Visual Representation of Explainable Artificial Intelligence Methods: Design and Empirical Studies

Zur Erlangung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften

Dr. rer. pol.

von der KIT-Fakultät für Wirtschaftswissenschaften

des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Miguel Angel Meza Martínez, M.Sc.

Tag der mündlichen Prüfung:	17. Mai 2023
Referent:	Prof. Dr. Alexander Mädche
Korreferent:	Prof. Dr. Ali Sunyaev
Karlsruhe	November 2023

Abstract

Explainability is increasingly considered a critical component of artificial intelligence (AI) systems, especially in high-stake domains where AI systems' decisions can significantly impact individuals. As a result, there has been a surge of interest in explainable artificial intelligence (XAI) to increase the transparency of AI systems by explaining their decisions to end-users. In particular, extensive research has focused on developing "local model-agnostic" explainable methods that generate explanations of individual predictions for any predictive model. While these explanations can support end-users in the use of AI systems through increased transparency, three significant challenges have hindered their design, implementation, and large-scale adoption in real applications.

First, there is a lack of understanding of how end-users evaluate explanations. There are many critiques that explanations are based on researchers' intuition instead of end-users' needs. Furthermore, there is insufficient evidence on whether end-users understand these explanations or trust XAI systems. Second, it is unclear which effect explanations have on trust when they disclose different biases on AI systems' decisions. Prior research investigating biased decisions has found conflicting evidence on explanations' effects. Explanations can either increase trust through perceived transparency or decrease trust as end-users perceive the system as biased. Moreover, it is unclear how contingency factors influence these opposing effects. Third, most XAI methods deliver static explanations that offer end-users limited information, resulting in an insufficient understanding of how AI systems make decisions and, in turn, lower trust. Furthermore, research has found that end-users perceive static explanations as not transparent enough, as these do not allow them to investigate the factors that influence a given decision.

This dissertation addresses these challenges across three studies by focusing on the overarching research question of how to design visual representations of local model-agnostic XAI methods to increase end-users' understanding and trust. The first challenge is addressed through an iterative design process that refines the representations of explanations from four well-established model-agnostic XAI methods and a subsequent evaluation with end-users using eye-tracking technology and interviews. Afterward, a research study that takes a psychological contract violation (PCV) theory and social identity theory perspective to investigate the contingency factors of the opposing effects of explanations on end-users' trust addresses the second challenge. Specifically, this study investigates how end-users evaluate explanations of a gender-biased AI system while controlling for their awareness of gender discrimination in society. Finally, the third challenge is addressed through a design science research project to design an interactive XAI system for end-users to increase their understanding and trust.

This dissertation makes several contributions to the ongoing research on improving the transparency of AI systems by explicitly emphasizing the end-user perspective on XAI. First, it contributes to practice by providing insights that help to improve the design of explanations of AI systems' decisions.

Additionally, this dissertation provides significant theoretical contributions by contextualizing the PCV theory to gender-biased XAI systems and the contingency factors that determine whether end-users experience a PCV. Moreover, it provides insights into how end-users cognitively evaluate explanations and extends the current understanding of the impact of explanations on trust. Finally, this dissertation contributes to the design knowledge of XAI systems by proposing guidelines for designing interactive XAI systems that give end-users more control over the information they receive to help them better understand how AI systems make decisions.

Dedication

With immense gratitude and love, I dedicate this dissertation to the most important people in my life, whose love, support, and encouragement have been the driving force behind my academic and professional journey.

To my loving wife, *Judith Elena Heredia Lopez*, your faith in me and your words of motivation have been my constant source of strength. Your unwavering support, understanding, and sacrifices have sustained me through the long hours and countless challenges of this academic pursuit. There are not enough words to express how much I value everything you have done for me, and I will be eternally grateful to you for inspiring me and pushing me to follow this dream. This dissertation is a testament to the love and partnership that we share, and it is dedicated to you with all my heart.

To my parents, *María del Carmen Martínez Estrella* and *Miguel Angel Meza Villalvazo*, your relentless love and countless sacrifices have shaped me into the person that I am today and have shaped my path in ways I can never repay. This dissertation stands as a tribute to your enduring encouragement that has fueled my determination to reach this milestone, and I offer it as a token of my deep appreciation for all you have done.

To my dear sister, *Carmen Alejandra Meza Martinez*, your steadfast believe in me has inspired me to keep pushing my boundaries. I truly appreciate all the support, encouragement, and help you have given me during this journey. This dedication is a tribute to the special sibling bond we share.

Lastly, but by no means least, I want to express my gratitude to my friends and colleagues, who have stood by me offering a welcome respite from the demands of academia. Your friendship and belief in me have been a great support. Whether reading my drafts, participating in my experiment pretests, providing a listening ear when the pressure felt overwhelming, or simply being a source of laughter and relief during difficult times, your companionship and encouragement made a significant difference in my academic and professional life. I'm deeply grateful for the support you have provided me.

In sincere dedication, I offer this dissertation to my wife, parents, sister, and friends with profound love and gratitude for their pivotal roles in this journey.

Acknowledgments

Completing this dissertation has been a challenging yet rewarding journey, and I owe my heartfelt gratitude to the many individuals who have supported and guided me along the way.

First and foremost, I want to express my deepest appreciation to my academic advisors, *Prof. Dr. Alexander Mädche* and *JProf. Dr. Mario Nadj*, for their guidance and comprehensive support. I thank Alexander Mädche – my *Doktorvater* – for allowing me to continue my academic development in his chair. *Alexander Mädche* allowed me to explore the topics I felt passionate about and always supported me in the search for research collaborations that could strengthen my work. He was also very supportive of my efforts to balance my research at the institute and my activities at SAP. Further, I thank Mario Nadj for his encouragement and invaluable support in my research work. I feel fortunate to have had him as my competent advisor for this dissertation and my research. I truly value all the uncountable sessions in which he helped me review my experimental design and statistical analyses. *Mario Nadj* was always a source of inspiration for his dedication to detail and I appreciate everything that I've learned from him.

I also thank *SAP* for allowing me to join the company as a Ph.D. student and funding this dissertation. In particular, my manager *Thomas Volmering* for allowing me to join his team where I had the opportunity to continue my professional development learning and implementing artificial intelligence solutions. *Thomas Volmering* was very supportive of my research collaboration with the *Karlsruhe Institute of Technology (KIT)* and my pursuit of my doctoral degree. Also, my deepest gratitude to my colleagues *Prashant Gautam*, *Dennis Scherle*, and *Hannah Sperling* for their interest in my research and their invaluable support and constructive feedback on my research projects.

Moreover, I thank my colleagues and students with whom I have collaborated in the past five years. I feel very fortunate to have worked with you on many interesting research projects. My thanks to my colleague *Prof. Dr. Ekaterina Jussupow* for the incredible experience of collaborating closely on research projects critical for this dissertation. It was a great pleasure working alongside Ekaterina Jussupow complementing our technical and theoretical backgrounds to investigate the challenges that arise due to biases in artificial intelligence. Further, a special thanks to all my co-authors: *Dr. Peyman Toreini*, *M.Sc. Moritz Langner* and *Prof. Dr. Armin Heinzl*. It was a great pleasure working with you, sharing ideas, discussing research, and writing papers together.

Contents

List of Tables

List of Figures

List of Abbreviations

AdViCE ……. Aggregated Visual Counterfactual Explanations
AGFI ……. Adjusted Goodness of Fit
AI ……..……. Artificial Intelligence
ALE ……..……. Accumulated Local Effects
ANCOVA ……. Analysis of Covariance
ANN …… Artificial Neural Networks
ANOVA ……... Analysis of Variance
AOI ……..……. Area of Interest
AUC ……...……. Area Under the Curve
AVE …….. Average Variance Extracted
CFA ……..……. Confirmatory Factor Analysis
CFI ……. Comparative Fit Index
CI ……..……. Confidence Interval
CR ……..……. Composite Reliability
DARPA ……. Defense Advanced Research Projects Agency
DF ……..……. Design Feature
DICE ……..……. Diverse Counterfactual Explanations
DP ……..……. Design Principle
DPD ……..……. Partial Dependence Plots
DSR ……..……. Design Science Research
GAM …..……. Generalized Additive Models
GFI ……..……. Goodness of Fit
GLM ……. General Linear Models
HCI ……. Human-computer Interaction
ICE ……. Individual Conditional Expectation
IDC …..……. International Data Corporation
IT ……..……..……. Information Technology
LIME ……..……. Local Interpretable Model-agnostic Explanations
LLCI ……. Lower-Level Confidence Interval
MANOVA ……... Multivariate Analysis of Variance
ML ……...……. Machine Learning
MR ……...……. Meta-requirement
PCV ……..……. Psychological Contract Violation
QII ……. Quantitative Input Influence
RA ……. Recommendation Agents

RISE …….……. Randomized Input Sampling for Explanation
RQ ……...……. Research Question
SC …….……. Stigma Consciousness
SD ……..……. Standard Deviation
SHAP ……..……. Shapley Additive Explanations
SIDU ……. Similarity Difference and Uniqueness
SMOTE ……..……. Synthetic Minority Oversampling Technique
SPSS ……. Statistical Package for Social Sciences
ULCI ……. Upper-Level Confidence Interval
XAI …….……. Explainable Artificial Intelligence

1 Introduction[1]

1.1 Motivation

Due to their capability of decoding complex data relationships to model behavior and to automatically learn and improve from experience without the need for explicit programming (Dudley & Kristensson, 2018), Artificial Intelligence (AI) systems based on Machine Learning (ML) algorithms have become ubiquitous and have a significant impact across all industries. As a result, AI systems are transforming every aspect of our lives by integrating and analyzing vast amounts of information, supporting human decision-making, and sometimes even taking over the decision-making process. The scale of AI systems' economic and societal potential is reflected by their successful implementation across an extensive spectrum of applications in fields such as e-commerce, financial services, manufacturing, entertainment, and customer service. To put this economic impact into a global perspective, the International Data Corporation (IDC) forecasts that revenues for the AI market worldwide, including software, hardware, and services, will grow 19.6% year over year in 2022 to around $432 billion and are expected to break the $500 billion mark in 2023 (IDC, 2022). Furthermore, Gartner has included AI as an essential technology among the Gartner top 12 strategic technology trends for 2022 (Gartner, 2022).

The large availability of information in our digital era, significant technological developments in hardware, research on new optimized algorithms, and the extensive availability of high-quality open-source libraries have led to AI systems achieving very high performance in many tasks previously considered as computationally unattainable (Došilović et al., 2018; Lecun et al., 2015). For example, AI systems have already outperformed humans in complex tasks such as speech recognition, language translation, strategy games, and visual recognition (Mnih et al., 2015; Russakovsky et al., 2015; Silver et al., 2016). As a result, with the expectation of increasing decision-making quality and efficiency, AI systems are being increasingly developed and deployed in high-stake domains (Binns et al., 2018; Stowell et al., 2018). For instance, in the healthcare sector, Esteva et al. (2017) developed an AI system for skin cancer screening that can differentiate between images of benign and malignant skin lesions. Furthermore, Gulshan et al. (2016) developed an AI system that analyzes images to detect diabetic retinopathy, a diabetes complication that affects the eyes and can cause blindness (Mayo Clinic, 2021). Likewise, AI systems have been used in recruitment to find the best candidates among large volumes of curricula vitae or to analyze video interviews to assess the person-organization and person-job fit (Albert, 2019).

[1] This chapter is based on the following published papers and papers under review: (Jussupow et al., 2023; Jussupow, Meza Martínez, et al., 2021; Meza Martínez et al., 2023; Meza Martínez & Maedche, 2023)

Regardless of the success that AI systems have achieved across multiple fields and applications over the years, it has repeatedly been shown that AI systems are prone to deliver biased decisions that can have considerable consequences for individuals directly affected by these decisions. For instance, while developing an AI system for diagnosing malignant skin tumors from images, an AI system inadvertently learned that rulers appearing in images were an indication of malign tumors (Narla et al., 2018). Moreover, it was shown that an AI system for risk assessment of recommitting crimes disproportionately misclassified black individuals with a higher risk than white individuals (Angwin et al., 2016). Furthermore, an AI system for the automatic review of job applicants' resumes in hiring processes learned to penalize resumes that included words associated with women (Dastin, 2018). These cases exemplify the potential negative consequences that biased decisions of AI systems can have on individuals. While humans can also fail in their judgment and provide biased decisions, it is at least possible to ask them for a rationale for their decision and hold them accountable (Binns et al., 2018).

Traditionally, in software used in high-stake domains, handcrafted rules that formalize the knowledge of domain experts are explicitly programmed into the decision logic of the software (Janiesch et al., 2021). Therefore, it is possible to examine which rules were used to make a particular decision. In contrast to traditional rule-based systems, AI systems building on ML can be extraordinarily complex and difficult to understand or even audit (Weld & Bansal, 2019). AI systems often incorporate hundreds or thousands of factors in their decision-making. Furthermore, AI systems are typically designed to process large amounts of training data to perform a complex optimization process for a specific performance measure. As a result, AI systems are considered "black boxes" where only the output of the underlying ML model is available to users. This lack of transparency leaves the inner working mechanisms of the AI system unclear to users. This problem has been reinforced by the popularity of deep learning models, which are hard to understand, even by experts (Dodge et al., 2019).

The effectiveness of AI systems is limited by their inability to explain their decisions to human users (D. Wang et al., 2019). In particular, for AI systems deployed in high-stake applications, it is critical that end-users understand the logic behind the decisions these systems make. Transparency on how decisions are made is essential for end-users to trust these systems and accept the decisions they provide. As a result, new regulations to provide the "right to explanation" of all decisions made or supported by AI systems have been proposed and implemented around the globe (Cheng et al., 2019; Dodge et al., 2019). The European Union General Data Protection Regulation (GDPR) is one of the most prominent examples of such regulations. This regulation requires organizations utilizing AI systems for decision-making to provide affected individuals with "meaningful information about the logic involved" in the decision-making process (Goodman & Flaxman, 2017). However, such regulations still need to be put into practice. Overall, it remains difficult for non-experts to understand the logic behind AI systems and figure out how specific inputs lead to a particular output (Cheng et al., 2019).

In this light, there has been a recent surge of interest in explainable artificial intelligence (XAI) among scholars and practitioners seeking to increase the transparency of AI systems (Miller, 2019). The goal of XAI is to produce systems that can explain their decisions in non-technical terms while at the same time maintaining a high prediction accuracy (Diakopoulos et al., 2017; Turek, 2018). XAI also aims to support human understanding and trust in the use of AI systems (Turek, 2018). There are many potential benefits of providing explanations of AI systems' decisions. For instance, explanations can help develop an AI system by enabling developers to understand the system's underlying logic (Kaur et al., 2019). Additionally, explanations can help analyze the system's behavior to ensure that the system is not only optimized for performance but also compliant with ethical and legal standards (Lipton, 2018). Thus, developers can identify and fix potential problems before deploying the system. Once an AI system has been deployed, explanations increase its transparency by making the underlying factors that influence a given decision visible to end-users. This transparency enables end-users to review and audit the decisions provided by the system (Tintarev & Masthoff, 2007). Furthermore, multiple studies have shown that providing explanations can increase end-users' trust in an AI system (Buçinca et al., 2020; Yang et al., 2020) and the likelihood that they accept the decision it provides (Cramer et al., 2008; Lim et al., 2009; Yeomans et al., 2019).

Explainability is being considered to be more and more a critical component of AI systems that can help to monitor and prevent the undesired consequences of biased decisions. For instance, in the published document *Ethics Guidelines for Trustworthy AI*, the High-Level Expert Group on artificial intelligence (HLEG) set up by the European Commission lists seven essential requirements of trustworthy AI systems (AI HLEG, 2020). Explainability is one of the three elements encompassed by transparency, identified as one of the seven essential requirements. Likewise, the National Institute of Standards and Technology (NIST) identifies explainability as one of the desirable characteristics of trustworthy AI systems (Phillips et al., 2021). Specifically, the NIST proposes four principles for explainable AI systems that can be summarized as a system that provides meaningful and accurate explanations and only operates under the conditions for which it was designed (Phillips et al., 2021).

Many companies have also identified the need for explainability in AI systems as a critical requirement. According to an IBM Institute for Business Value survey, 68 percent of business leaders expect that customers will demand more explainability from AI systems in the upcoming years (Mojsilovic, 2019). As a result, explainability is gaining more and more focus in commercialized AI systems as companies aim to adopt it to help them manage the risks of AI systems and improve their customers' trust in them (Chromik, 2021). For instance, the business AI platform *Watson AI OpenScale* from IBM, with features such as trust and transparency, explains outcomes to help mitigate bias (Smith, 2018). Similarly, the automatic ML platform *H2O Driverless AI* from H2O.ai, provides explainability as one of its main features (H2O.ai, 2022). Furthermore, several leading technology companies have developed open-source libraries and toolkits as a means to help gain a comprehensive understanding of AI systems and

the decisions they deliver. Examples of such efforts include AIX360,[2] Contextual AI,[3] InterpretML,[4] and Vizier.[5]

To tackle the lack of transparency in AI systems, extensive research has been conducted in several research communities. These efforts include academic institutions as well as government research projects. A prominent example is the XAI program for funding academic and military research created by the Defense Advanced Research Projects Agency (DARPA) (Gunning & Aha, 2019). Furthermore, the increasing attention on XAI has been reflected in the presence of this topic in major scientific conferences and journals (Adadi & Berrada, 2018). Some conferences, such as the ACM Conference on Fairness, Accountability, and Transparency (FAccT), focus exclusively on fairness and transparency topics in AI systems. Additionally, multiple workshops focused on explainability have been conducted at many conferences. Examples include the Fairness, Accountability, and Transparency in Machine Learning workshop (FAT/ML, 2022) and the Workshop on Human Interpretability in Machine Learning held at the International Conference on Machine Learning (WHI, 2020).

In particular, extensive research has focused on developing so-called "model-agnostic" explainable methods (Dodge et al., 2019). These methods are applied after the predictive model has been trained and can explain any predictive model (Adadi & Berrada, 2018; Molnar et al., 2020). Such methods are expected to gain further popularity due to the higher applicability and scalability offered by their decoupling from the prediction model (Ribeiro et al., 2016b). Some of these model-agnostic methods provide local explanations, which clarify how individual predictions are made. In contrast, others offer global explanations, which describe the entire model behavior across all instances for a given dataset (Molnar, 2020). Local explanations from model-agnostic methods are becoming increasingly important today, especially as AI systems are used in high-stake domains where organizations and individuals need to be able to understand why these systems made a particular decision.

Many innovative algorithms, visualizations, and interfaces have been developed due to extensive research in the XAI field. These achievements have been presented by several studies that have surveyed the literature (see, e.g., Adadi & Berrada, 2018; Carvalho et al., 2019; Guidotti, Monreale, Ruggieri, Turini, et al., 2018; Miller, 2019; Molnar, 2020; Tjoa & Guan, 2019). Nevertheless, despite the efforts and achievements in XAI, many of the proposed model-agnostic methods have not been evaluated with end-users (D. Wang et al., 2019). Instead, researchers often define custom measures to assess the quality of the local explanations generated by these methods (Doshi-Velez & Kim, 2017). Therefore, there have been many critiques that the developed model-agnostic XAI methods are based on researchers' intuition instead of a deep understanding of end-users' needs (Miller, 2019; D. Wang et al., 2019). So far, there

[2] https://github.com/Trusted-AI/AIX360
[3] https://github.com/SAP/contextual-ai
[4] https://github.com/interpretml/interpret
[5] https://github.com/google/vizier

is insufficient evidence on whether end-users understand and trust AI systems that provide local explanations from model-agnostic XAI methods (Abdul et al., 2018; Cheng et al., 2019). Moreover, it is unclear how end-users perceive these explanations and how they utilize and assess them while interacting with AI systems.

