held hostage by information overload. *XRDS: Crossroads, The ACM Magazine for Students*, 1(1), 11-15.

Okoli, C., & Schabram, K. (2010). A guide to conducting a systematic literature review of information systems research.

Panigutti, C., Perotti, A., & Pedreschi, D. (2020). Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 629-639.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Ploug, T., & Holm, S. (2020). The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI. *Artificial Intelligence in Medicine*, 107, 101901.

Prahalad, C. K., & Ramaswamy, V. (2004). Co-creation experiences: The next practice in value creation. *Journal of interactive marketing*, 18(3), 5-14.

Ras, G., Xie, N., Van Gerven, M., & Doran, D. (2022). Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73, 329-397.

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., Tengg-Kobligk, H. V., ... & Wiest, R. (2020). On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3), e190043.

Ribera, M., & Lapedriza García, À. (2019). Can we do better explanations? A proposal of user-centered explainable AI. CEUR Workshop Proceedings. Retrieve from https://openaccess.uoc.edu/bitstream/10609/99643/1/explainable_AI.pdf

Rowe, F. (2014). What literature review is not: diversity, boundaries and recommendations. *European Journal of Information Systems*, 23(3), 241-255.

Roy, S., Laberge, G., Roy, B., Khomh, F., Nikanjam, A., & Mondal, S. (2022) Why Don't XAI Techniques Agree? Characterizing the Disagreements Between Post-hoc Explanations of Defect Predictions. In *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 444-448. IEEE.

Sachan, S., Yang, J. B., Xu, D. L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, 144, 113100.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68.

Shankaranarayana, S. M., & Runje, D. (2019). ALIME: Autoencoder based approach for local interpretability. In *Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I 20*, 454-463. Springer International Publishing.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, 102551.

Singh, A., Sengupta, S., & Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6), 52.

Sperrle, F., El-Assady, M., Guo, G., Chau, D. H., Endert, A., & Keim, D. (2020). Should we trust (x) AI? Design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433*.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1), 44-56.

van der Linden, I., Haned, H., & Kanoulas, E. (2019). Global aggregations of local explanations for black box models. *arXiv preprint arXiv:1907.03039*.

Vassiliades, A., Bassiliades, N., & Patkos, T. (2021). Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36, e5.

Wang, J., Zhang, Y., Tang, K., Wu, J., & Xiong, Z. (2019). Alphastock: A buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 1900-1908.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, xiii-xxiii.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7-9.

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, 62(6), 70-79.

Yeom, S. K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K. R., & Samek, W. (2021). Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition*, 115, 107899.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*,1-12.

Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & technology*, 34(2), 265-288.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard?. *Philosophy & Technology*, 32, 661-683.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295-305.

Paper B

An Empirical Investigation in High-stakes Areas: User-based AI Explainability Development Principles

Shuren Yu

It has been submitted to the 32nd European Conference on Information Systems (ECIS 2024).

Abstract

Industry and academia have shown a growing interest in the explainability of AI. However, a lack of sufficient investigation into the actual needs of users for explainable solutions has resulted in some impractical approaches. Even though AI systems provide explanations, users may still not understand the reasons behind the decisions. This may be due to developers ignoring the actual needs of end users while relying on their intuition to provide explanations for AI systems. Users must understand AI decision-making and generate appropriate trust, especially in high-stakes areas. Therefore, this work focuses on conducting semi-structured interviews with AI users in four high-stakes areas: banking, education, healthcare, and justice. The aim is to understand their needs, satisfaction, and perspectives on the AI explainability of AI in their workflow. This work will provide design directions for AI developers to consider how to build trustworthy and understandable AI for users, particularly in high-stakes areas.

Keywords: User-based AI, Explainability principles, High-stakes areas, Automation-trust, XAI.

1. Introduction

The growth and contribution of Artificial Intelligence (AI) is evident in various fields, such as banking (Noreen et al., 2023), education (Hu et al., 2018), healthcare (Elemento et al., 2021), and justice (Campbell, 2020). As the practical applications of AI increase, more users are concerned about understanding the decision-making and workings of AI to ensure that they showcase their performance in the right way. Understanding how AI reaches conclusions is crucial in sensitive domains, as the consequences of errors can be fatal. As the pursuit of accuracy leads to the increased complexity of models and the technology-centric AI development blindly pursues model performance, understanding the models becomes increasingly challenging. This has sparked discussions on the establishment of methods and strategies for building explainable models.

Regarding the construction of explainability, most current research is based on the principles and techniques of Explainable AI (XAI) (Arrieta et al., 2020). Many studies have developed explainability strategies covering these principles and techniques to provide explanations of AI systems, both in theory and practice (Khan et al., 2022; Dazeley et al., 2021). However, a current issue is that most of these explanations are provided to developers and AI designers, such as heatmaps and analyses of key features, rather than to users in a way and language that they can understand. There are two significant reasons for the opacity of AI. First, "writing (and reading) code is a specialist skill" (Burrell, 2016, p.1-2), which most users do not possess. For them, reading these "explanations" is as challenging as reading the AI model itself. Second, there is a "mismatch between mathematical optimization in a high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of semantic interpretation." (p.1-2). The intent of AI professionals in developing explanations leans towards explaining the logic of machine reasoning in AI decision-making, which does not align with the actual needs for explanations from users, especially those without a technical background, and their ability to interpret these explanations. The underlying reason for this problem may be the cognitive gap between developers of AI explainability and users' understanding of the need for explanations. The evaluation of XAI solutions is usually conducted by their developers, i.e., AI or Machine Learning (ML) experts, and their intuition about what is a good explanation, rather than with end-users (van der Waa et al., 2021). Designing artificial intelligence systems that are more in line with users’ trust and understanding requires empirical investigations on the specific needs of users for explanations, especially in sensitive domains and high-stakes areas.

This paper reports on an empirical study involving nine AI system users in four high-stakes areas (banking, education, healthcare, and Justice) who participated through semi-structured interviews. The interviewees are professionals in their working field but do not have a technical background in AI. Our analysis of the interviews aims to unpack their needs, satisfaction, and perspectives regarding AI explainability at the three stages of conceptualization, construction, and measurement. Based on this analysis, this study presents a set of insights guided by the question: *How should AI developers design explainability for users who have no AI background in high-stakes areas to achieve trustworthy and understandable AI?* Following the presentation of these insights and based on the analysis of the interviews, this study discusses challenges, the broader issues of automation-trust, research limitations, and future work.

2 Background

This section, first, introduces developing explainability at the stages of conceptualization, construction, and measurement. (The conceptualization stage involves explaining the importance of constructing explanations for AI. The construction stage involves the selection of explainability methods. The measurement stage involves evaluation, providing feedback, and continuous improvement for better understanding). Following the discussion of literature related to these stages, I discuss existing explainability methods and strategies and literature on the necessity of understanding users' needs.

2.1 Three stages of understanding AI

2.1.1 Conceptualization

The integration of AI systems into various workflows and applications has become increasingly prevalent. However, before these systems are embedded in operational processes, it may be crucial to ensure that users possess a thorough understanding of how these AI systems operate and make decisions. This pre-emptive understanding can help users achieve best practices because perceived ease of use and perceived usefulness are significant factors in determining user acceptance of technologies (Davis, 1989; Bennani & Oumlil, 2014; Eze et al., 2021), while also having a positive impact on building trust and usage intentions (Choung et al., 2023; Ashfaq et al., 2020). Users are willing to use technologies they believe are beneficial to their work (Bölen, 2020) and maintain a positive attitude towards systems that make them feel easy to use and enjoyable (Ashfaq et al., 2020). Additionally, user satisfaction also depends on sufficient, precise, accurate, and up-to-date information on access to systems (Veeramootoo et al., 2018). Artificial intelligence systems sometimes have unexpected consequences, uncertainties,

and biases due to the data they are trained on. Preemptively educating users about system development information, potential pitfalls, and limitations can help alleviate these issues, as information about system accuracy, risk measurement, and timing can help users reduce overgeneralization and unintended use (Arnold et al., 2019). AI developers should share technical knowledge with the public, cultivate their computing awareness, and encourage their understanding of computing (ACM, 2018), which can help increase AI transparency and reduce ethical risks.

2.1.2 Construction

Explainability construction should not be limited to technical methods but should also consider other forms that allow users to gain a better understanding, such as stakeholder participation (Ansgar et al., 2017) and co-creation (Prahalad & Ramaswamy, 2004). Explainability should receive a broader definition. Explainability methods should be chosen with due consideration of the specific application domain. Each field has unique operational requirements, constraints, and interpretability needs. For instance, in healthcare, the IF-THEN rule may be more applicable to diagnosis explanations of diabetes (El-Sappagh et al., 2018) and allergy diagnosis (Kavya et al., 2021). In the field of education, capturing the causal relationship between input features and output labels, as well as extracting rules, can facilitate the interpretation of the model (Hooshyar et al., 2023). Attention weighting can highlight important parts of the case text to provide an explanation of the results of the judicial decision-making system (Branting et al., 2019). Both LIME and SHAP can provide effective explanations for credit risk management based on AI systems (Misheva et al., 2021). Moreover, different fields are subject to distinct regulatory frameworks and ethical considerations. Thus, the selection of explainability methods should adhere to domain-specific legal and ethical guidelines (Arrieta et al., 2020). The diversity of users is also a key factor in choosing explainability methods. The choice of an explainability method should be congruent with users' levels of technical expertise. Novices might benefit from visual explanations, while experts in data science or engineering may require more granular, technical explanations (Mohseni et al., 2021). Considering different users' knowledge structure and level can lead to more personalized explainability methods that meet their actual needs, rather than pursuing unified explainability standards.

2.1.3 Measurement

The active involvement of users in the evaluation and feedback process concerning AI system explainability in high-stakes areas is of profound significance for the continuous enhancement of AI explainability techniques and methods. A pivotal role for users is the regular evaluation of AI model

outputs. Assessing whether AI systems furnish comprehensible, relevant, and coherent explanations for their decisions or recommendations is essential (Doshi-Velez & Kim, 2017). The quality and clarity of these explanations must be scrutinized, with a focus on their effectiveness in fostering user understanding. Users can employ their domain-specific knowledge and context to gauge the adequacy of these explanations in practical scenarios (Szymanski et al, 2021). The feedback users provide regarding deficiencies in AI explainability is indispensable for the enhancement of AI systems (Chu & Shen, 2022). Addressing issues related to the intelligibility and precision of AI explanations necessitates the proactive participation of users in the improvement process. Moreover, any observed disparities or biases in AI explanations warrant diligent documentation and reporting to facilitate necessary adjustments. Another pivotal role for users may be to record and communicate the outcomes that transpire as a result of AI-driven decisions, including instances where such decisions have led to unintended, adverse, and biased effects (Barocas et al., 2017). These documented cases offer valuable insights for AI developers to rectify inadequacies in the decision-making processes, and subsequently enhance explainability. Additionally, the involvement of domain experts and ethicists who possess specialized knowledge and insights in the relevant field is a recommended approach for AI development (Ras et al., 2018). They can offer guidance and contextual understanding to users, enhancing their ability to interpret AI explanations and evaluate their appropriateness. The partnership between users and domain experts contributes to a more nuanced and thorough assessment of AI explainability.