Therefore, this dissertation aims to investigate how to design visual representations of local model-agnostic XAI methods to increase end-users' understanding and trust. To achieve this goal, this dissertation, in the first step, investigates how end-users evaluate these explanation representations and whether these explanations can help them understand how AI systems make decisions. Moreover, this dissertation also examines whether existing explanation representations alone can increase end-users' trust in AI systems or if further enhancements are necessary. On this basis, this dissertation contributes to both (1) academia by extending the understanding of the effects that local explanations from model-agnostic methods have on end-users' trust and (2) practice by providing empirical evidence on how end-users perceive the visual representation of explanations and by providing guidelines on how to design XAI systems.

1.2 Derivation of Research Gaps and Associated Research Questions

In the following sections, four sub-research (Sub-RQ) questions are derived and described from existing scientific literature before this dissertation's overall research question (RQ) is presented.

1.2.1 End-users' Evaluation of Different Local Model-agnostic Explanation Representations

The first two sub-research questions focus on how to design comparable local model-agnostic explanation representations and how end-users evaluate these designs. Recently there has been a rise in interest in XAI as a means to increase the transparency of AI systems by explaining their decisions in non-technical terms to end-users (Diakopoulos et al., 2017; Miller, 2019; Turek, 2018). In addition, by providing explanations of AI systems' decisions, researchers aim to support end-users' interaction with AI systems by increasing their understanding and trust in them (Turek, 2018). Besides, explainability is now considered a critical component of AI systems to help monitor their performance and prevent the undesired negative consequences that their wrong decisions can cause (AI HLEG, 2020; Phillips et al., 2021). It is also expected that in the upcoming years, there will be a higher demand from customers for more explainability in AI systems (Mojsilovic, 2019). As a result, extensive research has been conducted in the XAI field to develop several innovative model-agnostic explainability methods that provide local explanations, which are used to clarify the logic behind individual decisions made by AI systems (Dodge et al., 2019; Molnar, 2020)

Nevertheless, despite the advances in XAI, there is currently insufficient evidence on whether end-users understand and trust the local explanations provided by the several model-agnostic methods developed so far (Abdul et al., 2018; Cheng et al., 2019). For many of these novel developed explainability methods, researchers often perform evaluations without end-user involvement by proposing a custom interpretability metric to evaluate the quality of explanations (Doshi-Velez & Kim, 2017). For instance, Mothilal et al. (2020) decided to use the same distance function of their predictive model as a proximity measure to compare their explanations with other approaches. Therefore, there have been many critiques that the developed model-agnostic XAI methods are based on researchers' intuition instead of a deep understanding of end-users' needs (Miller, 2019; D. Wang et al., 2019).

Among others, Miller (2019) argues that experts who developed these model-agnostic methods may lack the judgment necessary to assess the usefulness of the generated explanation representations for end-users (see also Adadi & Berrada, 2018; Mittelstadt et al., 2019; Ribera & Lapedriza, 2019). Furthermore, although research in the literature agrees that explanations can help understand AI systems' decisions (Kulesza et al., 2013), there is a lack of research on how end-users evaluate the representation of local explanations from different model-agnostic XAI methods. Several studies have focused on evaluating the explanations of model-agnostic methods against non-model agnostic methods or no explanations at all (B. Kim et al., 2016; Ribeiro et al., 2016b). Furthermore, many studies have evaluated model-agnostic explanations with experts or practitioners instead of end-users (Jesus et al., 2021; Kaur et al., 2019). Moreover, from the limited studies that have evaluated explanations from multiple local explanations from different XAI methods with end-users, authors have used purely textual explanations to control the representations differences between the evaluated methods (Binns et al., 2018; Dodge et al., 2019).

Therefore, further research is needed to evaluate and compare holistically the representations of existing local model-agnostic explanations generated by different XAI methods to reach a consensus on which are more appropriate for end-users (D. Wang et al., 2019). To address these research challenges, the following two sub-research questions are proposed:

Sub-RQ1*: How to design comparable local model-agnostic explanation representations from different XAI methods following a user-centered approach?*

Sub-RQ2*: How do end-users perceive, evaluate, and visually attend to the designed local model-agnostic explanation representations from different XAI methods?*

1.2.2 The Effects of Explanations of Biased AI Systems' Decisions on End-users' Trust

The third sub-research question addresses end-users' evaluation of explanations for biased AI systems and the impact these have on their trust. Due to the very high performance that AI systems have achieved

in many complex tasks in the last decades (Došilović et al., 2018; Lecun et al., 2015), they are being increasingly deployed in high-stake applications such as healthcare, criminal justice, personal recruitment, and finance (Binns et al., 2018; Hartmann & Wenzelburger, 2021; Köchling & Wehner, 2020; Stowell et al., 2018). Nevertheless, many cases have been documented where AI systems have delivered biased decisions that have considerable consequences for individuals affected by those decisions (e.g., Angwin et al., 2016; Dastin, 2018; Narla et al., 2018). In particular, it has repeatedly been shown that AI systems are prone to amplifying gender biases by favoring men over women (Sharma et al., 2020). In contrast to traditional systems used in high-stake domains where explicit rules are coded for decision-making, AI systems are often too complex, which makes it challenging to understand the reasons behind the decisions they provide.

Explanations from XAI methods have been proposed as a helpful means to increase the transparency of AI systems in high-stake domains and enable end-users to understand the reasons behind these systems' decisions (Binns et al., 2018; D. Wang et al., 2019). Furthermore, multiple studies have revealed that providing explanations can increase end-users' trust in AI systems (W. Wang & Benbasat, 2007; Yang et al., 2020) and the likelihood that they will agree with the provided decision recommendations (Yeomans et al., 2019). Therefore, explanations can increase end-users' overall decision-making performance when these systems provide accurate decision recommendations.

Nevertheless, research investigating AI systems that provide biased decisions has found conflicting evidence on the effects that these explanations have. On the one hand, several studies have found that explanations can help end-users detect biases embedded in AI systems (Dodge et al., 2019; Law et al., 2020). For instance, Dodge et al. (2019) revealed that certain types of explanations were more effective than others in exposing case-specific bias issues. Therefore, by exposing inherent biases, explanations can reduce end-users' trust in AI systems.

On the other hand, studies have found that end-users are not always accurate in evaluating explanations of biased AI systems. For instance, Bussone et al. (2015) revealed that providing explanations can cause end-user overreliance on the context of clinical decision support systems. Similarly, Poursabzi-Sangdeh et al. (2021) demonstrated that when explanations were provided, individuals were less likely to detect biases in AI systems. Likewise, Law et al. (2020) demonstrated that individuals were less likely to consider attributes not displayed in explanations despite carefully auditing a biased AI system. Moreover, Lakkaraju and Bastani (2020) showed that individuals react differently to biased AI systems depending on whether or not explanations reveal discriminating attributes (e.g., race or sex). Other studies suggest that explanations of biased decisions can compensate for the adverse effects of biases on end-users' trust (Erlei et al., 2020; W. Wang et al., 2018; W. Wang & Wang, 2019).

Prior research on explanations of AI systems has mainly focused on the knowledge-based impact of explanations and neglected how individuals' emotional responses influence how they evaluate

explanations for systems' decisions (Kordzadeh & Ghasemaghaei, 2021; Starke et al., 2021). Nonetheless, prior research on social identity theory (Tajfel, 1982) suggests that individuals differ in their sensitivity to biases depending on their own past experiences and their social identification with the stigmatized group, i.e., their stigma consciousness (Pethig & Kroenung, 2020; Pinel, 1999). Therefore, individuals with a stronger identification with the stigmatized group are more sensitive toward these biases and respond more negatively to them. Hence, in the context of biased AI recommendations, individuals' social identity can strongly influence whether they perceive an AI system as biased.

The conflicting results in the literature show that there is very limited knowledge of how end-users cognitively evaluate explanations provided by XAI methods in the context of biased AI systems. Thus, it is unclear how these explanations can affect end-users' trust in AI systems. While explanations can influence some end-users to perceive an AI system as accurate despite its underlying biases, they can also cause other end-users to perceive it as less trustworthy. Additionally, individuals' stigma consciousness can play a role as a contingency factor in their evaluation of explanations that are provided by an AI system. Therefore, further research is needed to investigate how end-users evaluate explanations of biased systems and how their evaluations influence their trust in them. To address these research challenges, the following sub-research question is derived.

Sub-RQ3: *How do end-users differ in their evaluation of a biased XAI system's trustworthiness based on their level of stigma consciousness?*

1.2.3 Designing Interactive XAI Systems for End-users

The fourth sub-research question addresses how to design interactive XAI systems and how end-users evaluate these systems. As previously mentioned, many researchers and practitioners have resorted to the field of XAI to tackle the lack of transparency of AI systems. As a result of the extensive research in XAI in recent years, many innovative explainability methods have been developed to explain the logic behind AI systems' decisions to end-users (Carvalho et al., 2019). Nevertheless, many of these developed XAI methods can only deliver static explanations, which offer end-users a limited amount of information (Abdul et al., 2018; Ribera & Lapedriza, 2019). Providing these static explanations results in an insufficient end-users' understanding of how AI systems make decisions (Cheng et al., 2019; Liu et al., 2021). Moreover, research has found that end-users perceive static explanations as not transparent enough, as these do not allow them to investigate further the factors that influence a given decision (Sun & Sundar, 2022).

To address these challenges, researchers and practitioners have argued for enhancing XAI systems by allowing end-users to explore explanations, giving them more control over the information they receive (Krause et al., 2016; Miller, 2019). The expectation is that enabling interactive exploration of

explanations can give end-users a sense of agency and help promote trust in AI systems (Sun & Sundar, 2022). However, despite current efforts to explore how to design interactive XAI systems, most research studies have focused on developing these systems specifically for data scientists or domain experts (e.g., Hohman *et al.*, 2019; Spinner *et al.*, 2020). As a result, the explanations provided by these proposed systems are often too complex for end-users, making them very challenging to understand (Cheng et al., 2019; Miller, 2019). Therefore, it is necessary to design interactive XAI systems that provide explanations for end-users that support their understanding of AI systems' decisions. To address these research challenges, the following sub-research question is derived.

Sub-RQ4: *How to design interactive explainable artificial intelligence (XAI) systems to help end-users to better understand AI systems' decisions?*

1.2.4 Overall Research Question

Designing adequate explanations of AI systems' decisions is not a trivial task due to the complexity of the underlying ML algorithms that process large datasets (Haverinen, 2020). Furthermore, the requirements of what constitutes a good explanation might vary significantly depending on the target group. In particular, research has demonstrated that it is challenging to generate understandable explanations for end-users (Cheng et al., 2019).

Therefore, this dissertation takes note of the calls from researchers (e.g., Adadi & Berrada, 2018; Miller, 2019; Mittelstadt et al., 2019; Ribera & Lapedriza, 2019) to extend the current understanding of how end-users evaluate local explanations from model-agnostic methods and the effects that these explanations can have on their trust in AI systems. On this basis, this dissertation aims to design visual representations of local model-agnostic XAI methods that help end-users to better understand AI systems and trust these systems more. Thus, the following overall research question is formulated.

RQ: *How to design visual representations of local model-agnostic XAI methods to increase end-users' understanding and trust?*

1.3 Structure of the Dissertation

This dissertation's structure reflects the different studies conducted to address the overall research question and the four sub-research questions presented above.

Chapter 1 motivates the research performed in this dissertation by highlighting the importance of the necessity of explainability in AI systems. Additionally, it derives the sub-research questions and presents the structure of this dissertation. Subsequently, **Chapter 2** provides the conceptual foundations of XAI, the classification of XAI methods, and existing taxonomies to evaluate XAI explanations. Furthermore, it describes in detail the specific local model-agnostic XAI methods evaluated in this dissertation.

Chapter 3 presents **Study 1**, which investigates how to design comparable local model-agnostic explanation representations from different XAI methods following a user-centered approach. To achieve this goal, an iterative design process is conducted to refine the representations of local explanations from well-established model-agnostic XAI methods with end-users. Furthermore, the explanation representations of these explanations are evaluated with end-users using eye-tracking technology as well as self-reports and interviews.

Within **Chapter 4**, **Study 2** addresses the two conflicting effects that explanations of AI systems can have on end-users' trust. Explanations can increase trust through perceived transparency or decrease trust as end-users perceive the system as biased and unfair. A psychological contract violation theory and social identity theory perspective are incorporated into this study to investigate the contingency factors of these two opposing effects. Thus, two online experiments are conducted to evaluate a gender-biased AI system.

Chapter 5 presents **Study 3**, which investigates how to design interactive XAI systems to help end-users' to better understanding AI systems' decisions. To achieve this goal, this study conducts a design science research (DSR) project to design an interactive XAI system for end-users by proposing design principles and instantiating them in an initial prototype. Moreover, a qualitative evaluation of this prototype is conducted through interviews with end-users.

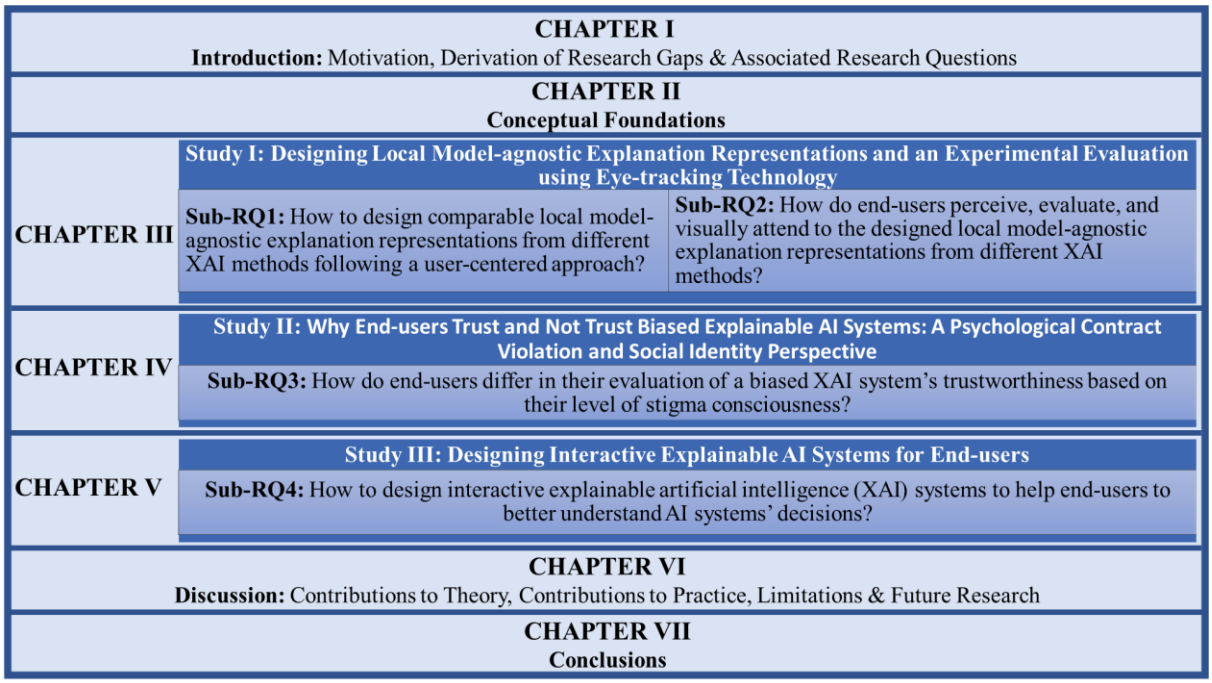


Figure 1: Structure of this dissertation.

Subsequently, **Chapter 6** discusses the contributions to theory and practice of this dissertation, its limitations, and potential future research directions. Finally, this dissertation concludes with closing remarks in **Chapter 7**. The structure of this dissertation is presented in Figure 1.

This dissertation is the result of extensive research conducted by the author in recent years. Parts of this dissertation's content have already been published in peer-reviewed journals or presented at peer-reviewed conferences. Therefore, this dissertation provides an overall framework for these studies and extends previously published content. In the following, an overview of the author's publications is presented that relates directly to this dissertation's contributions:

1. Jussupow, E., Meza Martínez, M. A., Maedche, A., & Heinzl, A. (2023). *Why Individuals Trust and Not Trust Biased Explainable AI Systems: A Psychological Contract Violation and Social Identity Perspective*. Working paper, to be submitted.
2. Meza Martínez, M. A., & Maedche, A. (2023). *Designing Interactive Explainable AI Systems for Lay Users*. Manuscript Accepted in the International Conference on Information Systems (ICIS 2023).
3. Meza Martínez, M. A., Nadj, M., Langner, M., Toreini, P., & Maedche, A. (2023). *Does This Explanation Help? Designing Local Model-Agnostic Explanation Representations and an Experimental Evaluation Using Eye-Tracking Technology*. ACM Transactions on Interactive Intelligent Systems (TiiS), Special Issue on Human-centered Explainable AI. Just Accepted. https://doi.org/10.1145/3607145
4. Jussupow, E., Meza Martínez, M. A., Maedche, A., & Heinzl, A. (2021). *Is This System Biased? – How Users React to Gender Bias in an Explainable AI System*. In Proceedings of the 42nd International Conference on Information Systems (ICIS 2021), Austin: AISel. https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/11

A complete list of the author's publications beyond the core scope of this dissertation can be found in the Appendix.

2 Conceptual Foundations[6]

This chapter presents the conceptual foundations of XAI that build the basis for this dissertation. First, the concept of XAI is presented. Then, a classification of XAI methods across multiple dimensions is presented to clarify the different approaches in the literature. Finally, the specific XAI methods relevant to this dissertation are presented in detail.

2.1 The Concept of XAI

The term XAI was first utilized in the AI field by Van Lent et al. (2004) to define an AI system's ability to provide an explanation for its behavior in the context of simulation games (Adadi & Berrada, 2018). Historically, there has been interest in the concept of explaining the decisions of intelligent systems for decades. This interest began in the 70s, as researchers investigated how expert systems could explain their reasoning process when providing recommendations (Moore & Swartout, 1988; Swartout, 1983). Similarly, some years later, researchers studied how some form of explanation capability could help Artificial Neural Networks (ANNs) reach their full potential (Andrews et al., 1995). Furthermore, researchers also investigated the influence of explanations on end-user trust and acceptance in recommender systems (Cramer et al., 2008; Herlocker et al., 2000). Nevertheless, the interest in explainability for AI systems diminished as the priority of AI research shifted and focused more on developing and implementing more accurate and efficient algorithms (Carvalho et al., 2019).

With the development of new ML algorithms, AI systems achieved high performance in many tasks previously considered unattainable (Došilović et al., 2018; Lecun et al., 2015). However, besides their impressive accuracy, many newly developed ML algorithms were also more complex and less transparent, as exemplified by deep learning models, which are hard to understand even by experts (Dodge et al., 2019). As these complex algorithms were progressively deployed in AI systems in high-stake domains, it became critical that these systems provide reasoning for their decisions, which could significantly impact individuals (D. Wang et al., 2019). Therefore, the topic of XAI has recently gained renewed attention from researchers and practitioners as a valuable means to improve the transparency of AI systems in critical decision-making processes (Miller, 2019).

The concept of XAI has evolved from research across multiple fields interested in increasing the transparency of AI systems. Therefore, there is no universally agreed definition of XAI. Instead, XAI and its terminology have been shaped across different research communities, creating a group of related concepts. Nonetheless, the XAI term among these communities tends to refer to the efforts to address the concerns about transparency, accountability, safety, and trust in AI systems (Adadi & Berrada, 2018;

[6] This chapter is based on the following published papers and papers under review: (Jussupow et al., 2023; Jussupow, Meza Martínez, et al., 2021; Meza Martínez et al., 2023; Meza Martínez & Maedche, 2023)

Miller, 2019). The main goal of XAI is to provide end-users with explanations of AI systems decisions in non-technical terms, enabling end-users to understand the factors that influenced that decision (Diakopoulos et al., 2017; Turek, 2018). Furthermore, XAI also aims to support human understanding and trust in the use of AI systems (Turek, 2018). Figure 2 illustrates the concept of XAI. It shows how end-users can interact with an XAI system through an explainable interface to obtain an explanation of how a given decision was made (DARPA, 2016). The representation of the system's explanation can vary according to the application's requirements and end-users' type (e.g., text, graphic, audio).

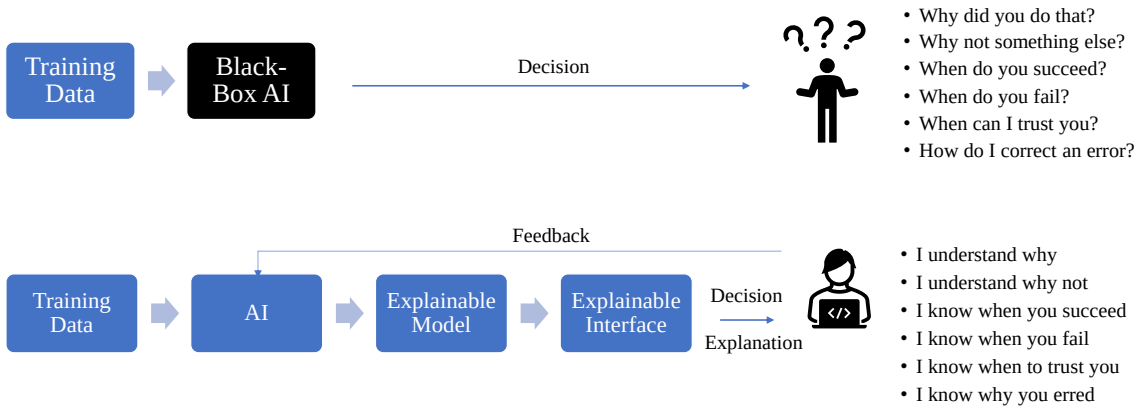


Figure 2: Concept of XAI (adapted from DARPA, 2016).

In the literature, there is ambiguity in terms of the terminology used to describe the capability of AI systems to provide decisions that are understandable to human users (Carvalho et al., 2019). For instance, the term XAI is closely related to the concept of interpretable ML. While interpretable ML is often used in the literature to refer to research on ML algorithms that are considered inherently interpretable, XAI is often used to refer to the generation of post-hoc explanations as a means of introspection for black-box models (Kaur et al., 2019; Rudin, 2019). Likewise, the terms explainability and interpretability express similar ideas. The term explainability has been defined as the ability of an ML model to explain the reason for its decisions to end-users accurately (Kulesza et al., 2015). Meanwhile, interpretability has been defined as the ability to present decisions to humans in understandable terms (Doshi-Velez & Kim, 2017). A system has also been defined as interpretable if its decisions can be easily understood by a human (Biran & Cotton, 2017). Many of these terms are used interchangeably in the literature.

In this dissertation, the term "explainability" describes the system's ability to explain the reasons for its decisions to end-users. Meanwhile, "interpretability" refers to the degree to which end-users can understand how a decision was made. Finally, the term "explanation" represents the details or reasons presented by the system to end-users in order to explain the factors behind an individual decision.

2.2 Classification of XAI Methods

As a result of the extensive XAI research performed in different research communities over the last few years, researchers and practitioners have developed many innovative algorithms, visualizations, interfaces, and toolkits. For instance, some methods extract easily interpretable rules from the predictive model and present them to users as an explanation of the model's decision (Deng & Brown, 2021; Ming et al., 2019; Thomas et al., 2021; Yuan et al., 2021). Alternatively, others highlight regions of an image to indicate which pixels were influential in the model's prediction (Landecker et al., 2013; Xu et al., 2015; B. Zhou et al., 2018). Several studies have surveyed the literature to provide a detailed overview of XAI by presenting the different approaches developed and providing classifications or conceptual frameworks. The following existing literature reviews provide a deeper understanding of XAI (see, e.g., Adadi & Berrada, 2018; Carvalho et al., 2019; Guidotti, Monreale, Ruggieri, Turini, et al., 2018; Miller, 2019; Molnar, 2020; Tjoa & Guan, 2019).

XAI methods developed by researchers and practitioners vary significantly in their approach in generating explanations. Therefore, several classification frameworks have been proposed in the literature across a set of dimensions. This dissertation relies on the classification of explainability methods across a set of dimensions proposed in the literature to introduce the conceptual foundations of XAI and define the research scope (Došilović et al., 2018; Molnar, 2020; Sokol & Flach, 2020). The four dimensions selected for this dissertation are (1) relationship to the predictive system, (2) explanation scope, (3) applicability, and (4) explanation family. The classification dimensions presented in this dissertation are neither mutually exclusive nor exhaustive. Instead, they are a helpful means to compare and understand the different approaches developed in the field of XAI. The following subsections describe the classification dimension, and examples of current research are provided.

2.2.1 Relationship to the Predictive System

A straightforward strategy to increase the transparency of AI systems is to use intrinsically interpretable ML models. This approach, called *ante-hoc*, uses the same ML model for predicting and explaining (Sokol & Flach, 2020). These intrinsically interpretable models are inherently transparent, as their parameters directly reveal how the model works (Carvalho et al., 2019). For instance, in linear and logistic regression models, the weights assigned to each model feature can be directly interpreted or processed to help understand how they influence the model's decisions (Molnar et al., 2020). Moreover, in decision tree models, it is possible to extract rules that can be used to explain the model's decisions.

Nonetheless, the interpretability of ante-hoc models is directly related to their complexity. Usually, the higher the complexity, the more difficult it is to explain its inner workings (Adadi & Berrada, 2018). These models' interpretability is frequently enhanced by implementing constraints such as monocity, model size, or sparsity (Došilović et al., 2018). However, there is generally a trade-off between

interpretability and accuracy for these models, as simpler models are usually not the most accurate ones (Breiman, 2001).

An alternative approach called *post-hoc* has been proposed to avoid the trade-off between interpretability and accuracy. In post-hoc explainability methods, predictions are made by a complex "black-box" model, while a simpler model generates explanations. The simpler model generates explanations after the predictive model has been trained by attempting to mimic the behavior of the complex model using feature summaries or visualizations thereof (Molnar et al., 2020; Sokol & Flach, 2020). Separating predictions and explanations brings much flexibility, as these methods can generate explanations for other similar predictive models. Nevertheless, it is necessary to develop such post-hoc methods carefully to avoid generating easily interpretable but misleading explanations (Došilović et al., 2018). Prominent examples of post-hoc explainability methods include Partial Dependence Plots (DPD) (Friedman 2001) and Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015). To explain the predictive model, DPDs show features' marginal effect on the predicted outcome, providing an overview of their relationship across all values. In contrast, ICE plots show how the prediction of a specific instance changes accordingly to changes in a particular feature (Molnar, 2020).

2.2.2 Explanation Scope

XAI methods can also be classified according to the scope of their explanations on *global* and *local*. On the one hand, global methods provide a comprehensive, holistic model explanation, which describes the entire model behavior across all instances for a given dataset (Guidotti, Monreale, Ruggieri, Turini, et al., 2018). Therefore, global methods can help investigate population-level effects, such as identifying factors influencing drug consumption or climate change (Adadi & Berrada, 2018). For instance, Accumulated Local Effects (ALE) plots describe how features influence the prediction of an ML model on average (Apley & Zhu, 2016).

On the other hand, local methods provide explanations for specific instance decisions, which means that explanations are generated considering the vicinity of the instance to be explained (Molnar, 2020). Local explanation methods utilize the idea that even complex models expose a more simple, comprehensible behavior locally around the instance of interest (Carvalho et al., 2019). A prominent example is Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016b), which locally approximates the behavior of a complex model with a simple, interpretable model. Likewise, Anchors generates rules that determine the prediction of the instance of interest (Ribeiro et al., 2018). In other words, Anchors prescribes a set of rules for specific feature values so that altering other features does not change the prediction outcome.

2.2.3 Applicability

Concerning their applicability, XAI methods are classified across three degrees of portability: *model-specific*, *model-class-specific*, and *model-agnostic* (Robnik-Šikonja & Bohanec, 2018). Model-specific methods are limited to providing explanations only for the specific predictive model on which they are trained. For example, rules extracted from a decision tree model to provide explanations would not be valid for other decision tree models. Alternative, model-class-specific methods are generalizable to provide explanations for a specific model family. Examples of model-specific methods can be found in computer vision, where approaches have been developed specifically to explain the behavior of neural network models (Guidotti, Monreale, Ruggieri, Turini, et al., 2018). In contrast, model-agnostic methods are not bound to any particular model or model family and can generate explanations for any model (Adadi & Berrada, 2018; Molnar et al., 2020).

Model-specific and model-class-specific methods limit the choice of ML models that can be used for a given task. In contrast, model-agnostic methods are usually applied post-hoc after the predictive model has been trained. Therefore, extensive research has focused on developing model-agnostic explainability methods due to the broader applicability offered by their decoupling from the prediction model (Ribeiro et al., 2016b).

2.2.4 Explanation Family

According to Johnson and Johnson (1993), explanation types can be grouped into three main categories: *association between antecedent and consequent*, *contrast and differences*, and *causal mechanisms*. Association between antecedent and consequent includes explanation approaches that utilize item(s)-predictions relations such as influential instances (Sokol & Flach, 2020). These methods select particular instances of the dataset to explain the underlying data distribution and are more suitable for data humans can easily understand (e.g., images or text) (Agnar & Plaza, 1994; Molnar, 2020). For example, Koh and Liang (2017) utilized influence functions to identify instances on the training dataset that are more responsible for a given prediction. Furthermore, this category also includes approaches that consider the relationship between features and predictions (Sokol & Flach, 2020). Some methods, such as DPD (Friedman, 2001) and ICE plots (Goldstein et al., 2015), describe how features influence all model's predictions to provide global explanations. In contrast, methods such as LIME (Ribeiro et al., 2016b) analyze the influence of features on a particular decision to provide local explanations. Alternatively, other methods, such as Shapley Additive Explanations (SHAP) (S. M. Lundberg et al., 2017), analyze features' effects on predictions to generate local and global explanations.

The category *contrasts and differences* includes approaches that evaluate the similarities and dissimilarities of instances in the dataset. For instance, B. Kim et al. (2016) proposed using representative instances called prototypes and instances not well represented by those prototypes

(criticisms) to provide global explanations. Likewise, Ribeiro et al. (2018) developed Anchors, a method that generates local explanations by analyzing similar instances to derive high-precision rules representing sufficient conditions for the prediction. Moreover, this category also includes approaches that utilize contrasts to present explanations. Class-contrastive counterfactual statements are a prominent example that has gained interest in the literature, as they are believed to be comprehensible, human-friendly explanations (Miller, 2019). Counterfactual explanations describe a causal relationship in the form of "if X had not occurred, Y would not have occurred" (Molnar, 2020). Examples of model-agnostic methods that provide counterfactual explanations include Aggregated Visual Counterfactual Explanations (AdViCE) (Gomez et al., 2021) and Diverse Counterfactual Explanations (DICE) (Mothilal et al., 2020).

Finally, the category of *causal mechanisms* considers approaches that generate explanations by analyzing causal relationships. Historically, researchers have used causal models to analyze the causal relationships from statistical data in an individual system or a population (Hitchcock, 2018). Some explainability approaches, such as counterfactual statements and DPD, are considered to have a causal interpretation because they analyze which changes to the input attributes lead to a given prediction (Molnar, 2020). Other examples of causal approaches include the work of Heskes et al. (2020) and Frye et al. (2020), which adapted the concept of Shapley Values to generate causal explanations.

2.3 Interpretability Framework

Doshi-Velez and Kim (2017) proposed a framework for assessing the interpretability of XAI methods, which has been widely accepted in the literature. This framework suggests three types of interpretability evaluations that help determine the efficiency of XAI methods: *application-grounded*, *functionally-grounded*, and *human-grounded*. Application-grounded evaluations involve conducting experiments in real applications with the end-users for which the AI system is intended. These evaluations should be performed with domain experts who test the explanations for the application they were designed for (Doshi-Velez & Kim, 2017). While application-grounded evaluations are the most accurate type, they are also the most expensive and rarely used in scientific research.

Human-grounded evaluations encompass experiments conducted with real humans on simplified tasks that involve XAI methods and capture the essence of their target application. The simplification of tasks allows the use of laypersons instead of domain experts, which significantly enhances the accessibility of these evaluations. These evaluations are most appropriate to assess the general quality of explanations. Furthermore, they represent an alternative when an evaluation with the target users is challenging. (Doshi-Velez & Kim, 2017).

Finally, functionally-grounded evaluations are conducted without humans on proxy tasks that capture the quality of an explanation according to custom interpretability metrics. These evaluations are often used when only a few resources are available and are most appropriate once an explainability method has been validated with users. (Doshi-Velez & Kim, 2017). Additionally, these evaluations can assess an explainability method that is not mature or when there are ethical challenges when experimenting with human users. Nevertheless, functionally-grounded evaluations cannot capture end-users' perceptions of provided explanations.

2.4 Investigated Local Model-agnostic XAI Methods

Throughout the research projects incorporated in this dissertation, four local model-agnostic XAI methods were investigated to understand how end-users evaluated the design of their visual representations. These methods are all available as open-source Python packages. The four local model-agnostic XAI methods are (1) LIME,[7] (2) Anchors,[8] (3) SHAP,[9] and (4) DICE.[10] The following sections provide an overview of these methods and provide information on how they generate explanations.

2.4.1 LIME

LIME (Ribeiro et al., 2016b) aims to find an interpretable model locally faithful to the underlying predictive model to explain individual predictions. Artificial data points are created by drawing perturbed versions of the training data distribution to train this interpretable model. The number of changed features is random for each perturbed instance, and each feature is perturbed individually. Then, a prediction function is used to compute the labels of these synthetic samples. LIME then uses these perturbed instances to train the interpretable model scaling each perturbed instance by a proximity measure so that data points closer to the instance of interest carry more weight. An example of a LIME explanation is presented in Figure 3. The explanation shows the weights of a linear model in a plot that represents each feature value's influence on the classifier's prediction.

LIME provides significant flexibility as it can handle tabular, image, and text data, influencing its popularity with practitioners. However, LIME's sampling strategy has received criticism due to its substantial amount of randomness, which results in a lack of robustness (Zhang et al., 2019). This means that the resampling of the instances can lead to different parameters of the interpretable model, which would generate different explanations. Additionally, it has been argued that the perturbed samples can contain many incorrectly classified instances, as the sampling process does not consider the density of the data (Guidotti et al., 2019), which can produce data points with high prediction uncertainty.

[7] https://github.com/marcotcr/lime
[8] https://github.com/marcotcr/anchor
[9] https://github.com/slundberg/shap
[10] https://github.com/interpretml/DiCE

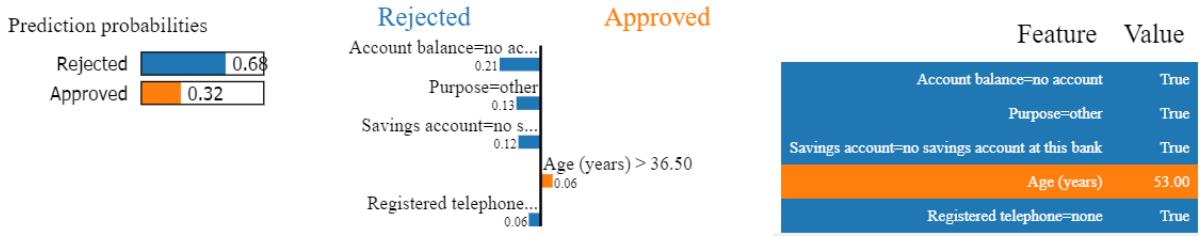


Figure 3: Example of a LIME explanation for a bank loan dataset instance.

2.4.2 Anchors

Anchors (Ribeiro et al., 2018) is an XAI method that addresses some of LIME's drawbacks by explaining the logic of a predictive model with high-precision rules representing local, sufficient conditions for predictions. In other words, Anchors prescribes a set of rules for specific feature values so that altering other features does not change the prediction outcome. These rules are computed for a single instance, so they "anchor" the respective model prediction for this instance. Figure 4 shows an example of an Anchors' explanation. It presents the list of rules that, if fulfilled, would lead to the AI system predicting the instance as rejected 95.4% of the time.

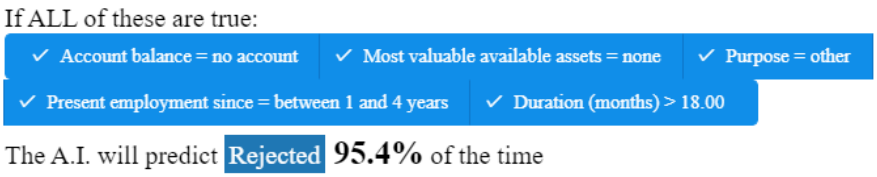


Figure 4: Example of an Anchors' explanation for a bank loan dataset instance.

The simple conditional rules Anchors generates are regarded as easy to interpret due to their clear coverage (Lakkaraju & Bastani, 2020). Thus, users know when to generalize an explanation given for an instance to other instances, which can help reduce the mental effort required to comprehend explanations. However, Anchors' parameters need to be carefully tuned to obtain concise rules (Guidotti, Monreale, Ruggieri, Pedreschi, et al., 2018). A drawback of Anchors is that it can generate specific rules for instances close to the boundary, which can be rather complex (Ribeiro et al., 2018).

2.4.3 SHAP

SHAP (S. M. Lundberg et al., 2017) can generate local and global explanations by providing feature attribution scores based on the concept of Shapley values from cooperative game theory. These scores estimate how to fairly distribute the contribution of features for a particular prediction (Weerts et al., 2019). SHAP is an adaptation from LIME that provides the Shapley values as the linear regression weights in an additive feature attribution method (S. M. Lundberg et al., 2017). Figure 5 shows an example of a SHAP explanation. The base value represents the expected value of the prediction function over the training dataset, and the arrows represent the influence of each feature value toward increasing

or decreasing the prediction score. These features' influences balance each other and produce the model's output score "f(x)", the classificatory prediction probability for the represented class.

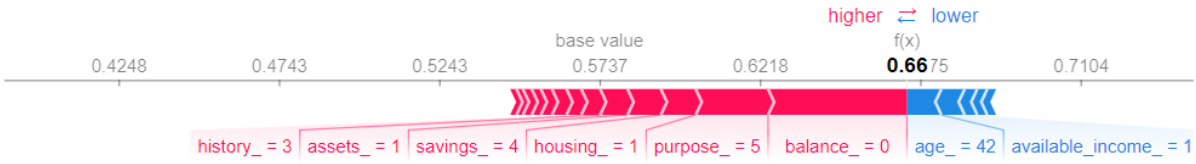


Figure 5: Example of a SHAP explanation for a bank loan dataset instance.