2.2 Explainability strategies and frameworks

Supported by XAI-related goals and technologies, many explainability strategies and frameworks have been proposed in the literature. Yu (2023) proposed a three-tiered explainability framework, including converting complex algorithms into simpler forms, appropriately disclosing algorithm information to increase comprehensibility, and effective human-computer communication and collaboration. Dazeley et al. (2021) proposed a five-order explanation framework, including reactive explanation, disposition explanation, social explanation, cultural explanation, and reflective explanation. Dazeley et al. also argued that an XAI system could be designed to provide a conversational explanation, so as to "better align the cognitive process of an AI system to that of people" (p. 23). Sperrle et al. (2020) reviewed the different design dimensions of the XAI and proposed a dependency model of the XAI process, which described the different stages and stakeholders of the XAI and discussed the process of bias propagation and trust construction in the XAI. According to their study, the design

dimension can guide future systematic evaluations of the XAI together with this dependency model. In addition to XAI, several works also consider users and participants more when designing interpretive approaches. Kim et al., (2022) emphasized the need for the end users to participate in the XAI design and investigated the user's different needs for explanations through interviews with the end users. According to them, such research can be used to improve the design of the XAI explanation. According to Jin et al. (2021), the design of EUCA (End-User-Centered explainable AI framework) considers the human-centered view and provides a practical prototype tool. It helps AI system designers understand users' requirements for interpretability, to provide support for developing XAI systems that conform to end users.

However, researchers must recognize that explainability design and development work still lacks practical verification (de Bruij et al., 2022) and testing in real environments (Jin et al., 2021). "Producing explanations that fully consider user contexts and tasks remains an understudied area" (Sanneman and Shah, 2020, p. 107).

2.3 Understanding users' needs

The development of XAI is a critical aspect of advancing AI technologies and ensuring they are not only accurate but also trustworthy. Most XAI development still remains at the algorithmic level, and despite considering user needs, it is also aimed at technical users rather than end users without technical background (Jin et al., 2021). XAI inherently needs to "consider its end-users" (p. 5) to avoid "harmful unintended consequences" (Bond et al., 2019, p. 2). This highlights a fundamental issue in the field of XAI: the need to bridge the gap between the highly technical nature of XAI development and the diverse range of end-users who will interact with AI systems. AI systems should "include end-user values throughout the AI development lifecycle" (Bond et al., 2019, p. 4). This involves understanding the needs, preferences, and expectations of end-users who may not have a technical background. The presentation of explainability should be intuitive and easy to comprehend, ensuring that users can trust and rely on AI systems. For instance, creating human-machine interaction (Reyes et al., 2020), clear visualizations based on users' perceptual understanding (Ribera & Lapedriza, 2019), and common language explanations for both end-users and XAI practitioners (Jin et al., 2021) can enhance the accessibility of XAI for a broader audience. To enable non-technical users to effectively engage with XAI, there is a need for education and training initiatives. This would involve teaching individuals the fundamental concepts of XAI, such as model transparency, explainability techniques, and the benefits and limitations of AI systems. Such initiatives can empower users to make informed decisions

when interacting with AI tools, which, in turn, promotes digital literacy (Long & Magerko, 2020), democratization of the discourse (Garvey, 2018), and literacy on AI capabilities and risks (Recki et al., 2023). Existing research suggests that in addition, research should delve into understanding user needs, their cognitive processes, and the contexts in which they engage with AI.

3 Method

To address the research question, the objectives of this work are as follows: (1) Describe how users understand AI decisions and explainability in four high-stakes areas at three stages: conceptualization, construction, and measurement; (2) Summarize and conceptualize users' needs, satisfaction, and perspectives. To achieve these two objectives, firstly, this study recruited users from these four fields and conducted semi-structured interviews with them; Secondly, this study analyzed the interview data; Thirdly, this study conceptualized and summarized these findings.

3.1 Recruting participants

The recruitment process involved convenience and snowball sampling (Creswell & Poth, 2016). The participants are all experienced professionals in one of four high-stakes areas. Table 1 shows detailed information about the nine participants (five females and four males). They come from eight different organizations in China. The participants cover four areas: banking, education, healthcare, and justice. In the context where they use AI, AI has embedded explainability methods in their work practices. The participants in these areas have rich experience in AI usage during their work, but most of them have basic knowledge of AI but do not have specific technical AI expertise.

Table 1: Participants recruitment in banking, education, healthcare, and justice.

Domain	PID	Company/Institution	Title	AI Environment
Banking	P1	Bank	Credit Evaluation Manager	Credit evaluation for loans
	P2	Bank	Key Account Manager	Customer evaluation

	P3	Bank	Credit Manager	Credit and risk evaluation for loans
Education	P4	Online Education School	Teacher	Analysis and evaluation of student learning and behavior
	P5	University	Student	University admission recommendation
	P6	University	Teacher	Teaching assessment
Healthcare	P7	Hospital	Urologist	Radiological imaging detection
	P8	Hospital	Dentist	Dental radiological imaging detection
Justice	P9	Law firm	Lawyer	Legal provisions searching and case risk assessment

From left: (1) Participant's domain, (2) Participant ID (PID), (3) Participant's company or institution information, (4) Participant's position, (5) Participant's specific context for using AI tools

3.2 Question design and data analysis

The design of the questions in the interviews was guided by the principles of human-centered explainable system designing by Mueller et al. (2021) (Appendix A.5). This served as a navigation for constructing a set of questions (Appendix A.6) about explainability at three stages: conceptualization, construction, and measurement. This work used interviews because they allowed us to engage with a wider range of users without the limitations of observational or participatory approaches to the environment and interviewees. The interviews were semi-structured and audio-recorded for subsequent analysis with the consent of the participants. The average recording time of the interview was half an hour, which did not include an introduction to the work before the interview, nor did it include claims of the participants' rights and privacy protection.

The interview recordings were transcribed, translated, and organized thematically. Audio recordings were manually transcribed into text by the

author. Next, the author translated the interview language (Mandarin) into English for subsequent analysis and presentation to readers. Therefore, in the "Findings" section of this work, all quotes regarding the participant's views are in English. After the translation was completed, the author checked the match between the translation and the original text to ensure correctness and accuracy. The analyzing method was drawn on Thematic Analysis (Clarke & Braun, 2017) and used induction to capture the potential meaning of the data. According to Clarke & Braun, the purpose of thematic analysis is not only to summarize data content but also to conduct high-quality analysis and identify and interpret key features of the data under the guidance of research questions. Theme analysis has a high degree of flexibility. Various data volumes and almost any type of data can be analyzed.

3.3 Ethical considerations

In this work, I carefully considered ethical factors to ensure the rights and confidentiality of participants were protected. Before data collection, informed consent was obtained from all participants. I provided detailed explanations of the study's objectives, procedures, potential risks, and benefits, as well as their rights as participants. Participation was voluntary, and they had the option to withdraw at any time during the process. To protect participant privacy, all personal identifying information was encoded (Table 1). Access and use of the research data were strictly limited to this research.

The data only involved the user's views and suggestions about AI explainability and did not involve any personnel information other than participants, such as bank customers and hospital patients. The data also did not involve any specific AI system, meaning that during the interviews, no specific system names, operations, or analysis results were mentioned. Therefore, it did not include any analysis, evaluation, or business information about specific systems.

4 Findings

The following section will use the three stages of explainability (conceptualization, construction, and measurement) to organize the findings of the study.

4.1 Conceptualization

The successful integration of AI systems into workflows requires preemptive user education to ensure a thorough understanding of system operations, mitigate unexpected consequences, and enhance user acceptance. The

following content demonstrated the viewpoints of participants, which have formed some considerations for AI developers to construct explainability during the conceptualization stage.

4.1.1 Pre-explanations

Among all the participants, only one participant (P3) had done technical work, but none of them, including P3, possessed professional knowledge of AI. For example, "*before using them, the understanding is not very extensive"* (P2), and "*I don't have a detailed understanding of how it works*" (P9). Some participants acquire knowledge of AI systems from a wide range of external resources or internal training within the enterprise, rather than disclosing information about models or algorithms provided by AI developers or suppliers. "*This AI system was introduced into my teaching, and before that, I did not retain information from sources like AI developers or suppliers; rather, I gathered information from websites, forums, and recommendations from peers, among other sources*" (P6). P2's bank provided them with written materials and professional training on the system, and P2 stated that this was very valuable. "*Before using it, typically, our bank provides promotional materials, there might be written descriptions, and there's training as well. It is very valuable we have a bit of training now. There might be some educational materials. These explanations come from our bank*" (P2). In most participants' environments, pre-explanations from AI developers or suppliers were lacking, leading to the occurrence of "*I don't have a very detailed understanding of the working principles of AI. I just only use it*" (P1).

4.1.2 Users' practices and behaviors

Through the embedding of AI models in practice, most participants could understand the operational principles of AI models based on big data collection and analysis. For instance, "*basically, my understanding is that it's data analysis, and through a significant amount of basic analysis, I can identify certain patterns... we don't know how they design their models; we only know that based on this data*" (P1). However, most participants emphasized that the important way to familiarize themselves with the mechanisms of AI was through their work practices. For example, "*I gradually understand how it works through practical use*" (P9). Moreover, P7 pointed out the importance of understanding AI mechanisms through interaction with developers in work practice. "*Some companies may provide explanations as we use the system, and we can also seek their help when encountering issues... I've gained a better understanding of the system's principles through ongoing interaction and problem-solving with the AI developers*" (P7).

Focusing on users' behavior was another significant consideration for developing AI, as some participants expressed their familiarity with AI systems by comparing their previous experience with other software. For example, "*I understand that, after a user inputs a command, the AI system processes a large amount of data and performs calculations to reach some form of conclusion or decision... it's somewhat similar to my previous experiences with certain tools*" (P4).

4.2 Construction

The construction of explainability in AI systems should involve not only technical methods but also stakeholder participation and co-creation. The stage of construction requires a broader definition and a tailored selection of methods based on specific application domains, operational requirements, needs of interpretability, and user diversity. The following themes showcased participants' perspectives on these aspects.

4.2.1 Comparison, validation, and auxiliary tools

Regardless of whether the AI system provided explainability methods and techniques, participants reported that the way to determine the correctness of AI decisions was through repeated verification and comparison of historical data. For example, "*I determine the correctness of the AI system based on the matching of its results with historical results. If the system's output aligns with previous decisions, I believe it should be consistent and correct in the future*" (P3). P7 expressed that comparing and verifying AI conclusions based on the actual situation during surgery was the best way to determine the correctness of AI conclusions. P9 mentioned that cross-validation of different cases could help determine the correctness of AI results.

However, it was worth noting that some participants did not argue that relying on AI decision-making was the right choice, regardless of whether the system provided explainability or not. They tended to use AI decision-making as a reference or auxiliary tool. For instance, "*In my understanding, AI provides an analysis or judgment, but the actual decision-making is done by humans... in my work serves as an assistant, and I still rely on human review to confirm if the results are correct*" (P1). Therefore, at least at this stage, considering AI decision-making as an auxiliary tool rather than a fully automated final decision-making might be a more appropriate approach for participants.

4.2.2 Direct or indirect participation

Some participants expressed that directly participating in AI development was an effective way to gain a good understanding. For example, "*I've been*

involved in some of the system development projects within our bank. The development process typically involves bringing business-oriented individuals like us together and also having technical development experts... I think that being part of this process is critical" (P1). Specifically, P4 stated that users should be involved in the early stages of AI development. Some participants argued that indirect participation was a way to obtain appropriate understanding. For example, "*I would not want to be directly involved in the development of the AI system. I consider this field to involve highly specialized knowledge in mathematics, algorithms, and computing, which I may not possess... I am open to offering suggestions and feedback to the developers to make the system more user-friendly and easier to understand*" (P5).

4.2.3 Explain training data and results in effective ways

For some participants, regardless of the explanation provided by AI, they might still consider AI as a black box. "*In my opinion, for end-users, an AI system is always a black box*" (P3). P1 stated that even if they asked the developer for an explanation of the results, the response was often 'This is a black box'. "*When we inquire about why the application was declined, we often receive responses from developers like 'it's a black box model', indicating that the model considers multiple complex factors. It's not easy to point out the exact reasons for the decision*" (P1).