SHAP is built on a solid theoretical foundation and produces explanations that can be contrasted. In other words, explanations for a given instance can be compared with other instances or a subset of instances because Shapley values result from all the feature value collisions in the dataset (Molnar, 2020). This contrastive property is not possible with local methods, such as LIME. Additionally, in theory, Shapley values can guarantee locally accurate and replicable explanations. This means that the linear model used to generate the explanations would recover the exact features' Shapley values and prediction score of the underlying model for a given instance. However, an exact computation of Shapley values is computationally expensive. Thus, approximate values are calculated, sacrificing some variability in recalculations (S. M. Lundberg et al., 2017).

2.4.4 DICE

Counterfactual explanations originated from research in social science and were first introduced as an XAI concept by Wachter et al. (2017). They describe a causal relationship in the form "if X had not occurred, Y would not have occurred" (Molnar, 2020). In the context of AI systems, they describe how the feature values of an instance would have to change to obtain a different, desirable output from the predictive model. Counterfactuals cannot reveal the internal mechanisms of a model. However, they can establish some causal relationships, as the altered feature values directly influence the predictive model's outcome. Due to their solid theoretical foundation in the psychological and philosophical literature, counterfactuals are believed to be comprehensible, human-friendly explanations (Miller, 2019).

Mothilal et al. (2020) acknowledged that not all counterfactual explanations might be feasible for users due to the proposed modifications to the values of the features. Building on the premise that presenting a diverse set of information items to users provides benefits in other domains of information search, the authors suggest that diversity could be beneficial when users are shown counterfactual explanations. Thus, they developed a method for generating a diverse set of counterfactual explanations (i.e., DICE). As diversity and proximity of counterfactuals come with a natural trade-off, the authors extended the work from Wachter et al. (2017) to explicitly include diversity in the counterfactuals search, in addition to proximity. DICE allows the input of relative difficulty in changing a feature by specifying feature

weights. The specification of the difficulty in changing a feature can be helpful to restrict the search for counterfactuals and avoid generating explanations that include immutable feature changes. DICE presents counterfactual explanations in a tabular form where the counterfactual changes for each feature can be visualized. Figure 6 shows an example of a DICE counterfactual explanation with only a subset of the features from the bank loan dataset.

	Amount (EUR)	Duration (months)	Purpose	Account Balance	Employment
Original Input Outcome: Rejected	4870	24	other	no account	between 1 and 4 years
Counterfactuals Outcome: Approved	4522	--	new car	--	--
	5080	26	--	above 200 EUR	--
	4878	23	--	--	--

Figure 6: Example of a DICE counterfactual explanation for a bank loan dataset instance.

3 Study I: Designing Local Model-agnostic Explanation Representations and an Experimental Evaluation using Eye-tracking Technology[11]

3.1 Introduction

AI is playing an increasingly important role in high-stake domains such as finance (Binns et al., 2018), criminal justice (Hartmann & Wenzelburger, 2021), and healthcare (Stowell et al., 2018). Nevertheless, the effectiveness of AI systems is limited by their inability to explain their decisions to human users (D. Wang et al., 2019). Specifically, many AI systems are opaque, and their underlying ML models are considered "black boxes" where only the model's output is available. This lack of transparency leaves the inner working mechanisms of the AI system unclear to end-users. This problem is reinforced by the popularity of deep learning models, which are hard to understand even by experts (Dodge et al., 2019). Due to the need for users to understand the logic of these systems, new regulations have been enacted that provide the "right to explanations" for all decisions made or supported by AI systems (Cheng et al., 2019; Dodge et al., 2019). Nevertheless, such regulations still need to be put into practice. Overall, it remains difficult for non-experts to understand the logic behind AI systems and figure out how specific inputs lead to a particular output (Cheng et al., 2019).

In this light, there has been a surge of interest in XAI among scholars and practitioners seeking to produce models that can explain their decisions in non-technical terms while maintaining a high prediction accuracy (Diakopoulos et al., 2017; Turek, 2018). XAI aims to support human understanding and trust in the use of AI systems (Turek, 2018). In particular, extensive research has focused on developing so-called "model-agnostic" explainable methods. These methods are applied after the predictive model has been trained and can explain any predictive model (Sokol & Flach, 2020). Some of these model-agnostic methods provide local explanations, which clarify how individual predictions are made. In contrast, others offer global explanations, which describe the entire model behavior across all instances for a given dataset (Molnar, 2020). This study focuses on model-agnostic methods that provide local explanations (e.g., LIME (Ribeiro et al., 2016b)) as they provide great flexibility by isolating the explanations from the underlying predictive model (Ribeiro et al., 2016a). As a result, practitioners can assess and compare different models and even migrate to a new model later (Adadi & Berrada, 2018; Molnar, 2020). Furthermore, users can understand why AI systems make particular decisions (Buçinca et al., 2020; Sokol & Flach, 2020).

[11] This chapter is based on the following paper (Meza Martínez et al., 2023). The data analysis and experimental material can be found in the following GitLab repository and RADAR archive:
https://git.scc.kit.edu/h-lab/research/1897_meza_miguel_design-of-xai-representations-and-eye-tracking-analysis
https://radar.kit.edu/radar/de/dataset/vHLviZcXQbjMvNOB

However, despite the advances within the XAI field, challenges remain. First, strong critiques have been made that explanation representations are still based on researchers' intuition rather than a deep understanding of users' needs (Miller, 2019; D. Wang et al., 2019). Among others, Miller (2019) argues that experts who train and evaluate models may lack the judgment necessary to assess the usefulness of the generated explanation representations for users (see also Adadi & Berrada, 2018; Mittelstadt et al., 2019; Ribera & Lapedriza, 2019). Thus, it is necessary to involve users in the design process when creating or refining local model-agnostic explanation representations. An appropriately designed explanation representation can help increase users' trust and understanding of the AI system's decisions (Bussone et al., 2015; Dodge et al., 2019).

Second, there is a lack of empirical research on how actual users evaluate representations of local model-agnostic explanations from different XAI methods (Kaur et al., 2019). Many studies have focused on evaluating explanations without users (D. Wang et al., 2019), using some formal definition of interpretability as a proxy for the quality of the explanations they provide (Doshi-Velez & Kim, 2017). However, this type of evaluation seems most appropriate when a method is not yet mature or has already been evaluated by users (Doshi-Velez & Kim, 2017). Other studies have rather focused on evaluating model-agnostic explanations with experts (Jesus et al., 2021; Kaur et al., 2019) or have evaluated model-agnostic explanations against a non-model-agnostic baseline or no explanations (B. Kim et al., 2016; Ribeiro et al., 2016b). Finally, some studies that evaluated local model-agnostic explanations from different XAI methods with users relied solely on textual representations (Binns et al., 2018; Dodge et al., 2019). Thus, further research is needed to evaluate and compare holistically the representations of local model-agnostic explanations generated by different XAI methods to reach a consensus on which are more appropriate for users (D. Wang et al., 2019).

To evaluate XAI explanations with end-users, researchers commonly rely on a combination of objective performance measures on an evaluation task and subjective self-reported measures of constructs such as trust and understandability (Buçinca et al., 2020; Doshi-Velez & Kim, 2017). Nevertheless, studies have found that objective performance measures might not generalize well from the evaluation task and that subjective measures might not predict the utility of explanations in decision-making tasks (Buçinca et al., 2020). Furthermore, subjective measures only provide a limited view of how end-users utilize explanations during an evaluation task, as their data collection occurs after task completion (Naiseh et al., 2022). Therefore, researchers have argued for incorporating complementary data collection methods based on non-intrusive measures that provide insight into how end-users actually utilize explanations during the evaluation task (Naiseh et al., 2022). In particular, there seems to be an increasing interest in incorporating the use of eye-tracking technology in order to collect data on end-users' visual attention (see, e.g., Barria-Pineda et al., 2021; Evans et al., 2022; Naiseh et al., 2022; Schneider & Handali, 2019). The use of eye-tracking technology in research has considerably increased in recent years with the development of more accurate and affordable devices due to their capacity to provide insights into end-

users' cognitive processes and, specifically, visual attention (Duchowski, 2017; Hayhoe & Ballard, 2005; Liversedge & Findlay, 2000; Rayner, 1998). For instance, eye-tracking has been used to analyze end-users' comprehension of graphs (Abdul et al., 2020; Strobel et al., 2016) and to investigate how end-users evaluate visual analytics (Kurzhals et al., 2015). Nevertheless, Nevertheless, the use of eye-tracking technology has not received much attention in evaluating local model-agnostic explanation representations of XAI methods.

Against this backdrop, the goal of this paper is to address the following research questions:

RQ1*: How to design local model-agnostic explanation representations from different XAI methods following a user-center approach?*

RQ2*: How do end-users perceive, evaluate, and visually attend to the designed local model-agnostic explanation representations from different XAI methods?*

To achieve this goal, first, representations of explanations from four well-established local model-agnostic XAI methods were refined following an iterative design process involving end-users (Iivari, 2015). This iterative refinement increased the explanation representations' comparability while controlling for confounding factors due to their different explanation approaches. The selected methods are Anchors (Ribeiro et al., 2018), DICE (Mothilal et al., 2020), LIME (Ribeiro et al., 2016b), and SHAP (S. M. Lundberg et al., 2017) (a detailed overview of each method is provided in Section 2.4). After the iterative refinement of the explanation representations, these were evaluated with 19 participants in a laboratory experiment using eye-tracking technology and self-reports, followed by interviews. The evaluation was centered in the bank loan applications domain, where an AI system predicts the decision to approve or reject a loan application by evaluating its risk using a set of attributes. Details regarding the selected domain and dataset used are found in Section 3.3.1.

The contributions of this study are two-fold. First, an iterative design process involving end-users in refining the representations of local explanations from well-established model-agnostic methods was performed. This iterative design process provides insightful information on how researchers can increase the comparability of explanation representations for well-established XAI methods. Moreover, end-users' evaluations throughout the iterative design process inform how they perceive the explanation representations. Second, leveraging eye-tracking technology, a laboratory experiment was conducted to evaluate how end-users visually attend to these explanation representations. This approach extends the work from a limited number of XAI studies that integrated eye-tracking technology by evaluating multiple model-agnostic XAI methods with users (Bigras et al., 2019; Coba et al., 2019; Conati et al., 2021; Karran et al., 2022; Muddamsetty et al., 2022; Polley et al., 2021). As a result, this study provides an interesting perspective on how end-users utilize different explanation representations and which ones they prefer. Additionally, as a practical contribution, this study provides an open-source reference

implementation of the refined explanation representations and the implementation of the selected model-agnostic methods.[12]

The remainder of the paper is as follows. First, this study provides background information and related work on local model-agnostic explanations from XAI methods and eye-tracking technology to evaluate end-users' visual attention. After that, the task domain, dataset, ML model used in the experiments, four selected XAI methods, and the pursued iterative design approach are described. Next, the iterative design process results and refined explanation representations are presented. Then, the eye-tracking laboratory study and its results are described in detail. Finally, results are discussed, and possible future research directions are suggested.

3.2 Related Work

As a result of the extensive research performed in XAI in different research communities over the last few years, researchers and practitioners have developed many innovative algorithms, visualizations, interfaces, and toolkits. For instance, some methods extract easily interpretable rules from the predictive model and present them to end-users as an explanation of the model's decision (Deng & Brown, 2021; Ming et al., 2019; Thomas et al., 2021; Yuan et al., 2021). Alternatively, others highlight regions of an image to indicate which pixels were influential in the model's prediction (Landecker et al., 2013; Xu et al., 2015; B. Zhou et al., 2018). Several studies have surveyed the literature to provide a detailed overview of XAI by presenting the different approaches developed and providing classifications or conceptual frameworks (e.g., Adadi & Berrada, 2018; Carvalho et al., 2019; Guidotti, Monreale, Ruggieri, Turini, et al., 2018; Miller, 2019; Molnar, 2020; Tjoa & Guan, 2019). This section first presents some of the different model-agnostic explanations proposed in the literature. After that, the use of eye-tracking in evaluating XAI explanations is discussed. Finally, related work that evaluates local explanations from multiple XAI model-agnostic methods is presented. Further information on the general classification of XAI methods is presented in Section 2.2.

3.2.1 Model-agnostic XAI Methods

As previously mentioned, many model-agnostic XAI methods have been proposed in the literature to provide explanations of AI systems predictions. These methods vary significantly in their approach to generating explanations. To provide an overview of the explainable approaches used by some model-agnostic XAI methods, this study relies on a categorization across three categories (1) association between antecedent and consequent, (2) *contrast and differences*, and (3) *causal mechanisms* (H. Johnson & Johnson, 1993; Sokol & Flach, 2020).

[12] https://github.com/miguelmezamartinez/Local-Model-agnostic-Explanations-Representations

Association between antecedent and consequent includes explanation approaches that utilize item(s)-predictions relations such as influential instances (Sokol & Flach, 2020). These methods select particular instances to explain the underlying data distribution and are more suitable for data humans can easily understand (e.g., images or text) (Agnar & Plaza, 1994; Molnar, 2020). For example, Koh and Liang (2017) utilize influence functions to identify instances on the training dataset that are more responsible for a given prediction. Furthermore, this category also includes approaches that consider the relationship between features and predictions (Sokol & Flach, 2020). Some methods, such as DPD (Friedman, 2001) and Individual Conditional Expectation (ICE) plots (Goldstein et al., 2015), describe how features influence all model's predictions to provide global explanations. In contrast, methods such as LIME (Ribeiro et al., 2016b) analyze the influence of features on a particular decision to provide local explanations. Alternatively, other methods, such as SHAP (S. M. Lundberg et al., 2017), analyze features' effects on predictions to generate local and global explanations.

The category contrasts and differences includes approaches that evaluate the similarities and dissimilarities of instances in the dataset. For instance, B. Kim et al. (2016) proposed using representative instances called prototypes and instances not well represented by those prototypes (criticisms) to provide global explanations. Likewise, Ribeiro et al. (2018) developed Anchors, a method that generates local explanations by analyzing similar instances to derive high-precision rules representing sufficient conditions for the prediction. Moreover, this category also includes approaches that utilize contrasts to present explanations. Class-contrastive counterfactual statements are a prominent example that has gained interest in the literature, as they are believed to be comprehensible, human-friendly explanations (Miller, 2019). Counterfactual explanations describe a causal relationship in the form of "if X had not occurred, Y would not have occurred" (Molnar, 2020). Examples of model-agnostic methods that provide counterfactual explanations include AdViCE (Gomez et al., 2021) and DICE (Mothilal et al., 2020).

Finally, the category of causal mechanisms considers approaches that generate explanations by analyzing causal relationships. Historically, researchers have used causal models to analyze the causal relationships from statistical data in an individual system or a population (Hitchcock, 2018). Some explainability approaches, such as counterfactual statements and DPD, are considered to have a causal interpretation because they analyze which changes to the input attributes lead to a given prediction (Molnar, 2020). Other examples of causal approaches include the work of Heskes et al. (2020) and Frye et al. (2020), which adapted the concept of Shapley Values to generate causal explanations.

3.2.2 Eye-Tracking Technology in XAI Research

Since eye-tracking was first used to investigate human visual perception over a century ago, many methods have been proposed to track eye movement and utilize it with various goals (Rayner, 1998). Generally, eye-tracking has been used for interactive and diagnostic purposes (Duchowski, 2002). While

users' eye movement data is used as an input modality in an interactive role, in a diagnostic role, it is used as a cue for estimating their intentions and cognitive states (Duchowski, 2017; Holmqvist et al., 2011). For evaluating the usability of human-computer interfaces, eye-tracking has been identified as an objective source of data that provides an understanding of users' visual information processing (Poole & Ball, 2006). In recent years, the development of more accurate and affordable eye-tracking devices and their non-intrusive data collection approach has increased their utilization in research.

To gain an overview of studies in the literature incorporating eye-tracking in the evaluation of explanations, a search was performed in established digital libraries for studies focusing on "explainable artificial intelligence" or "interpretable machine learning" systems, together with eye-tracking terminology. Table 1 presents a summary of studies found along the following attributes: (1) context domain, (2) eye-tracking measures, and (3) evaluated explanations.

Bigras et al. (2019) incorporated eye-tracking technology in an e-commerce context to investigate users' behavior toward recommendation agents (RA). Their research controlled for RA's transparency through model-specific feature attribution explanations. They utilized two established standard eye-tracking metrics (i.e., number of fixations and fixation duration) to investigate users' cognitive effort when interacting with RAs (Lorigo et al., 2008). Likewise, Coba et al. (2019) utilized eye-tracking data to analyze users' decision-making strategies when evaluating model-specific summarizations of rating distributions in an e-commerce context. Besides the number of fixations and fixation duration, they also incorporated the analysis of transitions and revisits (Payne, 1976) between areas of interest (AOIs) to investigate how users examined alternatives when making decisions.

Conati et al. (2021) incorporated eye-tracking in their investigation of the value of model-specific explanations of AI-driven hints in the context of intelligent tutoring systems. They analyze the number of fixations and fixation duration to capture how much time participants spent looking at explanations. Likewise, Polley et al. (2021) incorporated eye-tracking data to evaluate their proposed model-class-specific global and local explanations in the context of search systems. They generated heatmaps from users' eye-tracking data to explore scanning patterns and investigate users' attention on regions of the provided explanations.

Karran et al. (2022) utilized eye-tracking technology in the context of image classification to investigate how different model-class-specific explanation visualization strategies impact users' trust (Selvaraju et al., 2017; Sundararajan et al., 2017). Specifically, they used pupil dilatation to infer users' cognitive load when evaluating explanations. Meanwhile, Muddamsetty et al. (2022) followed an alternative approach to evaluate the quality of model-class-specific visual explanations generated by an XAI method utilizing eye-tracking data (Muddamsetty et al., 2022; Petsiuk et al., 2018; Selvaraju et al., 2017). In particular, they evaluated explanations generated by the XAI method Similar Difference and Uniqueness (SIDU), which provides visual saliency maps highlighting regions responsible for the

prediction in image classification. They gathered eye-tracking data from users evaluating images for object recognition and generated heatmaps representing the users' fixations on different image regions. Subsequently, they compared these user-generated heatmaps against the visual saliency maps to evaluate the quality of the explanations.

Table 1: Summary of evaluation studies using eye-tracking in XAI research (sorted by publication date).

	Context Domain					Eye-tracking Measures				Evaluated Explanations					
										Model-specific	Model-class-specific				
Studies	E-Commerce	Image Classification	Search Systems	Text Analysis	Tutoring Systems	Fixation-based	Heatmap-based	Pupil Dilatation	Transition-based	Feature attribution	Feature attribution	Grad-CAM (Selvaraju et al., 2017)	SIDU (Muddamsetty et al., 2022)	RISE (Petsiuk et al., 2018)	Integrated Gradients (Sundararajan et al., 2017)
Bigras et al., 2019	X					X				X					
Coba et al., 2019	X					X			X	X					
Conati et al., 2021					X	X				X					
Polley et al., 2021			X				X				X				
Karran et al., 2022		X						X				X			X
Muddamsetty et al., 2022		X					X					X	X	X	

In summary, the search results of the literature review indicate that research that leverages eye-tracking to evaluate explanations has mainly focused on model-specific or model-class-specific methods. However, the results indicate a lack of studies leveraging eye-tracking for evaluating explanations from model-agnostic XAI methods.

3.2.3 Local Model-agnostic Explanations

To gain an overview of similar work conducted in the literature so far, a search was conducted in established digital libraries for studies that conducted evaluations with different types of participants of local model-agnostic explanations from at least two XAI methods. Furthermore, research that evaluated global model-agnostic explanations or evaluated model-agnostic explanations against non-model agnostic explanations was not considered. Table 2 presents a summary of the comparative studies found along the following attributes: (1) XAI methods, (2) evaluation context, (3) type of participants, (4) representation used, and (5) evaluation measures.