Most participants reported that the focus of explaining AI systems was on the source and structure of training data, as well as the results, rather than the model. "*Based on my work experience, the focus should be on simplifying and clarifying historical data and telling me the structure of historical data, and what it is consistent with. Additionally, the results should be easily matched and compared with previous historical data, and the model itself may not be as critical as long as the results are correct*" (P3). Specifically, as a doctor, P7 argued that the authenticity of training data was the most important. Developers should be responsible for interpreting data and results. "*I think developers should provide users with necessary, comprehensive, and easily understandable explanations and descriptions of data, models, and results. Additionally, as I mentioned, the explanation of the result is particularly critical. From it, I can determine whether the result has any biases*" (P4). P9 suggested two effective ways: "*If a developer came to our law firm to introduce an AI software or program, having a presentation would be a quick way to grasp the concept...One more thing I'd like to mention is that in the legal field, cases can differ significantly...If developers could somehow be involved in our law firm, perhaps for a certain period, to help us adapt to AI usage, that would be great*."

4.2.4 Personalized explanatory approaches for different scenarios

For the credit evaluation of banks, P1 said that the most effective way to explain was interaction, as the large number of files and rapidly updated business and systems make it impossible for both new and experienced employees to master all knowledge. "*Having an AI with which employees can interact in a dialogue to get immediate results and explanations of these results would greatly enhance business management efficiency and customer service*" (P1). For customer evaluation, P2 stated that video and text are the best methods because "*they might be more detailed and analytical*" (P2).

As an online teacher, P4 described that AI systems should provide application manuals presented in videos, text, and images. Similar to P1's perspectives, P4 also stated that AI decision-making should be explained through real-time interaction. When explaining the results of university admission recommendations, P5 argued that text was the best way because it was more intuitive and easier to verify. When conducting teaching and student learning assessments, P6 expressed that videos and images could provide human-centered explanations.

For the medical field, "*I think images are the most suitable format. When it comes to surgery, images provide the clearest information for decision-making, surgical choices, and various aspects during surgery. Images are the most effective way to convey this information*" (P7). As a dentist, P8 also agreed that images were the best way to explain. Specifically, it should be noted that "*patients sometimes do not want to see overly accurate images and explanations about lesions, as they may feel afraid of the unknown*" (P8).

As a lawyer, P9 said: "*In terms of the form of explanation, visual path visualization would work best for my work… Having a visual representation, like a tree diagram, showing which legal statutes it cites, would be more helpful. I believe it's better than text or dialogue. Human-machine dialogue might not be as effective since the range of responses can be limited, and AI might give the same answers to different questions. I prefer a visual knowledge map with logical connections that represent the analysis process leading to the AI's decision.*"

4.3 Measurement

Active user involvement in evaluating and providing feedback on AI system explainability, especially in high-stakes domains, is crucial for continual improvement. This involvement includes regular assessment of the quality, clarity, and effectiveness of explanations, as well as addressing deficiencies in AI explainability. Additionally, documenting biases, collaborating with

domain experts, and recording outcomes of AI-driven decisions contribute to a more nuanced and thorough assessment of AI explainability. The following content demonstrated the participants' perspective during the measurement stage.

4.3.1 Long-term verification in practice

P1 described that when a customer receives non-human credit consultation and evaluation if they receive an explanation of the evaluation results at the same time instead of only one result, the customer expresses trust in the system most of the time. "*We initially had a level of mistrust in the model itself... as we continue to use the system and receive some explanations about the results, we find that it is indeed helpful in our daily work. So, our trust in the system gradually increases*" (P3). P2 added that trust in AI came from its efficient work, as it could present complex tasks in an organized manner. Furthermore, P7 stated that whether explanations could increase trust was also related to the form of explanation. For example, the three-dimensional imaging of the lesion and the explanation of the lesion description would compensate for the information that might be missed in the two-dimensional imaging. From another perspective, as P4 described: "*I don't have much trust in the systems I'm currently using, and I don't rely on them heavily either. It doesn't significantly increase or decrease my trust in the system; it just requires time for me to verify its results*." P5 and P8 also expressed similar views. Therefore, whether explanations of AI decision-making could increase trust requires long-term verification in practice.

4.3.2 Meeting users' roles

When evaluating the explainability of AI systems, P1, P4, P5, P6, P7, and P9 stated that the user's role should be as a feedback provider. "*My role in the evaluation process is to provide feedback with the hope that the system developers will consider and adopt our feedback to further enhance the system*" (P1). P2 expressed that if a system could satisfy both the roles of an operator (interacting with the system and developers) and a beneficiary (Improving work efficiency through human-computer interaction), the system was valuable. P3 described the users' role as a 'bridge' in evaluating AI systems, as P3 collected feedback from clients and sent feedback to developers. P8 emphasized that doctors should be the 'leader' in AI system development and developers must develop systems and explainability based on the actual needs of doctors to meet their work practices. In addition to providing information feedback, P9 also hoped to become a co-creator of AI explainability and participate in actual development.

4.3.3 Factors of trust or distrust

Based on participants' work practices, when it came to which factors can trigger user trust/distrust in AI, P1 emphasized the verifiability of training data and ensured that the system was always associated with the latest data, which could increase user trust in AI. The key to increasing trust in AI systems for P2 was: "*It can have contingency plans or anticipate and address uncontrollable factors in advance*." P3 stated that the most critical factor in trusting an AI system was the response speed. P3 explained that banks need a large amount of data to evaluate a company's credit level. Calculating these data often took hours or even a day. Some companies might make significant changes during the AI evaluation period, and AI might make incorrect credit evaluations due to not adjusting calculations promptly based on these changes. This would also pose great risks to banks. P7 and P9 described that AI systems that can detect omissions in work and provide explanations could increase their trust. P4 also introduced the importance of privacy protection in increasing AI trust. On the contrary, "*if the AI system has previously provided incorrect responses, it could decrease my level of trust*" (P1).

4.4 The synthesis of findings

At the beginning of this paper, I posed the question: *How should AI developers design explainability for users who have no AI background in high-stakes areas to achieve trustworthy and understandable AI?* Based on the findings of the study (see Table 2), several suggestions can be made.

Table 2. The explainability of AI design principles based on user needs, satisfaction, and perspectives

Three stages of explainability	**Principles**	**Users' needs, satisfaction, and perspectives**
Conceptualization	Pre-explanations	Pre-explanations should be provided by AI developers or suppliers before AI is embedded in the workflow
	Users’ practices and behaviors	Before designing AI, developers should have a deep understanding of users' work practices and behavior using digital tools

Construction	Comparison, validation, and auxiliary tools	Users tend to use the repeated verification and comparison of historical data and results to verify the accuracy of AI. Developers should consider AI decision-making as an auxiliary tool rather than a fully automated final decision that instead users.
	Direct or indirect participation	For better understanding, developers should consider allowing users to directly or indirectly participate in the development process of AI
	Explain training data and results in effective ways	Developers should be responsible for explaining the composition of training data and the results of the model. Such explanations should be based on the actual work situation and process of users.
	Provide personalized explainability approaches for different scenarios	Developers should design AI based on user needs, workflows, and best practices for understanding AI decisions
Measurement	Long-term verification in practice	Whether explanations of AI decision-making can increase trust requires long-term verification in practice.
	Meeting users' roles	Allowing users to play various roles in evaluating explainability and continuous improvement, such as feedback provider, beneficiary, bridge, leader, and co-creator, can increase users' trust in the AI systems.
	Factors of trust or distrust	It is crucial to determine the trust and distrust factors of AI systems based on user-specific work practices for designing trustworthy AI.

In the *conceptualization* stage, AI developers should ensure that these systems are explainable and accessible to users without expertise in artificial intelligence. This involves the provision of clear, jargon-free explanations or pre-explanations before AI integration, helping users understand the AI system's purpose and potential consequences. Additionally, AI developers should have a deep understanding of how users work with digital tools and make decisions in high-stakes domains. This understanding forms the basis for tailoring AI systems to meet the specific needs and requirements of these users.

Moving on to the *construction* phase, AI developers should consider designing AI as a supplementary tool rather than a full-auto black-box decision-maker. Decision-making in high-stakes areas still relies heavily on manual evaluation. For non-technical users, decision-making needs to entail incorporating features that enable them to compare and validate AI-generated outcomes against historical data or other sources, facilitating the assessment of the accuracy and reliability of AI outputs. AI developers also should encourage direct or indirect user participation in an AI development process, fostering a sense of ownership and trust. Furthermore, compared to explaining AI models and complex parameters, developers should prioritize transparency by explaining the composition of training data and how it influences AI decisions. These explanations should be relatable to real-world work situations and processes of users. Additionally, AI systems should also offer personalized explanatory methods that cater to various high-stakes scenarios, ensuring that information aligns with users' unique needs and workflows.

In the final stage, *measurement* revolves around evaluating the efficacy and trustworthiness of the AI system's explainability in high-stakes areas for non-technical users. AI developers need to consider long-term validation in real-world scenarios to determine whether the explanations provided by AI systems can indeed increase trust. Continuous monitoring and evaluation are essential to refine the system and ensure it meets users' expectations. AI developers should also be aware that users should be given various roles in evaluating explainability, such as providing feedback, benefiting from the AI system, or even participating in its co-creation. This engagement enhances trust and user involvement in the system's evolution. Moreover, based on the specific work practices of users in high-risk areas, AI developers also should identify and collect factors that affect trust or distrust in artificial intelligence systems. Understanding these factors can guide AI developers to continuously improve AI systems to meet user expectations better.

5 Discussion

In this part, firstly, I argue the reflections on the findings based on the existing literature discussed above; Secondly, I explore the challenges faced by AI developers; Thirdly, I discuss how research results affect trust in AI usage; Finally, I elaborate on the limitations of this study and future work that is indicated by the findings.

5.1 Reflections on the principles

Based on the findings (Table 2), in the conceptualization stage, "Pre-explanations" can enhance the perceived ease of use and usefulness (Davis, 1989; Bennani & Oumlil, 2014; Eze et al., 2021). This can involve providing instructions about AI before integrating AI. By understanding "Users' practices and behaviors" in key domains, AI developers can Inform users about the benefits of AI in their work (Bölen, 2020) and try to make them feel enjoyable (Ashfaq et al., 2020). This will boost their willingness and attitude towards using AI. Moreover, designs based on the above two principles can provide accurate system access information, thereby increasing user satisfaction with the system (Veeramootoo et al., 2018). In the construction stage, developers should recognize that AI is "Comparison, validation, and auxiliary tools" for users during development rather than a decision-maker that makes a fully automated final decision instead of users. "Direct or indirect participation" encourages users to understand AI decision-making in the form of stakeholder participation (Ansgar et al., 2017) and value co-creation (Prahalad & Ramaswamy, 2004), while not limited to the affections of technical methods. Users need developers to "Explain training data and results in effective ways", allowing the system to comply with the legal and ethical constraints discussed by Arrieta et al. (2020). Developers should "Provide personalized explainability approaches for different scenarios" to meet the diverse needs of different users, as the knowledge structures of different users vary (Mohseni et al., 2021). Beginners who start their AI life and professionals who have been using AI for a long time may need to benefit from explanations at different levels and scopes. In the measurement stage, developers should recognize that whether explanations of AI decision-making can increase trust requires "Long-term verification in practice". Users can use their domain-specific knowledge and background to assess the adequacy of these explanations in practical situations (Szymanski et al., 2021). Developers should be aware that evaluation and feedback "Meet users' role" so that users can provide feedback of different levels and categories based on their position, to address unexpected, adverse, and biased impacts of decision-making (Barocas et al., 2017). The participation of domain experts and ethicists with relevant domain expertise and insights (Ras

et al., 2018) can also help evaluate "Factors of trust or distrust." Although these principles can help developers build explainability frameworks, they still require real-world testing (Jin et al., 2021), especially, how to promote digital literacy (Long & Magerko, 2020), and democratization of the discourse (Garvey, 2018).

5.2 Challenges for AI developers

Implementing AI design principles based on user needs, satisfaction, and perspectives in a real environment presents several difficulties and challenges for AI developers. These challenges arise from the complexities of creating AI systems that must be not only technically robust but also user-friendly and trustworthy in high-stakes scenarios.

Challenge 1: Crafting effective pre-explanations demands clear communication and the ability to translate complex AI concepts into plain language. This challenge becomes more pronounced in high-stakes domains where the potential consequences of AI decisions are significant. More importantly, helping users establish trust at the beginning is much more important than making amends after mistakes pumping out.

Challenge 2: Gaining a deep understanding of users' work practices and behaviors can be challenging, as evaluating users is often difficult. It requires developers to immerse themselves in various domains, which can be time-consuming and resource-intensive. Additionally, users may have diverse practices, making it challenging to create a one-size-fits-all AI solution.

Challenge 3: Achieving the right balance between automation and user control is another challenge. Users often require the flexibility to validate AI results and make manual decisions. Developers must design AI systems that offer this control while not overwhelming users with complex decision-making processes. This trade-off may require developers to be very familiar with business processes. However, given the vast amount of sensitive information, business processes in high-stakes areas, such as healthcare and judicial, are difficult to open to developers fully. On the other hand, people with dual backgrounds are scarce, such as those who are both doctors and developers.

Challenge 4: Allowing direct or indirect user participation in the development process is valuable but can be logistically challenging. Incorporating user feedback effectively, especially in real-time, may require well-defined processes and resources to make timely adjustments to the AI systems, which undoubtedly increases the complexity of the systems again.

Challenge 5: Creating personalized explainability approaches for different scenarios can be intricate. Developers must develop flexible AI systems that adapt to varying user needs and workflows. This challenge involves designing dynamic interfaces and explanation modules that cater to a broad range of user requirements. This challenge involves designing dynamic interfaces and interpretation modules to meet a wide range of user needs, which may involve high development costs. Moreover, some highly specialized modules may not be able to be ported to other systems after development, resulting in resource waste.

Challenge 6: Continuously verifying the effectiveness of AI explanations in practice over the long term can be resource-intensive. It demands ongoing monitoring and data analysis, which may be challenging to sustain, especially in dynamic, high-stakes environments.

Challenge 7: Encouraging users to play various roles in evaluating explainability and continuous improvement can be met with resistance or lack of user availability. Developers may struggle to convince users to actively engage in these roles, and some users may be hesitant to take on new responsibilities. After all, work outside of business flow undoubtedly increases users' workload and prolongs their working hours.

Challenge 8: Identifying the trust and distrust factors specific to user work practices can be a complex endeavor. Users' preferences and expectations can vary widely, and capturing these factors accurately necessitates in-depth research and data analysis. Developers cannot establish a standard to define, collect, and evaluate information that encompasses trust or distrust.

Challenge 9: AI developers may face resource constraints in terms of time, budget, and access to domain experts or users for feedback and validation. These constraints can limit the extent to which the principles can be implemented effectively, thereby reducing the value of evaluation and measurement.

5.3 Automation-trust calibration, resolution, and specificity

The relationship between the user's level of trust in automation and the capability of automation is called automation-trust calibration (Muir,1987). The complexity of the automation system puts the user almost always in an improper state of automation-trust calibration, namely, underusing the automation led by underestimating the ability of automation and discontinuation it, and overusing led by overestimating the ability of

automation and indiscriminately relying on it (Parasuraman & Riley, 1997). The automation-trust resolution has been described as the correspondence between trust and automation capabilities (Cohen et al., 1998). The resolution was expressed as a poor level when the range of automation capabilities does not match the range of trust and as a good level when the ranges of the two are the same. The automation-trust specificity is divided into functional specificity and temporal specificity (Lee & See, 2004). High functional specificity is described as the correspondence of trust to the subfunctions of the automated system. In contrast, low functional specificity is the correspondence of trust to the capabilities of the entire automation system. High temporal specificity means that trust corresponds to an immediate fluctuation in automation capability, whereas low temporal specificity means trust matches a long-term change in automation capability. Lee and See (2004) explained that good calibration, resolution, and high specificity could reduce the underuse and overuse of automation.

Creating appropriate trust and flexibly adjusting trust based on specific practices should be considered in the design and development of user-centered AI. In this study, the explainability based on user needs, satisfaction, and perspectives can connect to automation trust calibration, resolution, and specificity (Table 3). Providing pre-explanations about the AI system's functioning and decision-making processes before its integration into the workflow helps set clear expectations for users. This transparency can mitigate the uncertainty that often leads to inappropriate trust in automation, contributing to better trust calibration. Understanding users' work practices and behaviors allows developers to design AI systems that complement these practices, making automation a seamless part of their workflow. This user-centered approach can lead to better trust calibration and alignment with user needs. Acknowledging that users tend to verify AI outcomes through historical data and results, developers can design AI as an auxiliary tool that assists in decision-making rather than full auto-decision-making. This approach can prevent overreliance on automation, addressing trust calibration issues. Involving users in the development process can enhance their trust in the AI system. By allowing users to participate in decision-making or providing feedback, developers can better align automation capabilities with user expectations and, in turn, improve trust calibration. A clear explanation of the composition of training data and model results can help users establish an understanding of AI capability boundaries, prevent mismatches between user trust and AI capability ranges, and thus improve resolution. Moreover, developers can customize explanations based on the specific work situation of users, enhancing their trust and understanding of specific sub-functions. This can place functional specificity at a higher level. Similarly, continuously

verifying the effectiveness of explanations in real-world scenarios will place time specificity at a high level through timely correction. Allowing users to play various roles in evaluating explainability and continuous improvement can increase their trust calibration in the system and enhance temporal specificity. Determine specific factors that affect trust or distrust based on user work practices, enabling developers to customize artificial intelligence designs to adjust trust and control its matching with AI capabilities. This targeted approach enhances trust calibration and resolution.

Table 3. Design principles and automation-trust calibration, resolution, and specificity. The √ represents the three elements of automation-trust and their correlation with corresponding principles.

Three stages of explainability	**Principles**	**Calibration**	**Resolution**	**Specificity**
Conceptualization	Pre-explanations	√		
	Users' practices and behaviors	√		
Construction	Comparison, validation, and auxiliary tools	√		
	Direct or indirect participation	√		
	Explain training data and results in effective ways		√	
	Provide personalized explainability approaches for different scenarios			√
Measurement	Long-term verification in practice			√
	Meeting users' roles	√		√

	Factors of trust or distrust	√	√	

5.4 Limitations of research and future work

While the outlined research methods and ethical considerations provide a robust foundation for understanding how users perceive AI decisions and explainability in high-stakes domains, it is important to acknowledge certain limitations in this study. First, the sample size is relatively small, comprising nine participants from distinct fields. While these participants bring valuable insights, the findings may not capture the full spectrum of perspectives in each domain. Expanding the participant pool and including a more diverse range of individuals, such as those with varying levels of AI expertise or different roles within their organizations, could enhance the study's breadth. Similarly, research should also be extended to other high-risk areas to obtain more comprehensive investigations.

Second, this research primarily relies on semi-structured interviews, which may be subject to potential biases or recall errors. Multiple proofreading should be used to avoid misunderstandings caused by inaccurate wording when translating raw data. Future work should consider incorporating complementary data collection methods, such as surveys or observations, to cross-validate the findings. Future research methods should also consider quantitative analysis, which allows researchers to distinguish between the objective behavioral effects of explanation and self-perception (van der Waa et al., 2021). Additionally, the focus on high-stakes areas may limit the generalizability of the results to other domains where AI plays a significant role but may have different implications.

In future research, efforts should be made to explore the interplay between AI explainability and trust more deeply, potentially by conducting follow-up studies to understand how the identified user needs and perspectives can be translated into improved AI systems and continuously verified in real environments. Moreover, a longitudinal approach could be employed to track changes in user perceptions over time as AI technology evolves. Finally, the study could benefit from more extensive collaboration with AI developers to bridge the gap between user expectations and system design, ultimately contributing to the development of more user-centered and trustworthy AI systems.

6. Conclusion

By analyzing interviews with nine non-technical professionals from four high-risk areas, this study conceptualized some AI explainability development principles. These principles can help developers better design trustworthy and understandable AI for high-stakes areas. Implementing these principles in AI development can also be used to improve the three elements of automation-trust: calibration, resolution, and specificity. Good calibration, resolution, and high specificity could reduce the underuse and overuse of AI. However, it should be noted: Firstly, implementing these principles in reality still poses many challenges for AI developers, such as the increase in system complexity and the effective use of resources; Secondly, the limitations of research methods and data volume indicate that there is still a lot of work to be done in future related work, such as broader user surveys and domain extension; Thirdly, although this study is based on empirical evidence, the validation of these principles in real-world environments and whether they can help achieve understandable and trustworthy AI still require a long way to go, which is also a direction that future study can focus on.

Reference

ACM (2018). ACM Code of Ethics and Professional Conduct. URL: https://www.acm.org/code-of-ethics

Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019* (pp. 1078-1088). International Foundation for Autonomous Agents and Multiagent Systems.

Ansgar, K., Perez, V. E., Helena, W., Menisha, P., Sofia, C., Marina, J., & Derek, M. (2017). Editorial responsibilities arising from personalization algorithms. *The ORBIT Journal*, 1(1), 1-12.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5), 6-1.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.

Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. Telematics and Informatics, 54, 101473.
Bennani, A. E., & Oumlil, R. (2014). The Acceptance of ICT by Geriatricians reinforces the value of care for seniors in Morocco. *IBIMA Publ. J. African Res. Bus. Technol. J. African Res. Bus.* Technol, 2014(2014).

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1, 2017.

Bond, R. R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., ... & Adel, T. (2019, October). Human Centered Artificial Intelligence: Weaving UX into Algorithmic Decision Making. *In RoCHI*, 2-9.

Branting, K., Weiss, B., Brown, B., Pfeifer, C., Chakraborty, A., Ferro, L., ... & Yeh, A. (2019, June). Semi-supervised methods for explainable legal

prediction. In *Proceedings of the seventeenth international conference on artificial intelligence and law*, 22-31.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, *3*(1), 2053951715622512.

Bölen, M. C. (2020). Exploring the determinants of users' continuance intention in smartwatches. *Technology in Society*, 60, 101209.

Campbell, R. W. (2020). Artificial intelligence in the courtroom: The delivery of justice in the age of machine learning. *Colo. Tech. LJ*, *18*, 323.

Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, 39(9), 1727-1739.

Chu, H. Y., & Shen, Y. (2022, June). User Feedback Design in AI-Driven Mood Tracker Mobile Apps. In *International Conference on Human-Computer Interaction*, 346-358. Cham: Springer International Publishing.

Clarke, V., & Braun, V. (2017). Thematic analysis. *The journal of positive psychology*, 12(3), 297-298.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. In *in Proc. Command and Control Research and Technology Symp*.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.
Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, 319-340.

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2), 101666.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, *299*, 103525.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

Elemento, O., Leslie, C., Lundin, J., & Tourassi, G. (2021). Artificial intelligence in cancer research, diagnosis and therapy. *Nature Reviews Cancer*, *21*(12), 747-752.