Next, a search was performed in established digital libraries for related work that conducted evaluations of local model-agnostic explanations generated by XAI methods. The search focused on publications

that evaluated local model-agnostic explanations of at least two XAI methods. Furthermore, research that evaluated global model-agnostic explanations or evaluated model-agnostic explanations against non-model agnostic explanations was not considered. Table 2 presents a summary of the comparative studies found along with the following attributes: (1) XAI methods, (2) evaluation context, (3) type of participants, and (4) representation used.

Binns et al. (2018) conducted between-subjects experiments with students and users to evaluate the effect of explanation styles on perceived fairness in different contexts using self-reported measures. They compared input influence-based explanations (i.e., LIME and QII) (Datta et al., 2016; Ribeiro et al., 2016b), case-based explanations (Doyle et al., 2003), demographic explanations (Ardissono et al., 2003), and a type of counterfactual called sensitivity-based explanations (Rasmussen et al., 2012). Their analysis shows that sensitivity-based explanations led to a significantly higher fairness perception than case-based and demographic explanations, while the difference with input influence-based explanations was not significant. Dodge et al. (2019) also evaluated the same explanation types with users in a criminal justice context using self-report measures and additionally manipulated the underlying classifiers' fairness. They found that sensitivity-based explanations were most effective at exposing fairness discrepancies. Nonetheless, in both studies, the authors used purely textual explanations to control the representation difference between the evaluated methods.

Ribeiro et al. (2018) compared Anchors and LIME explanations in different contexts using the original explanation representations proposed in their developed libraries. In a within-subjects design using both self-report and performance measures, students were presented first predictions without explanations and then with explanations from one of the methods in a randomized order. Additionally, students had to guess the model's prediction for additional instances before and after each round of explanations. They found that students achieved a higher prediction accuracy using Anchors' explanations. For LIME, the authors found that the prediction accuracy varied drastically and, in some cases, was worse than for no explanations. Furthermore, it took significantly less time to understand and use Anchors' explanations, which the authors attribute to their simplicity and generalizability.

Table 2: Evaluation studies of local model-agnostic XAI explanations (sorted by publication date).

	XAI Methods											**Context Domain**					**Type of Participants**				**Repre-sentation**			**Evaluation Measures**		
Authors, year	Anchors (Ribeiro et al., 2018)	Case-based (Doyle et al., 2003)	CluReFI (El Bekri et al., 2019)	Decision Boundary (Hase & Bansal, 2020)	Demographic (Ardissono et al., 2003)	GAM (Friedman, 2001)	LIME (Ribeiro et al., 2016b)	Prototype (Hase & Bansal, 2020)	QII (Datta et al., 2016)	Sensitivity Analysis (Rasmussen et al., 2012)	SHAP (S. M. Lundberg et al., 2017)	Economy	Employment	Entertainment	Justice	Science	Data Scientist	Domain Experts	Students	End-users	Adapted	Original	Textual	Behavioral (e.g., eye-tracking)	Self-reported	Performance
Binns et al., 2018		X			X		X		X	X		X	X	X					X	X			X		X	
Ribeiro et al., 2018	X						X					X			X	X			X			X			X	X
Dodge et al., 2019		X			X		X		X	X					X					X			X		X	
Kaur et al., 2019						X					X		X				X					X			X	X
El Bekri et al., 2019			X				X					X								X		X			X	
Hase & Bansal, 2020	X			X			X	X					X	X					X		X		X		X	X
Jesus et al., 2021							X				X	X						X					X		X	X

Kaur et al. (2019) studied data scientists' use of Generalized Additive Models (GAMs) (Friedman, 2001) and SHAP explanations to uncover common issues when building a model in the context of salary predictions. Their research relied on GAMs' and SHAP's original explanation representations. Their analysis using self-report and performance measures revealed that data scientists often "misuse" and over-trust interpretability tools and that the representations of explanations were hard to understand and could be misleading. Additionally, the results show that participants using GAMs had significantly higher accuracy and confidence in their understanding and lower cognitive load than those using SHAP.

El Bekri et al. (2019) evaluated explanations from LIME against their proposed method CluReFI and a baseline with users in the context of bank loan applications. CluReFI extended LIME by first clustering instances and then providing LIME explanations for an instance that is the cluster's representative. Their research relied on self-report measures to evaluate LIME's original explanation representation. Their results indicated that explanations increase trust in the system and that users preferred LIME explanations due to their balance between detail and simplicity.

Hase and Bansal (2020) investigated the effect of explanations from LIME, Anchors, a type of counterfactual explanation called Decision Boundary, a Prototype model, and a composite method that combined the other four explanations.[13] Their research relied on adaptations of the original explanation representations and textual representations and utilized performance and self-report measures. These adaptations include changing LIME's color-coding for the attributes' influence, presenting Anchors'

[13] Table 2 only includes single methods. Composite methods were not included to improve readability.

explanations as a probabilistic statement, and presenting counterfactuals as a series of attribute changes that lead to crossing the decision boundary. The authors conducted a between-subjects experiment with students in two contexts. Students had to guess the model's prediction without explanations and then with explanations of one type. They found that neither explanation type led to a significant increase in task performance.

Jesus et al. (2021) evaluated explanations from LIME, SHAP, and a non-model-agnostic method on a real-world fraud detection task with domain experts (i.e., fraud analysts). They investigated whether textual explanation representations from these methods could increase domain experts' performance compared to a baseline without explanations using self-report and performance measures. Their results show that presenting no explanations resulted in the highest precision and the slowest decision time. Furthermore, LIME was the least preferred explanation type due to the low diversity of features shown in its explanations.

In summary, the literature review results indicate that research that evaluates local model-agnostic explanations from multiple XAI methods is scarce (i.e., seven articles in total). Studies have focused on performing these evaluations along different types of participants and contexts. Moreover, three out of seven studies solely focused on textual representations. However, none of these studies has incorporated complementary data collection methods based on non-intrusive measures, such as eye-tracking, to provide insights into how end-users utilize explanations. This study aims to address this unfolding research gap by relying on eye-tracking technology.

3.3 Design Context and Methodology

This section presents an overview of important contextual constraints (i.e., domain, dataset, ML model, evaluated XAI methods) and the iterative design method followed in this study. First, the domain, dataset, and ML model used are presented. Afterward, the evaluated XAI model-agnostic methods and their out-of-the-box representations for providing local explanations are described. Then, the iterative evaluation design, measures, and analysis strategy are presented.

3.3.1 Domain, Dataset, and Model

The bank loan applications domain was selected to design and evaluate the explanation representations. Specifically, a scenario in which an AI system evaluates the risk of bank loan applications using a set of attributes and predicts the decision to approve or reject them was used. This scenario is commonly used in XAI research since bank loan decisions typically involve a notion of trust in the AI system (Adadi & Berrada, 2018; Aggarwal et al., 2019; Binns et al., 2018; Chakraborty et al., 2020) and end-users are familiar with the general process of requesting a loan in a bank. Furthermore, this domain is

highly relevant as financial institutions increasingly use AI systems to evaluate loan applications (Binns et al., 2018), and the resulting decisions can significantly impact loan applicants.

Moreover, the publicly available, open-source German Credit dataset was used (Dua & Graff, 2017), which contains 1,000 instances of loan applications, each represented by 20 attributes describing the details of the loan application and the applicant's financial and personal information. The dataset was modified to adjust it for the research goal. For instance, the attributes "personal status and sex" and "foreign worker" were removed as these are considered sensitive and are prohibited in many countries legislations as they could induce unlawful discrimination. It is worth mentioning that it was decided to keep the attribute "job" to maintain the accuracy of the prediction model even though this attribute contains some categories that provide information about the residence status of the loan applicant. Additionally, the original attributes' names and descriptions were modified to improve end-users' understandability. The attributes used and their descriptions are presented in Appendix A1.

To create the predictive model used in this study, an exploratory data analysis was performed using Jupyter notebooks and the programming language Python. This analysis observed that the dataset suffers from a class imbalance, with 70% of loan applications approved and only 30% rejected. This class imbalance can result in a bias towards a majority class in the predictive model. On the one hand, tackling class imbalance in the training dataset has been shown to improve the predictive model's performance and generalizability (Chawla et al., 2002). On the other hand, it can also affect how the attributes influence the model's predictions and the explanations generated for them. Thus, it was decided to evaluate three approaches to tackle this class imbalance to improve the model's performance despite the influence it can have on the explanations shown to end-users. The three evaluated approaches were Balance Class Weights (J. M. Johnson & Khoshgoftaar, 2019) using the "compute_sample_weight" functionality of the scikit-learn library,[14] as well as Random Oversampling (ROS) (Yap et al., 2014) and Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002) from the imbalanced-learn library.[15] The default parameters provided by the libraries were utilized for the three approaches.

Using the popular Python deep learning libraries Keras[16] and TensorFlow,[17] a grid parameter search was performed using cross-validation for a neural network model and the approaches to tackle imbalance to determine the best hyperparameters and architecture. The resulting model with the highest score was a neural network with 2-hidden layers, each with 65 and 33 neurons, and the SMOTE approach. Table 3 illustrates the predictive model's performance metrics.

[14] https://scikit-learn.org/
[15] https://imbalanced-learn.org/stable/
[16] https://keras.io/
[17] https://www.tensorflow.org/

Furthermore, a clustering approach was required in the evaluation of the third design iteration with end-users to control for the generalizability of explanations from one of the evaluated methods. Details regarding the reasoning for this implementation are presented in Section 3.4.3.1. Clustering algorithms were evaluated, and it was decided to implement a k-medoids approach as it led to the highest variability and, thus, the most meaningful clusters concerning the predicted class (Kaufman & Rousseeuw, 2009). The Gower distance was used to calculate the distance between the points, as it can handle numerical and categorical variables (Gower, 1971). As a result, the instances of the dataset were divided into 13 cluster groups. For the evaluation of the third design iteration, similar bank loan applications were selected from the cluster with the highest average prediction score, which means that the loan applications were more likely to be rejected.

Table 3: Precision, recall, F1-score, and accuracy of the neural network predictive model.

	Precision	**Recall**	**F1-score**
Approved	0.85	0.81	0.83
Rejected	0.60	0.67	0.63
Macro average	0.72	0.74	0.73
Weighted average	0.78	0.77	0.77
Accuracy			0.77

3.3.2 Selected Local Model-agnostic Methods

To select the XAI methods to be evaluated in this study, existing methods proposed in the literature were analyzed considering a set of criteria. Only methods implemented as open-source Python packages were considered to integrate them with the predictive model. Additionally, their relevance in the literature and among practitioners was evaluated. Finally, the type of explanation these methods provide was analyzed to integrate a diverse set of approaches to this study. On this basis, the selected methods are (1) LIME, (2) Anchors, (3) SHAP, and (4) DICE. In particular, LIME and SHAP were selected since they are popular and widely implemented in research and practice, as exemplified by toolkits such as AIX360 (Arya et al., 2019) or InterpretML (Nori et al., 2019). Anchors was chosen because it is supposed to provide explanations that are easy for end-users to understand (Ribeiro et al., 2018). Lastly, DICE was selected due to the solid theoretical foundations for counterfactual explanations in psychological and philosophical literature. Moreover, counterfactuals are believed to be comprehensible and human-friendly (Miller, 2019). Section 2.4 provides an overview of these methods by referring to the original studies and presents the out-of-the-box representations for providing local explanations from each library.

3.3.3 Iterative Evaluation Process

To investigate how end-users perceive, evaluate, and visually attend to representations of different local explanations from model-agnostic XAI methods, this study relied on a design process that iteratively refined the representations and evaluated them with end-users. Then these refined representations were evaluated with end-users in a laboratory experiment using eye-tracking technology. Throughout the research journey in this study, the experimental design and evaluation measures were adjusted according to the focus of each evaluation. This section presents an overview of this evolution and the underlying analysis strategy.

3.3.3.1 Evaluation Design

This study relied on a between-subjects experimental design to conduct three online evaluations (one in each iteration of the design process), with participants randomly assigned to one of four groups corresponding to the explanation representations for Anchors, DICE, LIME, and SHAP. These three evaluations were conducted online due to the COVID-19 pandemic. Each of these online evaluations consisted of four phases. First, participants were introduced to the bank loan application scenario, a description of the loan applications' attributes, and information on how the AI system was trained and how it provides decision recommendations to approve or reject loan applications. Additionally, they were required to answer attention-check questions to verify their understanding. Second, participants were presented with a description of the corresponding explanation representation and examples of the explanation for an approved and a rejected loan application. Third, participants performed a forward-prediction task divided into two stages, training, and testing (Doshi-Velez & Kim, 2017). During training, participants were presented, in the same order, a set of eight loan applications (four approved and four rejected), each with the model's decision and corresponding explanation. In the testing stage, participants were asked to guess the model's prediction for eight new loan applications that displayed the application's attributes but no explanation (four approved and four rejected). The same set of loan applications was presented to participants in the same order. Fourth, participants were asked to respond to questions regarding their demographics and their evaluation of explanations in the form of short interviews or self-reported measures.

In contrast to the experimental design for the iterative design process, for the eye-tracking laboratory experiment, this study relied on a within-subjects design, showing each participant one approved and one rejected loan application with the explanation representations of each evaluated XAI method (Anchors, DICE, LIME, and SHAP). This experimental design allowed to understand better how participants evaluated and utilized each type of explanation representation and which one they preferred. Moreover, participants' visual attention during their interaction with these explanation representations was tracked using a Tobii Eye Tracker 4C configured with a frequency of 90 Hz and its corresponding relevant research license for recording and analyzing data. In addition, a forward-prediction task was

not performed in this laboratory experiment. Thus, the laboratory experiment consisted of four phases. First, similarly to the online evaluations, participants were introduced to the bank loan scenario, a description of the loan applications' attributes, and information on how the AI system was trained and how it provides decision recommendations to approve or reject loan applications. Second, in a randomized order, participants were shown a description of one of the explanation representations followed by an approved and a rejected loan application together with that explanation representation. This process was repeated for each explanation representation. Thus, participants received a total of two explanation representations of each of the four XAI methods evaluated. Third, participants were asked to respond to questions regarding their demographics and their evaluation of explanations using self-reported measures. Fourth, semi-structured interviews were conducted with participants (see Appendix A8).

3.3.3.2 Evaluation Measures

The evaluation measures were adjusted throughout the iterative design process and the laboratory experiment according to the focus of the evaluation. Table 4 presents a summary indicating which measures were used in each user evaluation round. For more details on the evaluation measures, see Appendix A2.

Table 4: Matrix of evaluation measures used in each evaluation round conducted with end-users.

Category	**Measure**	**Iteration of Design Process (Between-subjects)**			**Laboratory Experiment (Within-subjects)**
		1st	**2nd**	**3rd**	
Performance	Forward-prediction Score		X	X	
Self-report	Rank				X
	Satisfaction (Hoffman et al., 2018)			X	X
	Trust (Hoffman et al., 2018)		X	X	
	Understandability (Madsen & Gregor, 2000)		X	X	
	Usefulness				X
Behavioral	Fixation duration				X
	Number of fixations				X

The number of correct guesses of the model's prediction each participant made during the forward-prediction task was used as an objective performance measure. This forward-prediction score has been commonly used in research as a means to investigate the quality of explanations (Buçinca et al., 2020; Cheng et al., 2019; Hase & Bansal, 2020; Poursabzi-Sangdeh et al., 2021; Ribeiro et al., 2018). The idea behind this measure is that participants first build a mental model of how the ML predictive model makes decisions by observing the explanations for those decisions in a training phase. Afterward, they apply this mental model to estimate the predictive model's decision on new instances in a testing phase.

Subjective self-reported measures of constructs commonly utilized in XAI research were also incorporated. These self-reported measures are collected using Likert scales to capture participants'

agreement with a series of statements representing each construct. The evaluations performed in this study relied on the following constructs established in the literature: (1) satisfaction (Hoffman et al., 2018), (2) trust (Hoffman et al., 2018), and (3) understandability (Madsen & Gregor, 2000). Custom self-reported scales were also utilized to measure participants' perceived usefulness of each explanation representation and their rank according to their preference. In the evaluations performed in the iterative design process, the constructs' reliability and validity were examined by performing confirmatory factor analyses (CFA) to ensure good model properties. Self-reported control variables were also used in the laboratory experiment.

Finally, participants' eye movement data were integrated into the laboratory experiment to derive two behavioral measures commonly used in eye-tracking research (Lorigo et al., 2008; Muddamsetty et al., 2022; Polley et al., 2021), fixation duration and the number of fixations. Similarly to Polley et al. (2021), these measures were used to investigate which regions of the explanation representations received more attention from end-users. Specifically, the number of fixations and fixation duration between the explanation representations were compared, and heatmaps to analyze end-users' visual attention focus were generated. In addition, these behavioral measures were combined with self-reports and interviews to better understand how end-users utilize and perceive the refined representations from the iterative design process.

3.3.3.3 Analysis Strategy

Throughout the evaluations of the iterative design process and in the laboratory experiment, a series of tests were performed to investigate if there were any statistically significant differences across the performance, self-reported, and behavioral measures. SPSS was used to perform these statistical analyses.[18]

In the between-subjects online evaluations performed in the iterative design process, participants evaluated one of the four explanation representations. For these evaluations, a combination of self-reported and performance measures was utilized. While the self-reported measures were collected using a 7-point Likert scale, the forward-prediction score ranged from zero to eight. Therefore, these measures were analyzed separately in a multivariate analysis for self-reported measures and univariate analysis for the forward-prediction score. Nevertheless, it was found that the parametric one-way MANOVA and one-way ANOVA assumptions were not met. So instead, each self-reported measure and the forward-prediction score were analyzed with the nonparametric test Kruskal-Wallis provided by the "NPTESTS" function in SPSS. All post-hoc pairwise comparisons for the different explanation representations were performed automatically by SPSS using a Bonferroni correction.

[18] https://www.ibm.com/products/spss-statistics

In contrast, participants evaluated each explanation representation in the within-subjects laboratory experiment. Self-reported and eye-tracking measures and self-reported control variables were used for this evaluation. Multiple measurements were created in related groups across the within factors for each of the self-reported and eye-tracking measures. Due to differences in measures' scales, independent repeated-measures univariate analyses were conducted with each measure as a dependent variable and the related groups as the within-subjects factor. Additionally, control variables were incorporated as covariates when supported by the statistical model. First, a check was performed for common assumptions necessary in parametric tests. These included no significant outliers, normality, and sphericity. Depending on the number of within factors, a one-way repeated measures ANCOVA or a two-way repeated measures ANOVA were utilized when the necessary assumptions were met. These tests were conducted using the SPSS General Linear Models (GLM) function, which automatically performed all necessary post-hoc pairwise comparisons with a Bonferroni correction. Nonparametric Friedman tests were used for each measure in case one or more assumptions were violated. When statistically significant differences were found for the within-subjects factors, a follow-up analysis with a pairwise Wilcoxon signed-rank test with a Holm-Bonferroni correction was conducted (Holm, 1979).

3.4 Results of the Iterative Design Process

To increase the comparability of the out-of-the-box explanation representations from the selected methods and control for any confounding factors due to their different explanation approaches, these explanation representations were refined in an iterative design process involving end-users. This was achieved following the strategy Iivari (2015) proposed, which aims to provide a new solution to a general problem identified by researchers. In this strategy, although researchers may be informed about some specific problems, they face uncertainties regarding the most appropriate general solution. As a result, they must identify potential end-users to develop and evaluate conceptual artifacts.