El-Sappagh, S., Alonso, J. M., Ali, F., Ali, A., Jang, J. H., & Kwak, K. S. (2018). An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. *IEEE Access*, 6, 37371-37394.

Eze, N. U., Obichukwu, P. U., & Kesharwani, S. (2021). Perceived usefulness, perceived ease of use in ICT support and use for teachers. IETE Journal of Education, 62(1), 12-20.

Garvey, C. (2018, December). AI risk mitigation through democratic governance: Introducing the 7-dimensional AI risk horizon. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 366-367.

Hooshyar, D., Azevedo, R., & Yang, Y. (2023). Augmenting deep neural networks with symbolic knowledge: Towards trustworthy and interpretable AI for education. *arXiv preprint arXiv:2311.00393*.

Hu, S., Bhattacharya, H., Chattopadhyay, M., Aslam, N., & Shum, H. P. (2018, December). A Dual-Stream Recurrent Neural Network for Student Feedback Prediction using Kinect. In *2018 12th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, 1-8. IEEE.

Jin, W., Fan, J., Gromala, D., Pasquier, P., & Hamarneh, G. (2021). EUCA: The end-user-centered explainable AI framework. *arXiv preprint arXiv:2102.02437*.

Kavya, R., Christopher, J., Panda, S., & Lazarus, Y. B. (2021). Machine learning and XAI approaches for allergy diagnosis. *Biomedical Signal Processing and Control*, 69, 102681.

Khan, M. S., Nayebpour, M., Li, M. H., El-Amine, H., Koizumi, N., & Olds, J. L. (2022). Explainable AI: A Neurally-Inspired Decision Stack Framework. *Biomimetics*, *7*(3), 127.

Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2022). " Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. *arXiv preprint arXiv:2210.03735*.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.

Long, D., & Magerko, B. (2020, April). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1-16.

Misheva, B. H., Osterrieder, J., Hirsa, A., Kulkarni, O., & Lin, S. F. (2021). Explainable AI in credit risk management. *arXiv preprint arXiv:2103.00949*.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4), 1-45.

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human-AI systems. *arXiv preprint arXiv:2102.04972*.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6), 527-539.

Noreen, U., Shafique, A., Ahmed, Z., & Ashfaq, M. (2023). Banking 4.0: Artificial intelligence (AI) in banking industry & consumer's perspective. *Sustainability*, *15*(4), 3682.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2), 230-253.

Prahalad, C. K., & Ramaswamy, V. (2004). Co-creation experiences: The next practice in value creation. *Journal of interactive marketing*, 18(3), 5-14.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and interpretable models in computer vision and machine learning*, 19-36.

Recki, L., Lawo, D., Krauss, V., & Pins, D. (2023, August). A Qualitative Exploration of User-Perceived Risks of AI to Inform Design and Policy. *In Fröhlich, Cobus (Hg.): Mensch und Computer 2023–Workshopband, 03.-06. September 2023, Rapperswil* (SG). Gesellschaft für Informatik eV.

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F. M., Tengg-Kobligk, H. V., ... & Wiest, R. (2020). On the interpretability of artificial

intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3), e190043.

Ribera, M., & Lapedriza García, À. (2019, March). Can we do better explanations? A proposal of user-centered explainable AI. *CEUR Workshop Proceedings*.

Sanneman, L., & Shah, J. A. (2020, May). A situation awareness-based framework for design and evaluation of explainable AI. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, 94-110*. Springer, Cham. p.107.

Sperrle, F., El-Assady, M., Guo, G., Chau, D. H., Endert, A., & Keim, D. (2020). Should we trust (x) AI? Design dimensions for structured experimental evaluations. *arXiv preprint arXiv:2009.06433*.

Szymanski, M., Millecamp, M., & Verbert, K. (2021, April). Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*, 109-119.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 103404.

Veeramootoo, N., Nunkoo, R., & Dwivedi, Y. K. (2018). What determines success of an e-government service? Validation of an integrative model of e-filing continuance usage. Government information quarterly, 35(2), 161-174.

Yu, S. (2023, October). Towards Trustworthy and Understandable AI: Unraveling Explainability Strategies on Simplifying Algorithms, Appropriate Information Disclosure, and High-level Collaboration. In *Proceedings of the 26th International Academic Mindtrek Conference*, 133-143.

APPENDIX

A.1: Search string for Scopus

TITLE-ABS-KEY("explainable strategy" OR "interpretable strategy" OR "explainability" OR "interpretability" AND "AI") AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (PUBYEAR,2023) OR LIMIT-TO (PUBYEAR,2022) OR LIMIT-TO (PUBYEAR,2021) OR LIMIT-TO (PUBYEAR,2020) OR LIMIT-TO (PUBYEAR,2019)) AND (LIMIT-TO (LANGUAGE,"English"))

A.2: 100 most cited papers in journals

Author	Title	Number of citations
Barredo Arrieta A., 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	2114
Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.-I.	From local explanations to global understanding with explainable AI for trees	1508
Miller T.	Explanation in artificial intelligence: Insights from the social sciences	1486
Kelly C.J., Karthikesalingam A., Suleyman M., Corrado G., King D.	Key challenges for delivering clinical impact with artificial intelligence	546
Holzinger A., Langs G., Denk H., Zatloukal K., Müller H.	Causability and explainability of artificial intelligence in medicine	518
Ting D.S.W., Pasquale L.R., Peng L., Campbell J.P., Lee A.Y., Raman R., Tan G.S.W., Schmetterer L., Keane P.A., Wong T.Y.	Artificial intelligence and deep learning in ophthalmology	493
Linardatos P., Papastefanopoulos V., Kotsiantis S.	Explainable ai: A review of machine learning interpretability methods	462

Shin D.	The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI	209
Amann J., Blasimme A., Vayena E., Frey D., Madai V.I., the Precise4Q consortium	Explainability for artificial intelligence in healthcare: a multidisciplinary perspective	209
Yang K.-C., Varol O., Davis C.A., Ferrara E., Flammini A., Menczer F.	Arming the public with artificial intelligence to counter social bots	208
Singh A., Sengupta S., Lakshminarayanan V.	Explainable deep learning models in medical image analysis	184
Ntoutsi E., Fafalios P., Gadiraju U., Iosifidis V., Nejdl W., Vidal M.-E., Ruggieri S., Turini F., Papadopoulos S., Krasanakis E., Kompatsiaris I., Kinder-Kurlanda K., Wagner C., Karimi F., Fernandez M., Alani H., Berendt B., Kruegel T., Heinze C., Broelemann K., Kasneci G., Tiropanis T., Staab S.	Bias in data-driven artificial intelligence systems—An introductory survey	173
Lim B., Arık S.Ö., Loeff N., Pfister T.	Temporal Fusion Transformers for interpretable multi-horizon time series forecasting	167
Zhou S.K., Greenspan H., Davatzikos C., Duncan J.S., Van Ginneken B., Madabhushi A., Prince J.L., Rueckert D., Summers R.M.	A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies with Progress Highlights, and Future Promises	166
Ghassemi M., Oakden-Rayner L., Beam A.L.	The false hope of current approaches to explainable artificial intelligence in health care	162
Holzinger A., Malle B., Saranti A., Pfeifer B.	Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI	158
Shrestha Y.R., Ben-Menahem S.M., von Krogh G.	Organizational Decision-Making Structures in the Age of Artificial Intelligence	151
Reyes M., Meier R., Pereira S., Silva C.A., Dahlweid F.-M., Tengg-Kobligk H.V., Summers	On the interpretability of artificial intelligence in radiology: Challenges and opportunities	149

R.M., Wiest R.		
Ma L., Sun B.	Machine learning and AI in marketing – Connecting computing power to human insights	134
Yang G., Ye Q., Xia J.	Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond	129
Castiglioni I., Rundo L., Codari M., Di Leo G., Salvatore C., Interlenghi M., Gallivanone F., Cozzi A., D'Amico N.C., Sardanelli F.	AI applications to medical images: From machine learning to deep learning	126
Ding Y., Zhu Y., Feng J., Zhang P., Cheng Z.	Interpretable spatio-temporal attention LSTM model for flood forecasting	112
Yang Y.J., Bang C.S.	Application of artificial intelligence in gastroenterology	111
Zerilli J., Knott A., Maclaurin J., Gavaghan C.	Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?	106
Markus A.F., Kors J.A., Rijnbeek P.R.	The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies	105
Chen M., Liu Q., Chen S., Liu Y., Zhang C.-H., Liu R.	XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System	101
Angelov P., Soares E.	Towards explainable deep neural networks (xDNN)	101
Shaw J., Rudzicz F., Jamieson T., Goldfarb A.	Artificial Intelligence and the Implementation Challenge	99
Gao K., Su J., Jiang Z., Zeng L.-L., Feng Z., Shen H., Rong P., Xu X., Qin J., Yang Y., Wang W., Hu D.	Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images	96
Riedl M.O.	Human-centered artificial intelligence and machine learning	96
Arnold M., Piorkowski D., Reimer D., Richards J., Tsay J., Varshney K.R., Bellamy R.K.E., Hind M., Houde S., Mehta S.,	FactSheets: Increasing trust in AI services through supplier's declarations of conformity	91

Mojsilovic A., Nair R., Ramamurthy K.N., Olteanu A.		
Angelov P.P., Soares E.A., Jiang R., Arnold N.I., Atkinson P.M.	Explainable artificial intelligence: an analytical review	90
Thieme A., Belgrave D., Doherty G.	Machine Learning in Mental Health: A systematic review of the HCI literature to support the development of effective and implementable ML Systems	89
Spinner T., Schlegel U., Schäfer H., El-Assady M.	ExplAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning	88
Páez A.	The Pragmatic Turn in Explainable Artificial Intelligence (XAI)	87
Belle V., Papantonis I.	Principles and Practice of Explainable Machine Learning	87
De Bruyn A., Viswanathan V., Beh Y.S., Brock J.K.-U., von Wangenheim F.	Artificial Intelligence and Marketing: Pitfalls and Opportunities	82
Wang F., Preininger A.	AI in Health: State of the Art, Challenges, and Future Directions	80
Baryannis G., Dani S., Antoniou G.	Predicting supply chain risks using machine learning: The trade-off between performance and interpretability	79
Shi Z., Yao W., Li Z., Zeng L., Zhao Y., Zhang R., Tang Y., Wen J.	Artificial intelligence techniques for stability analysis and control in smart grids: Methodologies, applications, challenges and future directions	78
Coeckelbergh M.	Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability	78
Antoniadi A.M., Du Y., Guendouz Y., Wei L., Mazo C., Becker B.A., Mooney C.	Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review	74
Garcez A.D., Gori M., Lamb L.C., Serafini L., Spranger M., Tran S.N.	Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning	73
Jiang Y., Yang M., Wang S., Li X., Sun Y.	Emerging role of deep learning-based artificial intelligence in tumor pathology	72
Bera K., Braman N., Gupta A., Velcheti V., Madabhushi A.	Predicting cancer outcomes with radiomics and artificial intelligence in radiology	70
Siau K., Wang W.	Artificial intelligence (AI) Ethics: Ethics of	68

	AI and ethical AI	
Shan T., Tay F.R., Gu L.	Application of Artificial Intelligence in Dentistry	67
Confalonieri R., Coba L., Wagner B., Besold T.R.	A historical perspective of explainable Artificial Intelligence	67
Guo W.	Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine	66
Shin D.	User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability	66
Yeom S.-K., Seegerer P., Lapuschkin S., Binder A., Wiedemann S., Müller K.-R., Samek W.	Pruning by explaining: A novel criterion for deep neural network pruning	66
Lo Piano S.	Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward	65
Magesh P.R., Myloth R.D., Tom R.J.	An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery	64
Meske C., Bunde E., Schneider J., Gersch M.	Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities	63
Cao R., Yang F., Ma S.-C., Liu L., Zhao Y., Li Y., Wu D.-H., Wang T., Lu W.-J., Cai W.-J., Zhu H.-B., Guo X.-J., Lu Y.-W., Kuang J.-J., Huan W.-J., Tang W.-M., Huang K., Huang J., Yao J., Dong Z.-Y.	Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer	63
Wang Y., Wang H., Peng Z.	Rice diseases detection and classification using attention based neural network and bayesian optimization	61
Schwendicke F., Singh T., Lee J.-H., Gaudin R., Chaurasia A., Wiegand T., Uribe S., Krois J.	Artificial intelligence in dental research: Checklist for authors, reviewers, readers	60
Hong S.R., Hullman J., Bertini E.	Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs	60
Nensa F., Demircioglu A., Rischpler C.	Artificial intelligence in nuclear medicine	57
van de Poel I.	Embedding Values in Artificial Intelligence (AI) Systems	55

Papadimitroulas P., Brocki L., Christopher Chung N., Marchadour W., Vermet F., Gaubert L., Eleftheriadis V., Plachouris D., Visvikis D., Kagadis G.C., Hatt M.	Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization	53
de Fine Licht K., de Fine Licht J.	Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy	53
Calegari R., Ciatto G., Omicini A.	On the integration of symbolic and sub-symbolic techniques for XAI: A survey	53
Jin Y., Qin C., Huang Y., Zhao W., Liu C.	Multi-domain modeling of atrial fibrillation detection with twin attentional convolutional long short-term memory neural networks	52
Singh R.K., Pandey R., Babu R.N.	COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays	49
Sachan S., Yang J.-B., Xu D.-L., Benavides D.E., Li Y.	An explainable AI decision-support-system to automate loan underwriting	49
Wang H., Wang L., Lee E.H., Zheng J., Zhang W., Halabi S., Liu C., Deng K., Song J., Yeom K.W.	Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures	48
Pessach D., Singer G., Avrahami D., Chalutz Ben-Gal H., Shmueli E., Ben-Gal I.	Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming	48
Kuzlu M., Cali U., Sharma V., Güler Ö.	Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools	47
Cearns M., Hahn T., Baune B.T.	Recommendations and future directions for supervised machine learning in psychiatry	47
Buhrmester V., Münch D., Arens M.	Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey	46
Giudici P., Raffinetti E.	Shapley-Lorenz eXplainable Artificial Intelligence	45
Kim B., Park J., Suh J.	Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information	45
Montani S., Striani M.	Artificial Intelligence in Clinical Decision Support: a Focused Literature Survey	45

Holzinger A., Dehmer M., Emmert-Streib F., Cucchiara R., Augenstein I., Ser J.D., Samek W., Jurisica I., Díaz-Rodríguez N.	Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence	45
Vassiliades A., Bassiliades N., Patkos T.	Argumentation and explainable artificial intelligence: A survey	43
Heinrichs B., Eickhoff S.B.	Your evidence? Machine learning algorithms for medical diagnosis and prediction	43
Henman P.	Improving public services using artificial intelligence: possibilities, pitfalls, governance	41
Islam M.R., Ahmed M.U., Barua S., Begum S.	A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks	40
Mousavi S., Afghah F., Acharya U.R.	HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks	40
Ploug T., Holm S.	The four dimensions of contestable AI diagnostics- A patient-centric approach to explainable AI	40
Gomolin A., Netchiporouk E., Gniadecki R., Litvinov I.V.	Artificial Intelligence Applications in Dermatology: Where Do We Stand?	40
Gilvary C., Madhukar N., Elkhader J., Elemento O.	The Missing Pieces of Artificial Intelligence in Medicine	40
Buiten M.C.	Towards intelligent regulation of artificial intelligence	40
Ras G., Xie N., van Gerven M., Doran D.	Explainable Deep Learning: A Field Guide for the Uninitiated	39
Salahuddin Z., Woodruff H.C., Chatterjee A., Lambin P.	Transparency of deep neural networks for medical image analysis: A review of interpretability methods	39
Joshi G., Walambe R., Kotecha K.	A Review on Explainability in Multimodal Deep Neural Nets	39
Hacker P., Krestel R., Grundmann S., Naumann F.	Explainable AI under contract and tort law: legal incentives and technical challenges	39
Doleck T., Lemay D.J., Basnet R.B., Bazelais P.	Predictive analytics in education: a comparison of deep learning frameworks	39
Emmert-Streib F., Yli-Harja O., Dehmer M.	Explainable artificial intelligence and machine learning: A reality rooted perspective	38
Yeung C., Tsai J.-M., King B., Kawagoe Y., Ho D.,	Elucidating the Behavior of Nanophotonic Structures through Explainable Machine	38

Knight M.W., Raman A.P.	Learning Algorithms	
Arya V., Bellamy R.K.E., Chen P.-Y., Dhurandhar A., Hind M., Hoffman S.C., Houde S., Liao Q.V., Luss R., Mojsilović A., Mourad S., Pedemonte P., Raghavendra R., Richards J.T., Sattigeri P., Shanmugam K., Singh M., Varshney K.R., Wei D., Zhang Y.	Ai explainability 360: An extensible toolkit for understanding data and machine learning models	38
Wing J.M.	Trustworthy AI	36
Sermesant M., Delingette H., Cochet H., Jaïs P., Ayache N.	Applications of artificial intelligence in cardiovascular imaging	36
Trocin C., Mikalef P., Papamitsiou Z., Conboy K.	Responsible AI for Digital Health: a Synthesis and a Research Agenda	36
Song X., Yu A.S.L., Kellum J.A., Waitman L.R., Matheny M.E., Simpson S.Q., Hu Y., Liu M.	Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction	36
Holzinger A.	Explainable AI and Multi-Modal Causability in Medicine	36
Aizenberg E., van den Hoven J.	Designing for human rights in AI	35
Ward T.M., Mascagni P., Ban Y., Rosman G., Padoy N., Meireles O., Hashimoto D.A.	Computer vision in surgery	34
Razavi S.	Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling	33

50 most cited papers in conferences

Authors	Title	Number of citations
Mittelstadt B.; Russell C.; Wachter S.	Explaining explanations in AI	297

Liao Q.V.; Gruen D.; Miller S.	Questioning the AI: Informing Design Practices for Explainable AI User Experiences	245
Anjomshoae S.; Calvaresi D.; Najjar A.; Främling K.	Explainable agents and robots: Results from a systematic literature review	200
Zhang Y.; Vera Liao Q.; Bellamy R.K.E.	Efect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making	184
Xu F.; Uszkoreit H.; Du Y.; Fan W.; Zhao D.; Zhu J.	Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges	162
Sokol K.; Flach P.	Explainability fact sheets: A framework for systematic assessment of explainable approaches	126
Ehsan U.; Tambwekar P.; Chan L.; Harrison B.; Riedl M.O.	Automated rationale generation: A technique for explainable AI and its effects on human perceptions	103
Ehsan U.; Liao Q.V.; Muller M.; Riedl M.O.; Weisz J.D.	Expanding explainability: Towards social transparency in ai systems	98
Toreini E.; Aitken M.; Coopamootoo K.; Elliott K.; Zelaya C.G.; van Moorsel A.	The relationship between trust in AI and trustworthy machine learning technologies	92
Cambria E.; Liu Q.; Decherchi S.; Xing F.; Kwok K.	SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis	91
Panigutti C.; Perotti A.; Pedreschi D.	Doctor XAI An ontology-based approach to black-box sequential data classification explanations	89
Casalicchio G.; Molnar C.; Bischl B.	Visualizing the feature importance for black box models	82
Muhammad M.B.; Yeasin M.	Eigen-CAM: Class Activation Map using Principal Components	75
Clark P.; Tafjord O.; Richardson K.	Transformers as soft reasoners over language	74
Gade K.; Geyik S.C.; Kenthapadi K.; Mithal V.; Taly A.	Explainable AI in industry	73
Aggarwal A.; Lohia P.; Nagar S.; Dey K.; Saha D.	Black box fairness testing of machine learning models	72
Longo L.; Goebel R.; Lecue F.; Kieseberg P.; Holzinger A.	Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions	71
Hase P.; Bansal M.	Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?	66

Ribera M.; Lapedriza A.	Can we do better explanations? A proposal of user-centered explainable AI	63
Wolf C.T.	Explainability scenarios: Towards scenario-based XAI design	62
Schlegel U.; Arnout H.; El-Assady M.; Oelke D.; Keim D.A.	Towards a rigorous evaluation of XAI methods on time series	62
Weitz K.; Schiller D.; Schlagowski R.; Huber T.; André E.	I"do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design	60
Feng S.; Boyd-Graber J.	What can Ai do for me? Evaluating machine learning interpretations in cooperative play	55
Buiten M.C.	Towards intelligent regulation of artificial intelligence	54
Frye C.; Rowat C.; Feige I.	Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability	53
Adadi A.; Berrada M.	Explainable AI for Healthcare: From Black Box to Interpretable Models	51
Wang J.; Zhang Y.; Tang K.; Wu J.; Xiong Z.	AlphaStock: A buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks	50
Ehsan U.; Riedl M.O.	Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach	48
Holzinger A.	The Next Frontier: AI We Can Really Trust	47
Miranda-Escalada A.; Gonzalez-Agirre A.; Armengol-Estapé J.; Krallinger M.	Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at CodiEsp track of CLEF eHealth 2020	44
Lamb L.C.; d'Avila Garcez A.; Gori M.; Prates M.O.R.; Avelar P.H.C.; Vardi M.Y.	Graph neural networks meet neural-symbolic computing: A survey and perspective	43
Fidel G.; Bitton R.; Shabtai A.	When Explainability Meets Adversarial Learning: Detecting Adversarial Examples using SHAP Signatures	42
Jesus S.; Belém C.; Balayan V.; Bento J.; Saleiro P.; Bizarro P.; Gama J.	How can i choose an explainer?: An Application-grounded Evaluation of Post-hoc Explanations	40
Tamburri D.A.	Sustainable MLOps: Trends and Challenges	39
Ye Q.; Xia J.; Yang G.	Explainable AI for COVID-19 CT Classifiers: An initial comparison study	39

Schneeberger D.; Stöger K.; Holzinger A.	The European Legal Framework for Medical AI	36
Shankaranarayana S.M.; Runje D.	ALIME: Autoencoder based approach for local interpretability	36
Chromik M.; Eiband M.; Buchner F.; Krüger A.; Butz A.	I Think i Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI	34
Kasirzadeh A.; Smart A.	The use and misuse of counterfactuals in ethical machine learning	33
Holzinger A.; Weippl E.; Tjoa A.M.; Kieseberg P.	Digital Transformation for Sustainable Development Goals (SDGs) - A Security, Safety and Privacy Perspective on AI	33
Rodriguez-Diaz E.; Baffy G.; Lo W.-K.; Mashimo H.; Vidyarthi G.; Mohapatra S.S.; Singh S.K.	Real-time artificial intelligence–based histologic classification of colorectal polyps with augmented visualization	33
Ehsan U.; Wintersberger P.; Liao Q.V.; Mara M.; Streit M.; Wachter S.; Riener A.; Riedl M.O.	Operationalizing Human-Centered Perspectives in Explainable AI	33
Abdul A.; Von Der Weth C.; Kankanhalli M.; Lim B.Y.	COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations	31
Sanchez-Lengeling B.; Wei J.; Lee B.; Reif E.; Wang P.Y.; Qian W.W.; McCloskey K.; Colwell L.; Wiltschko A.	Evaluating attribution for graph neural networks	31
Zucco C.; Liang H.; Fatta G.D.; Cannataro M.	Explainable Sentiment Analysis with Applications in Medicine	31
Pal A.; Sankarasubbu M.	Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing	31
Szczepanski M.; Choras M.; Pawlicki M.; Kozik R.	Achieving Explainability of Intrusion Detection System by Hybrid Oracle-Explainer Approach	30
Dhanorkar S.; Wolf C.T.; 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	30
Wu Y.; Kuang K.; Zhang Y.; Liu X.; Sun C.; Xiao J.; Zhuang Y.; Si L.; Wu F.	De-biased court's view generation with causality	30