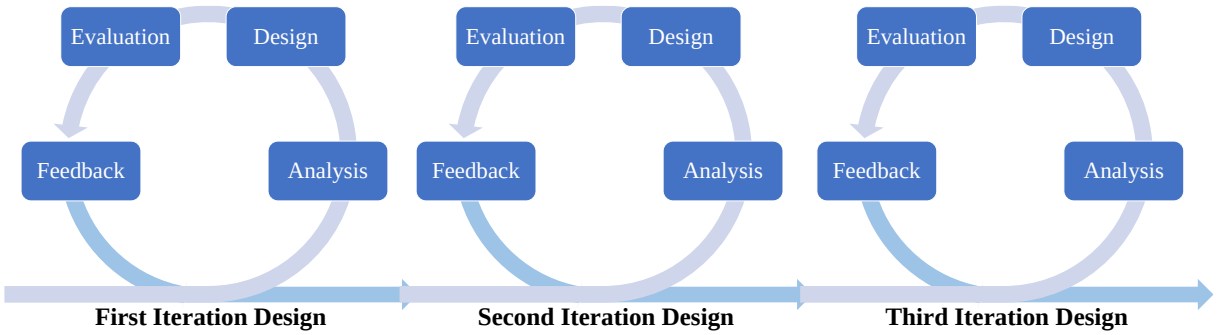


Figure 7: Iterative design process.

In this study, the iterative design process, which adhered to the methodology presented in Section 3.3, provided comparable representations after three iterations (see Figure 7). The following sections

describe the iterative refinements and evaluations from the design process. A summary of the iterative design process with the most relevant information is also provided.

3.4.1 First Design Iteration

First, the comparability of the explanation representations presented in Section 2.4 was analyzed. This analysis showed that an evaluation using these representations presents many challenges due to the differences in the amount of information presented, color coding, or layout, which could introduce additional confounding factors in the analysis. To tackle these issues, a design workshop took place with three data scientists from an information technology (IT) company and three human-computer interaction (HCI) researchers. This workshop analyzed each method's approach to generating explanations and the resulting representations. Specifically, the representation style, amount of information, color coding, and terminology were examined. The overlaps these explanation representations had in the mentioned criteria were discussed, and design modifications that would increase their comparability while reducing potential confounding factors due to their differences were proposed.

3.4.1.1 Representation Refinement

LIME's refined representation was based on a simplified version of the matplotlib bar plot provided by its library. A dedicated bar plot area allows to separate the attributes' details from the bars by placing them outside the plot on the y-axis to increase their readability. The class labels (i.e., approved and rejected) at the top were maintained so end-users could identify each attribute's influence towards a class. Each attribute's influence value next to each bar was also maintained to increase their comparability. LIME's standard color coding was replaced with SHAP's, highlighting attributes contributing to approval in blue and rejection in red. Figure 8 shows the first design for LIME's explanation representation.

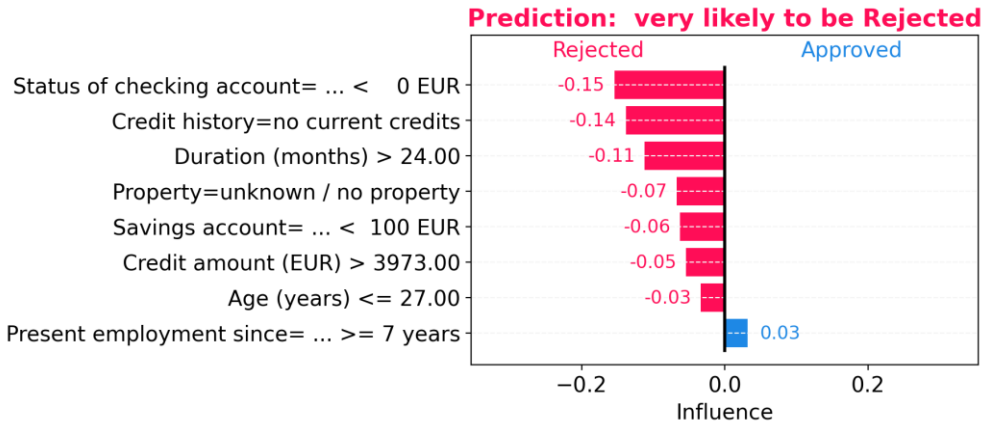


Figure 8: First design of LIME's explanation representation.

The same bar plot was used as a baseline to refine SHAP's representation. The stacked bars representing the influence of all attributes at the top of the plot from the original representation were maintained. Like LIME's representation, the bars and corresponding influence values were placed inside the plot to improve readability, and the attributes' details on the y-axis. The base and output values from the original representation were maintained, but their labels were replaced with "Base Probability" and "Decision Probability" correspondingly. In contrast to LIME, where the explanation representation for "Approved" or "Rejected" classes is the same, for SHAP, the explanation representation explicitly considers only one of the two classes. For the binary classification task in the selected scenario, the base and output values are different for explanations of each class, as they complement each other. To avoid providing an explanation with a decision probability below 0.5, the representation for each class was presented according to the model's decision. However, the color coding in SHAP's original representation is bound to the increment/decrement of the model's output score, which would present a contradictory color coding for the "positive" or "negative" contribution toward approved or rejected. To address this issue, there were two options, invert the scale of the x-axis on the plot to reverse the increment/decrement of the probability for one of the classes or invert the color-coding for the increment/decrement according to the model's decision. It was considered that the direction of the increment in the x-axis was a higher constraint. Thus, different color coding for the increment/decrement for each represented class was utilized. The first design for SHAP's explanation representation is shown in Figure 9.

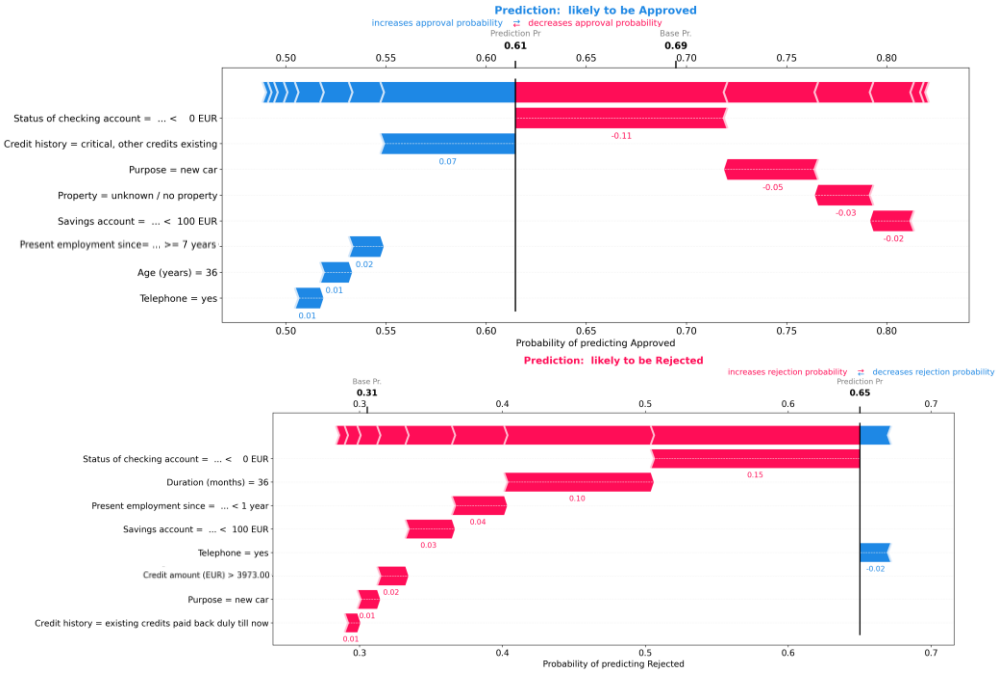


Figure 9: First design of SHAP's explanation representations for approved and rejected loan applications.

For Anchors and DICE, the possibility of presenting rules and counterfactuals similar to LIME and SHAP was evaluated during the design workshop. Specifically, the possibility of distributing the

influence of attributes contained in Anchors' rules and DICE counterfactuals and showing them on a bar plot with an equal magnitude was analyzed. Nevertheless, after careful analysis, it was decided that this representation could lead end-users to interpret that each attribute's influence is independent even though the model's decision is explained as a result of the exact combination of all attributes' values shown in the rules and counterfactuals. Therefore, a table was used to present the explanation representations of both methods.

The proposed explanation representation for Anchors consisted of a table containing the rules on the left and the model's prediction on the right. A header with the text "if" was presented on top of the rules to indicate that all the conditions need to be fulfilled. Each rule was then placed on an individual row. The prediction was presented on the right part of the table in a merged cell that extended across all rules with the header "Predict". The first design for Anchors' explanation representation is shown in Figure 10.

If	Predict
Duration (months) <= 18.00	Very likely to be Approved
Job = skilled	
Registered telephone = yes	

Figure 10: First design of Anchors' explanation representation.

For DICE, its library lists possible counterfactuals that would lead to the alternative decision as to the model's prediction. These counterfactuals are sorted by their distance to the original instance explained by the algorithm. It was decided to provide only the first counterfactual of the list, as it represents the scenario in which the model would predict the alternative class with the minimum number of attribute changes. Similar to Anchors, DICE counterfactuals were presented in a table. This table consisted of rows describing the necessary changes for each attribute in four columns: "Action", which showed "changing" for categorical attributes and "increase" or "decrease" for numerical attributes; "Attribute" with the name of the attribute; "Original Value" with the attribute's value of the loan application; and "Modified Value" with the attribute's value needed to be modified to obtain the alternative prediction. Figure 11 shows the first design for DICE's explanation representation.

The decision recommendation is "Very likely to be Approved"

The decision recommendation would be "Somewhat likely to be Rejected", if ALL the following attributes were modified:

Action	Attribute	Original Value	Modified Value
Changing	Purpose	used car	television
Changing	Housing	own	rent
Changing	Most valuable property	car	life insurance

Figure 11: First design of DICE's explanation representation.

In addition to the individual modifications to each explanation representation, the model's prediction probability in a text form was presented as proposed by Cheng et al. (2019). The motivation to show the model's probability to end-users is to communicate the model's confidence in each decision

recommendation. Thus, together with the model's decision, the text "somewhat likely to be" or "very likely to be" was presented depending on the prediction's probability.

3.4.1.2 Evaluation and Analysis

To evaluate the explanation representation designs proposed in the first design iteration, nine university graduates were invited to participate voluntarily in an online evaluation followed by a brief interview, as they could potentially apply for bank loans. As detailed in Section 3.3.3, in a between-subjects experimental design, participants evaluated one of the four explanation representations from the selected model-agnostic methods in a forward-prediction task. After the evaluation, participants were asked to compare the explanation representation they evaluated against the ones they were not shown.

Participants generally perceived the explanation representations as a useful help to understand why the AI system provided a given decision recommendation. All participants seemed to understand the basic notion of the explanation representation they evaluated and provided some feedback on how the loan application attributes were presented. Four participants recommended standardizing the use of color coding to highlight the model's decision recommendation for all explanation representations. Additionally, five participants recommended simplifying the model's decision recommendation with probability in text form for LIME and SHAP by removing the word "Prediction". Additional feedback regarding the SHAP representation was to rename the x-axis from "Probability of prediction Approved/Rejected" to "Probability of Approval/Rejection". Moreover, multiple participants provided feedback to make minor adjustments to the explanation representations layout for improvement. Finally, three participants brought to attention that showing eight attributes in LIME and SHAP explanation representations could not be a fair comparison considering the number of attributes shown in Anchors' rules and DICE's counterfactuals.

3.4.2 Second Design Iteration

Based on the first evaluation results, a second design workshop was conducted with two data scientists from an IT company and three HCI researchers. During the workshop, the feedback provided by participants in the first design iteration was analyzed, and possible modifications to the explanation representations were evaluated to improve their comparability further.

3.4.2.1 Representation Refinement

Concretely, the following modifications to the explanation representations were performed: (1) A standard color coding to the model's decision recommendation was implemented in all explanation representations. (2) Next, the probability in text form was simplified to remove the word "Prediction". (3) For Anchors, the text in the table's headers was improved to support the interpretation that all rules must be fulfilled for the model to provide the indicated decision recommendation. (4) Minor adjustments

were made to the axis labels and explanation representations' layout. And (5) moreover, the number of attributes shown in each explanation representation type was analyzed for all explanations in the dataset. For LIME and SHAP, it is possible to control how many attributes are shown in the explanation through configuration parameters. On the other hand, Anchors and DICE explanations introduce variability in the number of attributes shown in their explanations that depends on the instance being explained. It was found that Anchors' explanations had an average of 3.34 features (SD = 3.20), while DICE's explanations had an average of 2.96 features (SD = 1.56). It was decided to control the number of features shown in explanation representations to account for this. For LIME and SHAP, the number of features shown was limited to five. For Anchors and DICE, instances from the dataset that contained between three and five features were selected. Figure 12 shows the second design of explanation representations for Anchors, DICE, LIME, and SHAP.

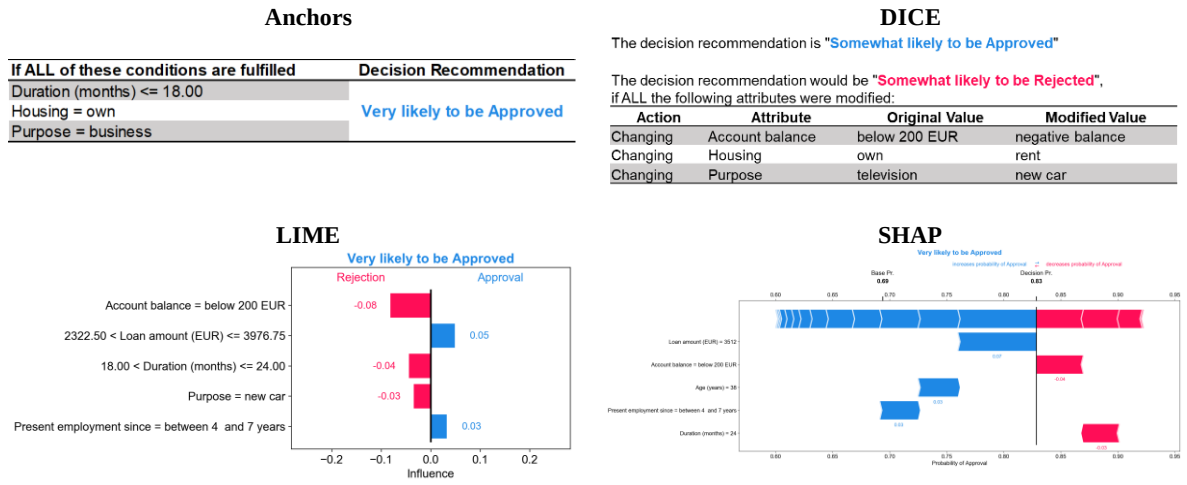


Figure 12: Second design of explanation representations for Anchors, DICE, LIME, and SHAP.

3.4.2.2 Evaluation and Analysis

An online evaluation to evaluate the explanation representations designs was conducted from the second iteration. As detailed in Section 3.3.3, in a between-subjects experimental design, participants evaluated one of the four explanation representations from the selected model-agnostic methods in a forward-prediction task. After the evaluation, participants were asked to answer a questionnaire to collect their demographic information and measure their perceived understandability and trust (see Appendix A2 for details). Details of the statistical analyses conducted are included in Appendix A3.

A total of 258 participants were recruited for the evaluation from a crowdsourcing website, as they could potentially apply for bank loans. The average time to complete the online evaluation was 28.82 minutes (SD = 10.78), and the average payment per hour was $9.72. Data from 23 participants were removed as multiple instances were found where answers provided were the same word by word, which indicated that the answers could be from the same participants completing the experiment in parallel from different

accounts. The final sample included 235 participants (Anchors = 58, DICE = 60, LIME = 60, SHAP = 57).

Table 5 shows the descriptive statistics of participants' perceived trust and understandability in the system and their forward-prediction score. Following the analysis strategy described in Section 3.3.3.3, it was found that the normality assumption was violated for the multivariate analysis of trust and understandability and the univariate analysis of the forward-prediction score. Therefore, independent Kruskal-Wallis tests with trust, understandability, and forward-prediction score as dependent variables and the explanation type as the independent variable were conducted. The analyses indicate that participants perceive all explanation representations similarly, as there is no statistically significant difference between groups for trust ($\chi2(3) = 1.751$, $p = 0.626$) or understandability ($\chi2(3) = 0.805$, $p = 0.848$). The analysis for the forward-prediction score indicates a significant difference between groups ($\chi^2(3) = 9.963$, $p = 0.019$). Post-hoc pairwise comparisons with Bonferroni correction reveal a significantly larger score for Anchors than SHAP ($p = 0.010$). It seems that the clear coverage of the Anchors' rules allowed participants in the Anchors group to generalize the explanations observed during the forward-prediction task to achieve a higher score than participants in the SHAP group.

Table 5: Descriptive statistics for the evaluation of the second iteration design.

	Group Means (SD)				
Variable	**Anchors**	**DICE**	**LIME**	**SHAP**	**Total**
Trust	5.04	5.07	5.03	5.23	5.09
	(1.38)	(1.12)	(1.35)	(1.25)	(1.27)
Understandability	4.92	5.20	4.99	5.17	5.07
	(0.44)	(1.04)	(1.32)	(1.24)	(1.26)
Forward-prediction score	4.31	3.95	3.88	3.52	3.92
	(1.26)	(1.27)	(1.39)	(1.40)	(1.35)

Through the provided open-text field, valuable feedback from participants was obtained. Several participants indicated that the explanation representations were useful, interesting, and helpful to know more about how the system makes decisions. Multiple comments were received that the text showing the model's probability in the form "very likely" or "somewhat likely" was confusing, and its meaning was not clear enough. For LIME, two participants highlighted that the interpretation of the influence values shown was unclear, and they questioned whether these values should be interpreted as a percentage. For SHAP, three participants commented that the attributes with a positive and negative influence were not always on the same side of the plot. This change in position made interpreting the explanation among all eight loan applications difficult, as they would have to change back and forth the direction of the influence. For LIME and SHAP explanation representations, several participants expressed their desire to see all attributes influencing the decision instead of only the five most relevant.

3.4.3 Third Design Iteration

Based on the second evaluation results, a third design workshop was conducted with two data scientists from an IT company and three HCI researchers. It was observed that there seemed to be good progress in improving the comparability of the explanation representations, considering how these were perceived similarly by participants in the different groups. Nevertheless, further potential improvements were identified from the feedback participants provided. Therefore, further modifications to the explanation representations were performed. Details of the statistical analyses conducted are included in Appendix A3.