Brennen A.	What do people really want when they say they want "explainable AI?" we asked 60 stakeholders	29

A.3: 73 papers in journals

Author	Title	Number of citations
Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.-I.	From local explanations to global understanding with explainable AI for trees	1508
Miller T.	Explanation in artificial intelligence: Insights from the social sciences	1486
Holzinger A., Langs G., Denk H., Zatloukal K., Müller H.	Causability and explainability of artificial intelligence in medicine	518
Shin D.	The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI	209
Amann J., Blasimme A., Vayena E., Frey D., Madai V.I., the Precise4Q consortium	Explainability for artificial intelligence in healthcare: a multidisciplinary perspective	209
Yang K.-C., Varol O., Davis C.A., Ferrara E., Flammini A., Menczer F.	Arming the public with artificial intelligence to counter social bots	208
Singh A., Sengupta S., Lakshminarayanan V.	Explainable deep learning models in medical image analysis	184
Ntoutsi E., Fafalios P., Gadiraju U., Iosifidis V., Nejdl W., Vidal M.-E., Ruggieri S., Turini F., Papadopoulos S., Krasanakis E., Kompatsiaris I., Kinder-Kurlanda K., Wagner C., Karimi F., Fernandez M., Alani H., Berendt B., Kruegel T., Heinze C., Broelemann K., Kasneci G., Tiropanis T., Staab S.	Bias in data-driven artificial intelligence systems—An introductory survey	173

Lim B., Arık S.Ö., Loeff N., Pfister T.	Temporal Fusion Transformers for interpretable multi-horizon time series forecasting	167
Ghassemi M., Oakden-Rayner L., Beam A.L.	The false hope of current approaches to explainable artificial intelligence in health care	162
Holzinger A., Malle B., Saranti A., Pfeifer B.	Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI	158
Shrestha Y.R., Ben-Menahem S.M., von Krogh G.	Organizational Decision-Making Structures in the Age of Artificial Intelligence	151
Reyes M., Meier R., Pereira S., Silva C.A., Dahlweid F.-M., Tengg-Kobligk H.V., Summers R.M., Wiest R.	On the interpretability of artificial intelligence in radiology: Challenges and opportunities	149
Ma L., Sun B.	Machine learning and AI in marketing – Connecting computing power to human insights	134
Yang G., Ye Q., Xia J.	Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond	129
Ding Y., Zhu Y., Feng J., Zhang P., Cheng Z.	Interpretable spatio-temporal attention LSTM model for flood forecasting	112
Zerilli J., Knott A., Maclaurin J., Gavaghan C.	Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?	106
Markus A.F., Kors J.A., Rijnbeek P.R.	The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies	105
Chen M., Liu Q., Chen S., Liu Y., Zhang C.-H., Liu R.	XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System	101

Angelov P., Soares E.	Towards explainable deep neural networks (xDNN)	101
Gao K., Su J., Jiang Z., Zeng L.-L., Feng Z., Shen H., Rong P., Xu X., Qin J., Yang Y., Wang W., Hu D.	Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images	96
Riedl M.O.	Human-centered artificial intelligence and machine learning	96
Arnold M., Piorkowski D., Reimer D., Richards J., Tsay J., Varshney K.R., Bellamy R.K.E., Hind M., Houde S., Mehta S., Mojsilovic A., Nair R., Ramamurthy K.N., Olteanu A.	FactSheets: Increasing trust in AI services through supplier's declarations of conformity	91
Angelov P.P., Soares E.A., Jiang R., Arnold N.I., Atkinson P.M.	Explainable artificial intelligence: an analytical review	90
Spinner T., Schlegel U., Schäfer H., El-Assady M.	ExplAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning	88
Páez A.	The Pragmatic Turn in Explainable Artificial Intelligence (XAI)	87
Belle V., Papantonis I.	Principles and Practice of Explainable Machine Learning	87
De Bruyn A., Viswanathan V., Beh Y.S., Brock J.K.-U., von Wangenheim F.	Artificial Intelligence and Marketing: Pitfalls and Opportunities	82
Baryannis G., Dani S., Antoniou G.	Predicting supply chain risks using machine learning: The trade-off between performance and interpretability	79
Coeckelbergh M.	Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability	78
Antoniadi A.M., Du Y., Guendouz Y., Wei L., Mazo C., Becker B.A., Mooney C.	Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review	74

Garcez A.D., Gori M., Lamb L.C., Serafini L., Spranger M., Tran S.N.	Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning	73
Bera K., Braman N., Gupta A., Velcheti V., Madabhushi A.	Predicting cancer outcomes with radiomics and artificial intelligence in radiology	70
Siau K., Wang W.	Artificial intelligence (AI) Ethics: Ethics of AI and ethical AI	68
Confalonieri R., Coba L., Wagner B., Besold T.R.	A historical perspective of explainable Artificial Intelligence	67
Guo W.	Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine	66
Shin D.	User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability	66
Yeom S.-K., Seegerer P., Lapuschkin S., Binder A., Wiedemann S., Müller K.-R., Samek W.	Pruning by explaining: A novel criterion for deep neural network pruning	66
Lo Piano S.	Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward	65
Magesh P.R., Myloth R.D., Tom R.J.	An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery	64
Meske C., Bunde E., Schneider J., Gersch M.	Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities	63
Cao R., Yang F., Ma S.-C., Liu L., Zhao Y., Li Y., Wu D.-H., Wang T., Lu W.-J., Cai W.-J., Zhu H.-B., Guo X.-J., Lu Y.-W., Kuang J.-J., Huan W.-J., Tang W.-M., Huang K., Huang J., Yao J., Dong Z.-Y.	Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in Colorectal Cancer	63

Wang Y., Wang H., Peng Z.	Rice diseases detection and classification using attention based neural network and bayesian optimization	61
Hong S.R., Hullman J., Bertini E.	Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs	60
Nensa F., Demircioglu A., Rischpler C.	Artificial intelligence in nuclear medicine	57
Papadimitroulas P., Brocki L., Christopher Chung N., Marchadour W., Vermet F., Gaubert L., Eleftheriadis V., Plachouris D., Visvikis D., Kagadis G.C., Hatt M.	Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization	53
de Fine Licht K., de Fine Licht J.	Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy	53
Jin Y., Qin C., Huang Y., Zhao W., Liu C.	Multi-domain modeling of atrial fibrillation detection with twin attentional convolutional long short-term memory neural networks	52
Singh R.K., Pandey R., Babu R.N.	COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X-rays	49
Sachan S., Yang J.-B., Xu D.-L., Benavides D.E., Li Y.	An explainable AI decision-support-system to automate loan underwriting	49
Pessach D., Singer G., Avrahami D., Chalutz Ben-Gal H., Shmueli E., Ben-Gal I.	Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming	48
Kuzlu M., Cali U., Sharma V., Güler Ö.	Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools	47
Buhrmester V., Münch D., Arens M.	Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey	46
Giudici P., Raffinetti E.	Shapley-Lorenz eXplainable Artificial Intelligence	45

Kim B., Park J., Suh J.	Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information	45
Montani S., Striani M.	Artificial Intelligence in Clinical Decision Support: a Focused Literature Survey	45
Holzinger A., Dehmer M., Emmert-Streib F., Cucchiara R., Augenstein I., Ser J.D., Samek W., Jurisica I., Díaz-Rodríguez N.	Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence	45
Vassiliades A., Bassiliades N., Patkos T.	Argumentation and explainable artificial intelligence: A survey	43
Heinrichs B., Eickhoff S.B.	Your evidence? Machine learning algorithms for medical diagnosis and prediction	43
Henman P.	Improving public services using artificial intelligence: possibilities, pitfalls, governance	41
Mousavi S., Afghah F., Acharya U.R.	HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks	40
Ploug T., Holm S.	The four dimensions of contestable AI diagnostics- A patient-centric approach to explainable AI	40
Gilvary C., Madhukar N., Elkhader J., Elemento O.	The Missing Pieces of Artificial Intelligence in Medicine	40
Buiten M.C.	Towards intelligent regulation of artificial intelligence	40
Ras G., Xie N., van Gerven M., Doran D.	Explainable Deep Learning: A Field Guide for the Uninitiated	39
Hacker P., Krestel R., Grundmann S., Naumann F.	Explainable AI under contract and tort law: legal incentives and technical challenges	39
Doleck T., Lemay D.J., Basnet R.B., Bazelais P.	Predictive analytics in education: a comparison of deep learning frameworks	39

Emmert-Streib F., Yli-Harja O., Dehmer M.	Explainable artificial intelligence and machine learning: A reality rooted perspective	38
Arya V., Bellamy R.K.E., Chen P.-Y., Dhurandhar A., Hind M., Hoffman S.C., Houde S., Liao Q.V., Luss R., Mojsilović A., Mourad S., Pedemonte P., Raghavendra R., Richards J.T., Sattigeri P., Shanmugam K., Singh M., Varshney K.R., Wei D., Zhang Y.	Ai explainability 360: An extensible toolkit for understanding data and machine learning models	38
Wing J.M.	Trustworthy AI	36
Song X., Yu A.S.L., Kellum J.A., Waitman L.R., Matheny M.E., Simpson S.Q., Hu Y., Liu M.	Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction	36
Holzinger A.	Explainable AI and Multi-Modal Causability in Medicine	36
Aizenberg E., van den Hoven J.	Designing for human rights in AI	35

18 papers in conferences

Authors	Title	Number of citations
Liao Q.V.; Gruen D.; Miller S.	Questioning the AI: Informing Design Practices for Explainable AI User Experiences	245
Panigutti C.; Perotti A.; Pedreschi D.	Doctor XAI An ontology-based approach to black-box sequential data classification explanations	89
Casalicchio G.; Molnar C.; Bischl B.	Visualizing the feature importance for black box models	82
Muhammad M.B.; Yeasin M.	Eigen-CAM: Class Activation Map using Principal Components	75
Clark P.; Tafjord O.; Richardson K.	Transformers as soft reasoners over language	74

Ribera M.; Lapedriza A.	Can we do better explanations? A proposal of user-centered explainable AI	63
Weitz K.; Schiller D.; Schlagowski R.; Huber T.; André E.	I"do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design	60
Feng S.; Boyd-Graber J.	What can Ai do for me? Evaluating machine learning interpretations in cooperative play	55
Buiten M.C.	Towards intelligent regulation of artificial intelligence	54
Frye C.; Rowat C.; Feige I.	Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability	53
Wang J.; Zhang Y.; Tang K.; Wu J.; Xiong Z.	AlphaStock: A buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks	50
Ehsan U.; Riedl M.O.	Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach	48
Ye Q.; Xia J.; Yang G.	Explainable AI for COVID-19 CT Classifiers: An initial comparison study	39
Shankaranarayana S.M.; Runje D.	ALIME: Autoencoder based approach for local interpretability	36
Chromik M.; Eiband M.; Buchner F.; Krüger A.; Butz A.	I Think i Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI	34
Rodriguez-Diaz E.; Baffy G.; Lo W.-K.; Mashimo H.; Vidyarthi G.; Mohapatra S.S.; Singh S.K.	Real-time artificial intelligence–based histologic classification of colorectal polyps with augmented visualization	33
Abdul A.; Von Der Weth C.; Kankanhalli M.; Lim B.Y.	COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations	31