3.4.3.1 Representation Refinement

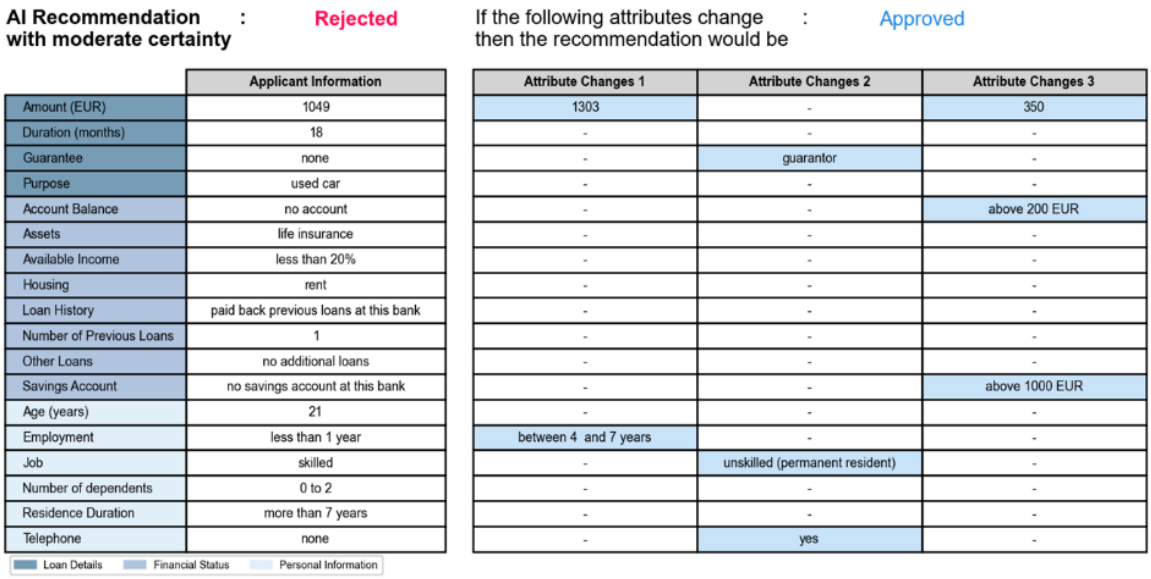
Since participants found the text showing the model's probability confusing, it was replaced with a text indicating the model's certainty in the decision recommendation as moderate or high. Considering participants' requests to see how all attributes influence the decision, an evaluation was performed on whether to increase the number of attributes shown for LIME and SHAP explanation representations and how it could affect the comparability with the other XAI methods. As mentioned in Section 3.4.2.1, controlling how many attributes are shown in the explanation through configuration parameters for LIME and SHAP is possible. With this in mind, an analysis of whether it was possible to increase the amount of information on Anchors and DICE explanation representations through a new design was performed.

For Anchors, the number of attributes depends on the rules generated by the algorithm, which can only be implicitly influenced by adapting the precision threshold of the algorithm. By changing the precision threshold, the number of features shown in the explanation could be increased. However, this could result in explanations that have specific rules with high complexity and lower coverage. An alternative would be to run the explanation algorithm multiple times to generate multiple sets of rules. Nevertheless, these different sets of rules could provide conflicting explanations. As Anchors' rules are interpreted as conditions that need to be fulfilled and would result in a particular decision, showing more than one set of rules could lead to different conditions that need to be fulfilled. On this basis, the number of attributes shown in Anchors' explanations was not modified.

For DICE counterfactuals, the layout was refined to present more than one counterfactual. Furthermore, the distribution of the number of attributes shown in DICE's counterfactuals was analyzed. It was found that, on average, roughly seven features are present when considering three counterfactual examples. Thus, DICE's explanation representation was refined to show three counterfactuals. DICE's new design was based on its original representation design. The original table was transposed, and a long table format was implemented in which each row represents one attribute of the bank loan application. Meanwhile, the columns represent the attributes' changes for the counterfactuals. Then, the cells that

contain attributes' changes were highlighted using the same color-coding as LIME and SHAP according to the alternative decision that the counterfactual changes would lead to. Additionally, a table containing loan application attributes' values was placed at the left of the explanation representation as a reference. This reference would allow end-users to relate the attributes' changes in the counterfactuals to the current values. The attributes were grouped into loan details, financial status, and personal information for the loan application details table. For each group, a different color code and alphabetical ordering were used. The third design of DICE's explanation representation is shown in Figure 13.

The loan application attributes table was integrated for all methods at the left of the explanation representation. To adhere to the number of attributes shown in DICE's new design, the number of attributes shown in the explanation representation was increased for LIME and SHAP to seven. In contrast, the number of attributes shown in Anchors was not controlled.



	Applicant Information	Attribute Changes 1	Attribute Changes 2	Attribute Changes 3
Amount (EUR)	1049	1303	-	350
Duration (months)	18	-	-	-
Guarantee	none	-	guarantor	-
Purpose	used car	-	-	-
Account Balance	no account	-	-	above 200 EUR
Assets	life insurance	-	-	-
Available Income	less than 20%	-	-	-
Housing	rent	-	-	-
Loan History	paid back previous loans at this bank	-	-	-
Number of Previous Loans	1	-	-	-
Other Loans	no additional loans	-	-	-
Savings Account	no savings account at this bank	-	-	above 1000 EUR
Age (years)	21	-	-	-
Employment	less than 1 year	between 4 and 7 years	-	-
Job	skilled	-	unskilled (permanent resident)	-
Number of dependents	0 to 2	-	-	-
Residence Duration	more than 7 years	-	-	-
Telephone	none	-	yes	-

Figure 13: Third design of DICE's explanation representations.

Moreover, Anchors' design was further simplified, and the text indicating that all rules must be fulfilled was placed at the top of the explanation representation with the rules below. Additionally, each attribute's influence value was recalculated as a percentage of the total influence from all attributes for LIME's explanations and was presented as a percentage instead. Lastly, for SHAP, it was considered how changing the position of the positive and negative attributes in the plot could increase participants' mental effort. To avoid this, a plot with the probability of rejection was always presented, even if this probability was below 0.50. Figure 14 shows the third design of explanation representations for Anchors, LIME, and SHAP but omits the loan application attributes table as it has already been shown for DICE in Figure 13.

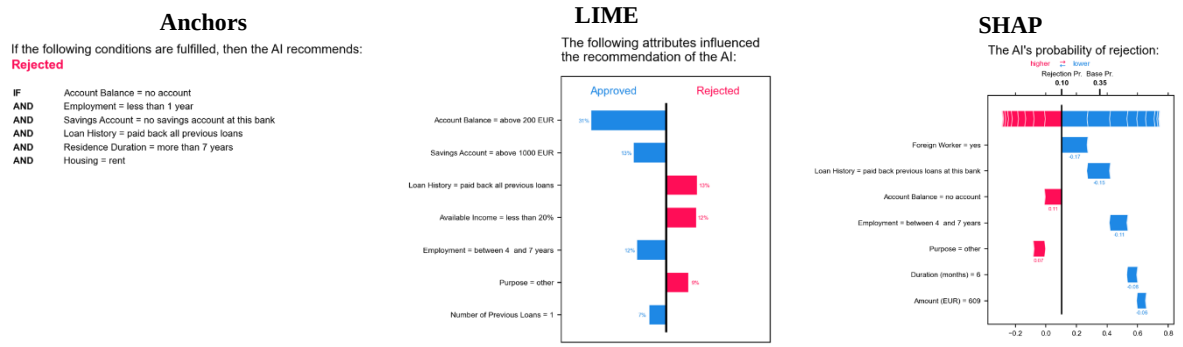


Figure 14: Third design of explanation representations for Anchors, LIME, and SHAP.

3.4.3.2 Evaluation and Analysis

As detailed in Section 3.3.3, an online evaluation in a between-subjects experimental design was conducted, with participants evaluating one of the four explanation representations from the selected model-agnostic methods in a forward-prediction task. Moreover, the generalizability of Anchors' explanations, visible in the higher forward-prediction score participants in the Anchors group achieved in the evaluation of the second iteration, was controlled. To achieve this, the clustering algorithm presented in Section 3.3.1 was implemented. This clustering approach should ensure that the local explanations seen by participants in the training step of the forward-prediction task can help them, to a certain degree, extrapolate to the test step. Thus, similar bank loan applications were selected from the cluster with the highest average prediction score. In other words, the selected loan applications for the forward-prediction task were more likely to be rejected by the predictive model.

After the forward-prediction task, participants were asked to answer a questionnaire to collect their demographic information and measure their perceived satisfaction, trust, and understandability (see Appendix A2 for details). A total of 261 participants were recruited for the evaluation from a crowdsourcing website, as they could potentially apply for bank loans. The average time to complete the online evaluation was 24.25 minutes (SD = 10.25), and the average payment per hour was $8.70. Data from nine participants who failed attention checks during the evaluation was removed. The final sample included 252 participants (Anchors = 66, DICE = 63, LIME = 62, SHAP = 61).

Table 6 shows the descriptive statistics of the evaluation measures. Following the analysis strategy described in Section 3.3.3.3, it was found that the normality assumption was violated for the multivariate analysis of satisfaction, trust, and understandability and the univariate analysis of the forward-prediction score. Therefore, independent Kruskal-Wallis tests were conducted for each measure. The analysis indicates that participants perceived all explanation representations similarly, as there was no statistically significant difference between groups for satisfaction ($\chi^2(3) = 1.415$, $p = 0.702$), trust ($\chi^2(3) = 1.897$, $p = 0.594$) or understandability ($\chi^2(3) = 1.381$, $p = 0.710$). In contrast to the evaluation of the second iteration, participants' trust and understandability were lower. These lower values could be

explained by the selection of loan applications with a higher probability of rejection. The forward-prediction score analysis also indicates no significant difference between groups ($\chi^2(3) = 6.056$, $p = 0.109$), which confirmed that using the clustering algorithm to select similar loan applications helped counterbalance the generalization of Anchors' explanations.

In contrast to the evaluations in the first and second design iterations, no feedback was received regarding the design of the explanation representations from participants. On the contrary, several participants indicated that they found the explanation representations understandable and well-designed.

Table 6: Descriptive statistics for the evaluation of the third iteration design.

	Group Means (SD)				
Variable	**Anchors**	**DICE**	**LIME**	**SHAP**	**Total**
Trust	3.72	3.76	3.58	3.80	3.71
	(1.22)	(1.29)	(1.31)	(1.15)	(1.24)
Understandability	4.08	3.94	4.19	3.89	4.03
	(1.36)	(1.40)	(1.35)	(1.38)	(1.37)
Satisfaction	4.19	4.29	4.37	4.48	4.33
	(1.38)	(1.46)	(1.41)	(1.38)	(1.40)
Forward-prediction score	5.73	5.39	5.90	5.54	5.64
	(1.14)	(1.38)	(1.14)	(1.34)	(1.26)

3.4.4 Summary of the Iterative Design Process

Throughout the iterative design process, the out-of-the-box explanation representations from Anchors, LIME, DICE, and SHAP were refined to increase their comparability and control for confounding factors that their original representations could induce. In the second and third online evaluations, similar levels of perceived satisfaction, trust, and understandability were observed among the groups. These results indicate that all explanation representations can, to a certain degree, help end-users understand how the AI system makes decisions. Nevertheless, since each participant evaluated only one of the four explanation representations in all evaluations, it was unclear which explanation type they would prefer and how they would utilize them.

Considering the results obtained in the third design iteration, it was decided to terminate the design process and utilize the third explanation representations' designs shown in Figure 13 and Figure 14 for the laboratory experiment using eye-tracking technology presented in the following section.

3.5 Eye-Tracking Laboratory Experiment

As detailed in Section 3.3.3, the iteratively refined representations of local model-agnostic explanations for Anchors, DICE, LIME, and SHAP were evaluated in a laboratory experiment with a within-subjects

design incorporating eye-tracking technology. In contrast to the previous between-subjects online evaluations in the iterative design process of this study, this experimental design allowed a better understanding of how satisfied participants were with each type of explanation. Moreover, this experiment investigated how participants visually attend to the information provided by the explanation representations by analyzing their eye-tracking data, providing insights into how they utilize them. After the evaluation, participants were asked to answer a questionnaire to collect their demographic information and other control variables (i.e., domain knowledge, ML knowledge, programming knowledge, gender, and study area). Additionally, participants were asked to rank the explanation representations from most to least preferred. Finally, semi-structured interviews with participants were conducted after the laboratory experiment to discuss their perceptions of the evaluated explanation representations. During these interviews, they were asked to grade the usefulness of each explanation representation. Details of the guide used in the semi-structured interviews can be found in Appendix A8.

The evaluation in this experiment relied on the results of the third design iteration of the explanation representations (see Figure 13 and Figure 14). However, there were some challenges in distinguishing the visual attention on the different elements of the explanation representations due to the accuracy of the eye-tracker. Thus, the explanation representations' layouts were slightly modified by increasing the spacing between elements to differentiate participants' visual attention on them. Specifically, the explanation representations were modified to increase the separation between some of their visual elements. For Anchors, the space between the rules shown was increased. For LIME and SHAP, the number of attributes shown was reduced from seven to six, and the bars' height in the plots was decreased. These modifications can be observed in Figure 18.

Twenty-two participants were recruited from the KD2Lab panel, which consists mainly of students studying in Karlsruhe, Germany (https://www.kd2lab.kit.edu/). Each participant was paid 12.00 € for their participation in the laboratory experiment. The average time of the actual experiment (first stage) was 29.59 (SD = 7.83) minutes, while the average time for the semi-structured interviews (second stage) was 11.71 (SD = 1.89) minutes. Data from three participants was removed due to an incomplete recording of the sessions. Thus, data from 19 participants were analyzed. All participants except one were university students. All details regarding participants' demographic information and the descriptive statistics of control variables are included in Appendix A3.

In the following, the results of the laboratory experiment are presented in three sections (1) analyses of satisfaction, usefulness, and ranks, (2) analyses of participants' eye-tracking data, and (3) results of the semi-structured interviews.

3.5.1 Analyses of Satisfaction, Usefulness, and Rank

Table 7 shows the descriptive statistics of the self-reported measures of the eye-tracking experiment. Following the strategy described in Section 3.3.3.3, repeated-measures univariate analyses were conducted with each measure as the dependent variable and the related groups for each explanation type as a within-subjects factor. When the necessary assumptions were met, one-way repeated measures ANCOVA analyses were conducted and included domain knowledge, ML knowledge, programming knowledge, technical literacy, gender, and age as covariates. In case any assumptions were violated, nonparametric Friedman tests were used. Details of the statistical analyses conducted are included in Appendix A5.

Table 7: Descriptive statistics of the self-reported measures for the eye-tracking experiment.

	Group Means (SD)			
Variable	**Anchors**	**DICE**	**LIME**	**SHAP**
Satisfaction	5.06	4.45	5.03	5.64
	(1.30)	(1.57)	(0.90)	(0.87)
Usefulness	5.08	5.30	6.50	7.08
	(2.29)	(2.85)	(1.81)	(2.08)
Rank *A lower rank is better.	1.84	2.47	2.74	2.95
	(0.77)	(1.26)	(1.10)	(1.08)

After ensuring the necessary assumptions were fulfilled, a one-way repeated measures ANCOVA analysis with satisfaction as a dependent variable was conducted. The results show that participants' satisfaction does not significantly differ across the four explanation representations ($F(3, 36) = 0.382$, $p = 0.766$), and only participants' ML knowledge has a significant influence on their satisfaction ($F(3, 36) = 3.001$, $p = 0.043$). To visualize the effect of ML knowledge on satisfaction, a follow-up analysis with satisfaction as a dependent variable, explanation representation as a within-subjects factor, and ML Knowledge as a between-subjects factor was conducted. Figure 15 shows participants' mean satisfaction for each explanation representation across the four levels of ML knowledge. Even though there are no statistically significant differences in satisfaction across the explanation representations, it is possible to observe how participants' ML knowledge level influences satisfaction. Satisfaction is similar for all explanation representations for participants with lower ML Knowledge. In contrast, there is a higher variance in satisfaction across explanation representations with higher ML knowledge.

For usefulness, the normality assumptions were violated. Thus, a nonparametric Friedman test was conducted with usefulness as the dependent variable. The results show that participants' perceived usefulness marginally differs across the four explanation representations ($\chi^2(3) = 7.190$, $p = 0.066$). Nevertheless, post-hoc pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction reveal no statistically significant differences between the pairwise comparisons.

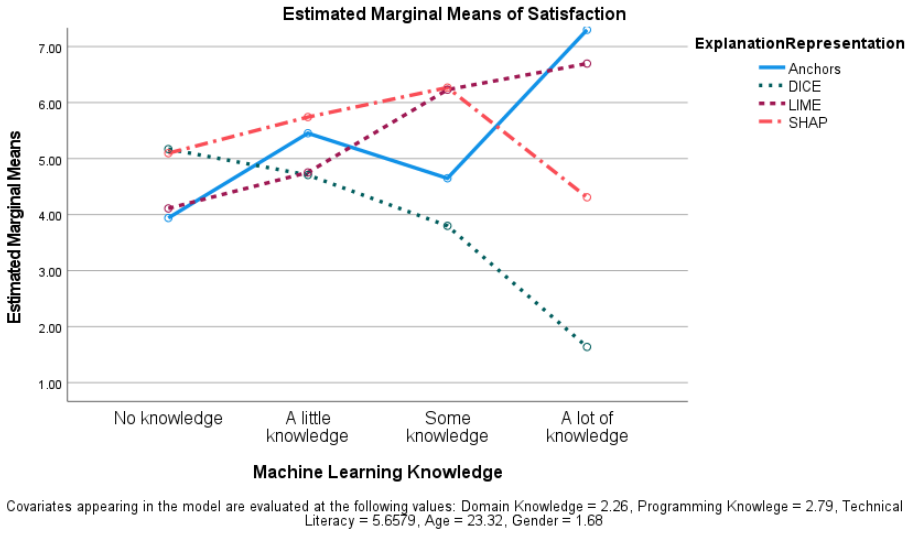


Figure 15: Interaction effect between satisfaction and participants' ML knowledge.

As an analysis with covariates is not possible with a Friedman test, a one-way repeated measures ANCOVA with the control variables as covariates was conducted to analyze if any of them significantly affected usefulness. The results indicate no significant difference in usefulness across the explanation representations ($F(3, 36) = 1.640$, $p = 0.197$). Nevertheless, there was a significant influence of participants' gender on their perceived usefulness ($F(3, 36) = 10.260$, $p < 0.001$). To visualize the effect of gender on usefulness, a follow-up analysis with usefulness as a dependent variable, explanation representation as a within-subjects factor, and gender as a between-subjects factor was conducted. Figure 16 shows the participants' usefulness for each explanation representation for females and males. It can be observed that females have higher perceived usefulness for Anchors and DICE on average. In comparison, males have higher perceived usefulness for LIME and SHAP.

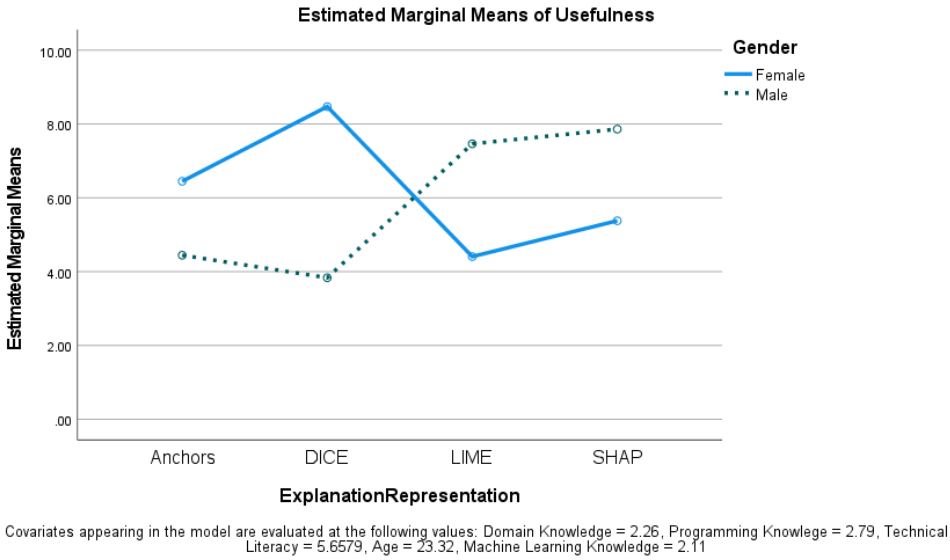


Figure 16: Interaction effect between usefulness and participants' gender (error bars not included for readability).