Wu Y.; Kuang K.; Zhang Y.; Liu X.; Sun C.; Xiao J.; Zhuang Y.; Si L.; Wu F.	De-biased court's view generation with causality	30

A.4: 25 papers in journals for further analysis and discussion

Author	Title	Number of citations
Lundberg S.M., Erion G., Chen H., DeGrave A., Prutkin J.M., Nair B., Katz R., Himmelfarb J., Bansal N., Lee S.-I.	From local explanations to global understanding with explainable AI for trees	1508
Holzinger A., Langs G., Denk H., Zatloukal K., Müller H.	Causability and explainability of artificial intelligence in medicine	518
Shin D.	The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI	209
Amann J., Blasimme A., Vayena E., Frey D., Madai V.I., the Precise4Q consortium	Explainability for artificial intelligence in healthcare: a multidisciplinary perspective	209
Lim B., Arık S.Ö., Loeff N., Pfister T.	Temporal Fusion Transformers for interpretable multi-horizon time series forecasting	167
Holzinger A., Malle B., Saranti A., Pfeifer B.	Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI	158
Reyes M., Meier R., Pereira S., Silva C.A., Dahlweid F.-M., Tengg-Kobligk H.V., Summers R.M., Wiest R.	On the interpretability of artificial intelligence in radiology: Challenges and opportunities	149

Zerilli J., Knott A., Maclaurin J., Gavaghan C.	Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?	106
Angelov P., Soares E.	Towards explainable deep neural networks (xDNN)	101
Gao K., Su J., Jiang Z., Zeng L.-L., Feng Z., Shen H., Rong P., Xu X., Qin J., Yang Y., Wang W., Hu D.	Dual-branch combination network (DCN): Towards accurate diagnosis and lesion segmentation of COVID-19 using CT images	96
Arnold M., Piorkowski D., Reimer D., Richards J., Tsay J., Varshney K.R., Bellamy R.K.E., Hind M., Houde S., Mehta S., Mojsilovic A., Nair R., Ramamurthy K.N., Olteanu A.	FactSheets: Increasing trust in AI services through supplier's declarations of conformity	91
De Bruyn A., Viswanathan V., Beh Y.S., Brock J.K.-U., von Wangenheim F.	Artificial Intelligence and Marketing: Pitfalls and Opportunities	82
Garcez A.D., Gori M., Lamb L.C., Serafini L., Spranger M., Tran S.N.	Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning	73
Guo W.	Explainable Artificial Intelligence for 6G: Improving Trust between Human and Machine	66
Yeom S.-K., Seegerer P., Lapuschkin S., Binder A., Wiedemann S., Müller K.-R., Samek W.	Pruning by explaining: A novel criterion for deep neural network pruning	66

Hong S.R., Hullman J., Bertini E.	Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs	60
Sachan S., Yang J.-B., Xu D.-L., Benavides D.E., Li Y.	An explainable AI decision-support-system to automate loan underwriting	49
Giudici P., Raffinetti E.	Shapley-Lorenz eXplainable Artificial Intelligence	45
Kim B., Park J., Suh J.	Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information	45
Vassiliades A., Bassiliades N., Patkos T.	Argumentation and explainable artificial intelligence: A survey	43
Heinrichs B., Eickhoff S.B.	Your evidence? Machine learning algorithms for medical diagnosis and prediction	43
Ploug T., Holm S.	The four dimensions of contestable AI diagnostics- A patient-centric approach to explainable AI	40
Buiten M.C.	Towards intelligent regulation of artificial intelligence	40
Holzinger A.	Explainable AI and Multi-Modal Causability in Medicine	36

Aizenberg E., van den Hoven J.	Designing for human rights in AI	35

12 papers in conferences for further analysis and discussion

Authors	Title	Number of citations
Liao Q.V.; Gruen D.; Miller S.	Questioning the AI: Informing Design Practices for Explainable AI User Experiences	245
Panigutti C.; Perotti A.; Pedreschi D.	Doctor XAI An ontology-based approach to black-box sequential data classification explanations	89
Casalicchio G.; Molnar C.; Bischl B.	Visualizing the feature importance for black box models	82
Clark P.; Tafjord O.; Richardson K.	Transformers as soft reasoners over language	74
Ribera M.; Lapedriza A.	Can we do better explanations? A proposal of user-centered explainable AI	63
Weitz K.; Schiller D.; Schlagowski R.; Huber T.; André E.	I"do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design	60
Feng S.; Boyd-Graber J.	What can Ai do for me? Evaluating machine learning interpretations in cooperative play	55
Frye C.; Rowat C.; Feige I.	Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability	53

Wang J.; Zhang Y.; Tang K.; Wu J.; Xiong Z.	AlphaStock: A buying-winners-and-selling-losers investment strategy using interpretable deep reinforcement attention networks	50
Ehsan U.; Riedl M.O.	Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach	48
Shankaranarayana S.M.; Runje D.	ALIME: Autoencoder based approach for local interpretability	36
Abdul A.; Von Der Weth C.; Kankanhalli M.; Lim B.Y.	COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations	31

A.5 The principles of human-centered explainable system designing by Mueller et al. (2021)

No.	Principles	Explanation
1	The property of being an explanation is not a property of statements, visualizations, or examples.	Explaining is a process by which an explainee and explainer achieve common ground. XAI Designers should recognize that an explanation is not merely an artifact delivered from an algorithm, but must be understood by a user to be effective.
2	Work matters.	It is impractical to develop a useful and usable explanation system outside of a work context. Explanation relates the tool to the user knowledge in the context of the goals and tasks, and whether a particular algorithm or visualization will support that work cannot be known in the abstract. Human-centered design principles suggest involving the user early and throughout the system development process
3	The importance of active self-explanation. The	The "spoon feeding" paradigm is oblivious to the fact that irrespective of whatever material people receive by way of explanation, they still engage in a motivated attempt to make sense of the AI and the explanatory material. Developers should recognize this and focus explanatory systems on information that empowers users to self-explain, rather than simply delivering an output of an algorithms that is intended to serve as explanatory.

4	Build explanatory systems, not explanations.	Rarely does the initial explanation coming directly from the AI algorithms provide a useful explanation, let alone an ideal one. It must be accompanied by other things (instructions, tutorial activities, comparisons, exploratory interfaces, user models, etc.) to succeed.
5	Combined methods are necessary.	Much work on XAI involves testing a particular concept or algorithm in isolation. But when designing an explanatory system, multiple kinds of information can complement one another. For example, both global and local explanations may be justifiable and reinforce one another. Showing examples that establish a pattern may play a different role than contrasting examples which establish a critical causal relationship. Using examples of related heatmaps may be more powerful than either examples or heatmaps in isolation. And it is crucial to keep in mind that actual work contexts involve multiple systems, so the user is continuously challenged with their own "system integration" task.
6	An explanation can have many different consequences.	Often, developers create and test explanations to determine whether they work. However, different explanations can have very different effects. This includes differing effects on qualitative assessments (satisfaction, trust), versus knowledge measures and performance measures. The explanations should be tuned to the goal, keeping in mind the fact that people may be led to trust and rely upon an AI system simply being given more and more information about it, whether or not that information leads to better or deeper understanding.
7	Measurement Matters.	Because explanations can have their impact on a number of ways, and so can be assessed along many dimensions (goodness, satisfaction and trust, knowledge/understanding, and performance). Designers should identify what consequences the explanation should have in order to develop an appropriate measurement and assessment approach.
8	Knowledge and understanding are central.	Much of the research on XAI focuses on algorithmic visualizations, which distracts from the fact that the focus of explanation is on developing a better understanding of the system. Understanding leads to appropriate trust and appropriate reliance, and therefore overall better performance with the system.

9	Context matters: Users, timing, goals.	An explanation is not a beacon revealing the truth. The best explanation depends on context: who the user is, what their goal is, when they need an explanation, and how its effectiveness is measured. Developers should consider use cases, user models, timeliness, and attention and distraction limitations for their explanations.
10	The power of differences and contrast.	A central lesson of XAI is the utility of contrast, comparison, and counterfactuals in understanding the boundary conditions of a system. A useful exercise is to first develop learning objectives for an explanatory system, and identify the contrasts necessary to support those objectives.
11	Explanation is not just about transparency.	If something is transparent, you cannot see it. This word is widely misused. What is needed are systems whose workings are apparent, that is, readily understood and not hidden (the "black box metaphor"). Especially in the context of fairness in algorithmic decision making, many have advocated “transparency” as an approach to explanation. This can never be enough, because a user may still not understand how a system works even if its algorithms are somehow apparent---observable and visible. Other methods (contrast, global explanation and local justification, examples, explorable interfaces that permit hypotheticals, etc.) will be necessary in most situations to harness apparency and develop understanding. In "real world" work contexts, people always feel some mixture of justified trust, unjustified trust and justified and unjustified mistrust. These attitudes are in constant flux and rarely develop in a smooth progression toward some ideal and stable point. Trust can come and go in a flash. When the AI fails in a way that a human would never fail, reliance can collapse.
12	The need for explanation is “triggered”.	Too often, XAI systems deliver an explanation regardless of whether one is needed. However, explanations are not always necessary. In normal human reasoning, explanation is triggered by states such as surprise and violations of expectation. Advances in XAI will come when systems begin to understand situations that are likely to engender surprise and violate user expectations.

13	Explanation is knowledge transformation and sensemaking.	The achievement of an understanding is not just the learning or incorporation of information; it involves changing previous beliefs and preconceptions. The acquisition of knowledge involves both assimilation and accommodation, to use Piagetian terminology. The power of explanation is that it can activate fast "System 2" learning modes that quickly reconfigure knowledge with minimal feedback, bypassing the slower "System 1" trial-and-error feedback-based learning often used to understand a system. XAI systems should harness this by attempting to identify the user's current understanding (so that it can better predict how to transform this knowledge), and support the information that will help make these transformations.
14	Explanation is never a "one-off."	Especially for AI systems that learn or are applied in dynamic contexts, users often need repeated explanations and re-explanations. How has the algorithm changed? Are these new data valid? XAI systems might benefit from considering the long-term interaction with users, even in simple ways like recognizing that once learned, an explanation may not need to be given again unless something has changed.

A.6 Questions in interviews

Stage	Principles in Appendix 1	Questions
Conceptualization	1, 8, 9 and 11	1. Before using an AI system, how well do you understand its working principle? 2. When you use an AI system, how do you understand the operating principles of the system?

Construction	1, 2, 3, 4, 5, 9, 10, 11, 13, and 14	3. After AI provides a decision, how do you judge the correctness of the decision made by AI? 4. Are you willing to participate in the development of AI? How do you think this participation can help you understand the AI system you are using? 5. The basic principle of artificial intelligence is to train a mathematical model with historical data to obtain results. This process includes three elements: historical data, the model, and results. How do you think one should understand these three elements? 6. If the AI system you are using is meant to help you understand the reasons behind decisions, what form do you think is best? Images, conversations, text, videos, or others? Why?
Measurement	2, 4, 6, 7, 9, 10, 11, and 12	7. Does the AI system you are using increase or decrease your confidence in the system's decision-making process when it helps you understand decisions? Why? 8. When evaluating the AI system you are using, what do you think your role is? 9. When evaluating the AI system you are using, what factors can make you trust the results?