A nonparametric Friedman test for participants' ranks on explanation representations shows a significant statistical difference between the related groups ($X^2(3) = 7.863$, $p = 0.049$). Nevertheless, post-hoc pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction reveal no statistically significant differences between the pairwise comparisons. Figure 17 shows participants' ranks for explanation representations.

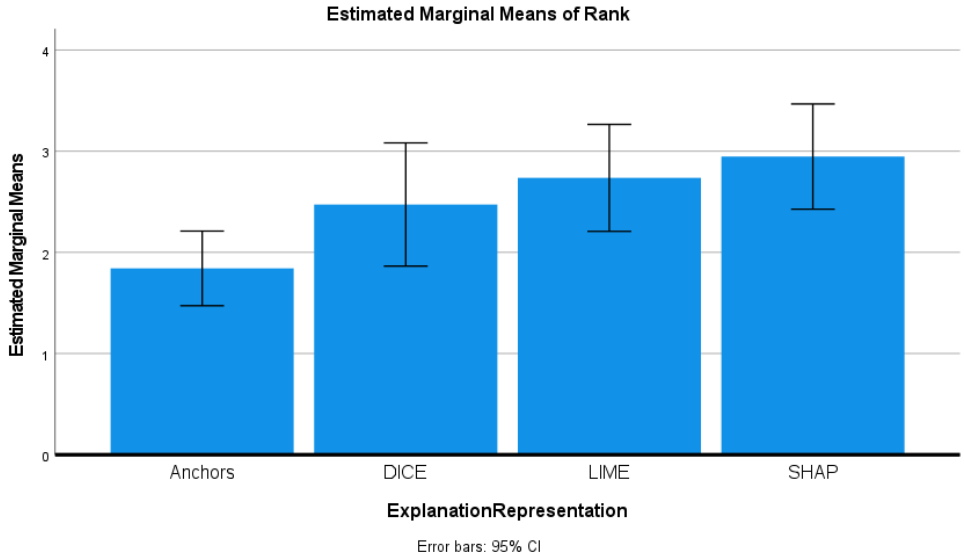


Figure 17: Rank for explanation representations. A higher rank indicates a higher preference.

3.5.2 Analyses of Eye-tracking Data

To analyze participants' visual attention on explanation representations from Anchors, DICE, LIME, and SHAP, the two commonly used eye-tracking measures, fixation duration and number of fixations, were extracted (Lorigo et al., 2008; Muddamsetty et al., 2022; Polley et al., 2021). These measures were utilized to investigate which regions of the explanation representations received more end-user attention (Polley et al., 2021).

As a result, the eight heatmaps shown in Figure 18 were generated by aggregating the fixations across each explanation representation and loan application decision to represent participants' visual attention focus. Users' attention levels are represented using a continuous color scale. Thus, blue stands for low attention, yellow for medium attention, and red for high attention.

Heatmaps provide an excellent visual overview of how participants utilize the explanations during the evaluation task. Moreover, an analysis can be performed to observe which regions of the explanation representations received more visual attention. Nevertheless, it is challenging to identify significant differences in visual attention between the explanation representations from observing and comparing the heatmaps visually.

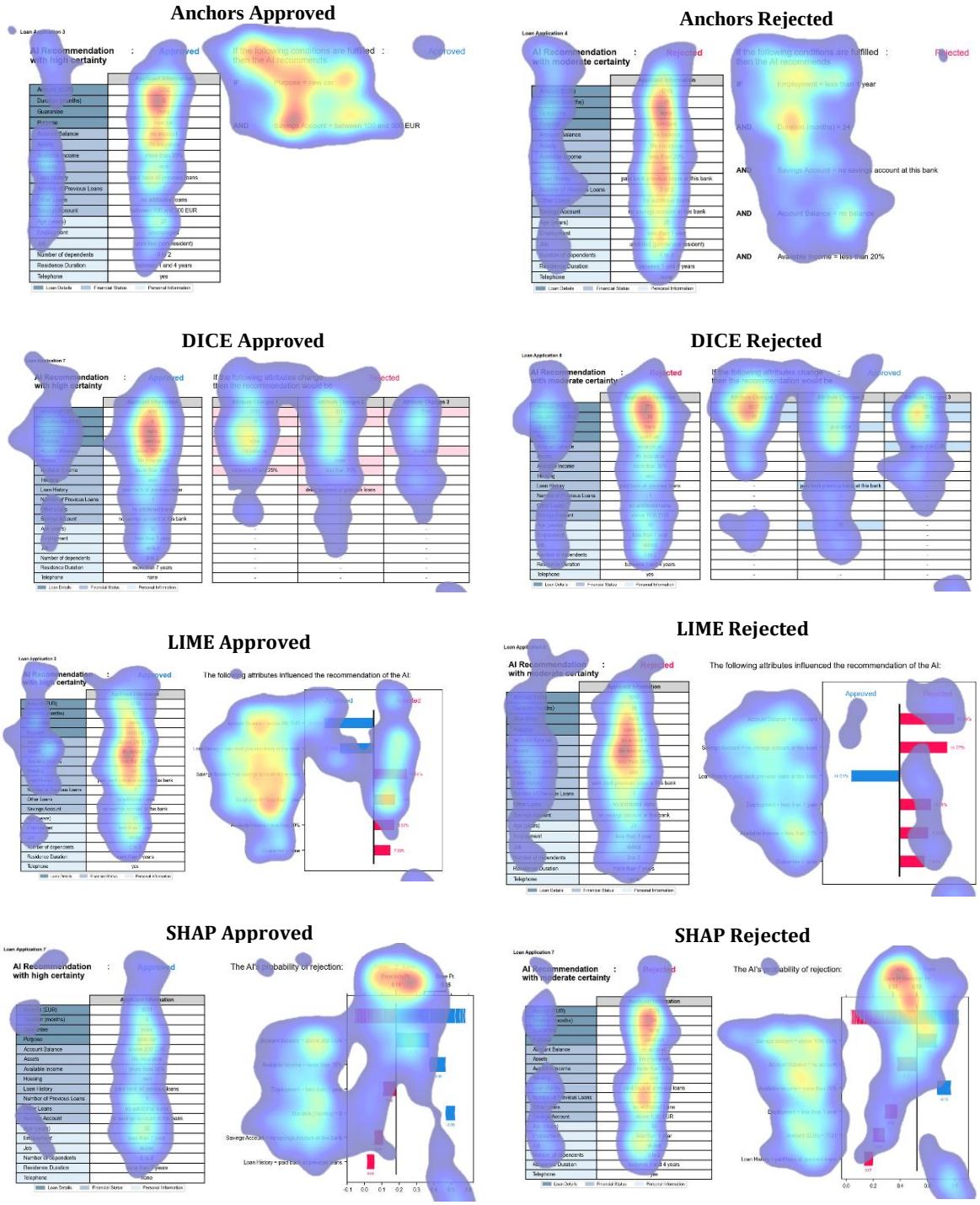


Figure 18: Aggregated heatmaps for explanation representations by loan decision for Anchors, DICE, LIME, and SHAP.

Therefore, following the strategy described in Section 3.3.3.3, statistical analyses were performed to investigate if there were any significant differences in participants' visual attention between the explanation representations. First, AOIs for regions of the explanation representations to be compared were defined. Afterward, repeated-measures univariate analyses with fixation duration or number of fixations on the AOIs as dependent variables and explanation type and loan applications as within factors

were conducted. Nonetheless, considering that both measures provided similar results in almost all the analyses, fixation duration was reported as a primary metric and the number of fixations was reported only when it provided additional interesting findings. Details of the description and visualization of the AOIs are found in Appendix A6.

Participants' visual attention on the AOIs was compared for each explanation type in a three-level top-down approach: (1) complete visualization, including attributes' table (AOI1) and explanation representation (AOI2); (2) explanation representation (AOI2); and (3) specific elements for each explanation type. The following subsections present the results of these analyses and a summary of the findings. Details of the statistical analyses conducted are included in Appendix A7.

3.5.2.1 Complete Visualization

The normality assumptions were violated for fixation duration on the complete visualization (AOI1 and AOI2). Thus, a nonparametric Friedman test was conducted with fixation duration on the complete visualization as the dependent variable. The results indicate statistically significant differences in fixation duration across the eight related groups representing explanation types and loan application decisions ($\chi^2(7) = 24.298$, $p = 0.001$). Post-hoc pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction reveal significantly lower fixation duration for Anchors approved compared to DICE approved ($p < 0.001$), SHAP rejected ($p < 0.001$), DICE rejected ($p = 0.001$), and LIME Approved ($p = 0.001$). A two-way repeated measures ANOVA analysis was conducted for fixation duration as dependent variable and explanation representation and loan decision as within factors to visualize the fixation duration in the complete visualization across the within factors. The results indicate a significant interaction effect between the within factors ($F(3,54) = 3.618$, $p = 0.019$). Figure 19 shows the interaction effect between explanation representation and loan decision on fixation duration on the complete visualization.

These differences can be explained by the amount of information presented in each explanation representation for approved and rejected applications and the time required to analyze them. For LIME and SHAP, the number of attributes for both approved and rejected applications was six, which resulted in a similar fixation duration for both decisions. For DICE, the average number of attribute changes for rejected applications was 8.00 and approved 11.75. In contrast, for Anchors, the average number of rules in rejected applications was 4.75 and approved 2.00. These differences are also visible in Figure 18, showing fewer regions with a concentration of visual attention on the complete visualization of Anchors approved compared to the rest of the heatmaps.

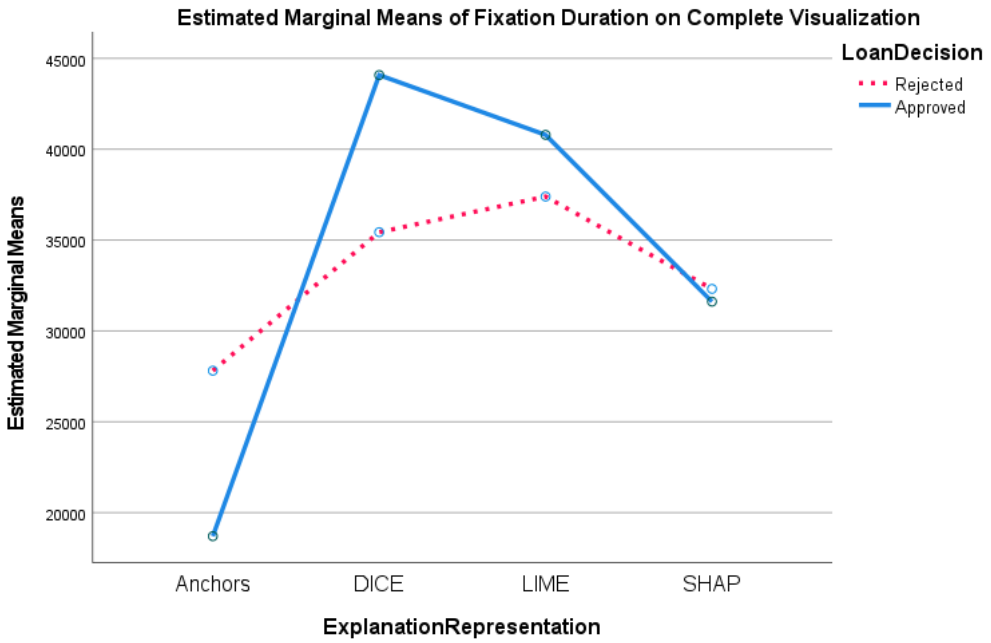


Figure 19: Interaction effect of explanation representation and loan decision on fixation duration on complete visualization (error bars not included for readability).

3.5.2.2 Explanation Representation

Comparing visual attention on the explanation representation for each type presents many challenges due to the differences in information they provide. Thus, visual attention on the explanation representations (AOI2) was analyzed as a percentage of the complete visualization (AOI1 and AOI2). After verifying the necessary assumptions, two-way repeated measures ANOVA analyses were conducted with fixation duration and number of fixations as dependent variables.

For fixation duration, the results show that the interaction effect between explanation representation and loan decision is not significant ($p = 0.808$). Moreover, there are statistically significant differences across explanation representations ($F(3,54) = 2.989$, $p = 0.039$) but no across loan decisions ($p = 0.535$). Post-hoc pairwise comparisons with Bonferroni correction revealed that the fixation duration for DICE is marginally lower than SHAP ($p = 0.094$). Figure 20 shows the fixation duration on explanation representation as a percentage of the complete visualization.

Similarly, the results show that the interaction effect between explanation representation and loan decision was not significant for the number of fixations ($p = 0.814$). Moreover, there were statistically significant differences across explanation representations ($F(3,54) = 6.097$, $p = 0.001$) but no across loan decisions ($p = 0.259$). Post-hoc comparisons with Bonferroni correction revealed that DICE's number of fixations is significantly lower than LIME ($p = 0.024$) and SHAP ($p = 0.003$). Figure 21 shows the number of fixations as a percentage of the complete visualization for each explanation type.

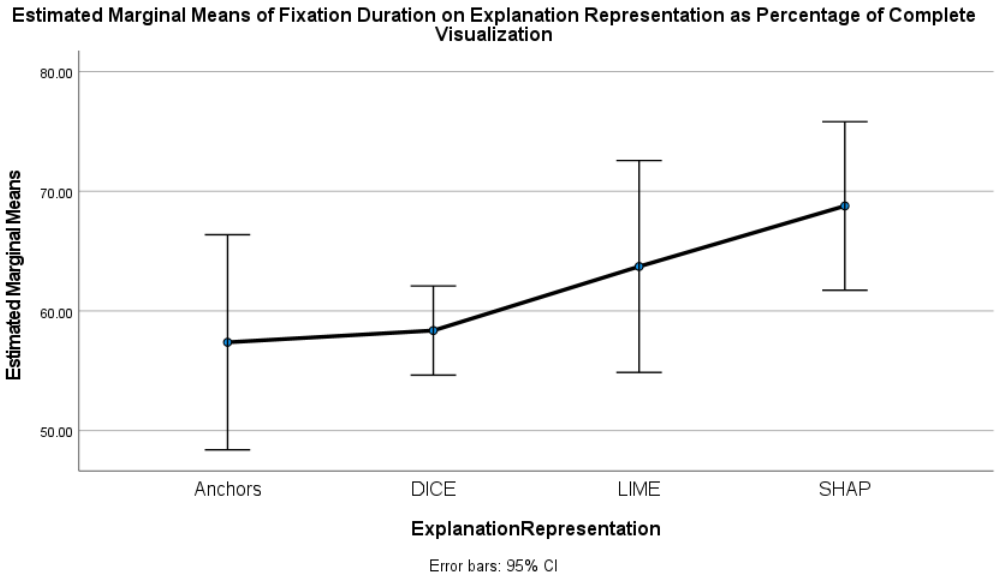


Figure 20: Fixation duration on explanation representation as a percentage of the complete visualization.

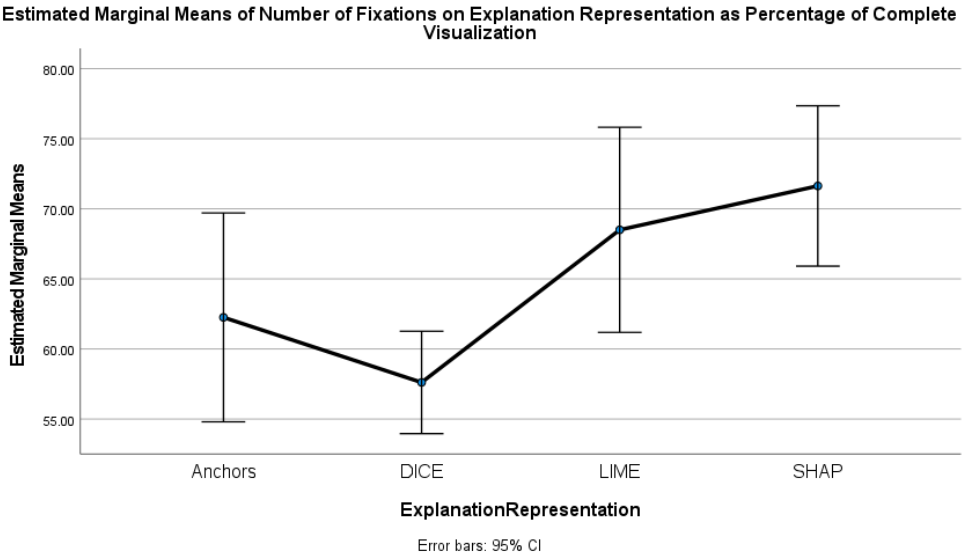


Figure 21: Number of fixations on explanation representation as a percentage of the complete visualization.

The lower values of visual attention as a percentage of the complete visualization for DICE indicate that participants had to check the attribute's table as a reference constantly. It is also possible to observe these differences in the heatmaps of Figure 18 by contrasting the distribution of visual attention between the attributes table and the representation of DICE. This design could significantly increase users' mental effort as they must transition between the attribute's table and the representation to explore and process DICE's explanations.

3.5.2.3 Elements of Explanation Representations

Similar to the approach for the explanation representations, fixation duration was analyzed on each DICE's counterfactuals as a percentage of the fixation duration in the three AOIs combined (i.e., AOI2.1, AOI2.2, and AOI2.3). After verifying the necessary assumptions, a two-way repeated measures ANOVA analysis with fixation duration as dependent variable and counterfactual and loan decision as within factors was conducted. The results show that the interaction effect between the counterfactual and the loan decision is not significant ($p = 0.616$). Moreover, there are statistically significant differences between the counterfactuals ($F(2,36) = 7.488$, $p = 0.002$). Post-hoc pairwise comparisons with Bonferroni correction revealed that fixation duration for the third counterfactual was significantly lower than the first ($p = 0.008$) and second ($p = 0.014$) (see Figure 22).

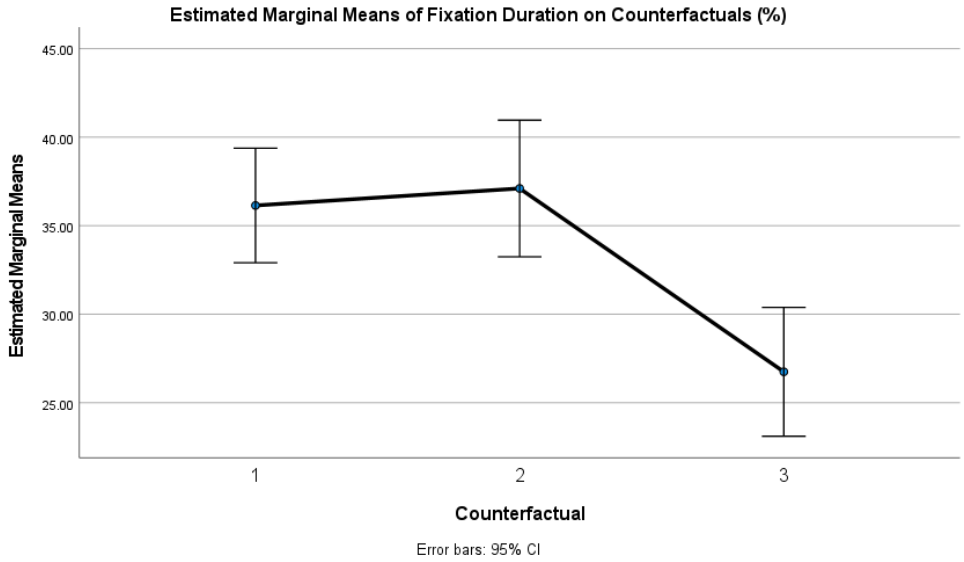


Figure 22: Percentage of fixation duration for each counterfactual in DICE explanation representation.

When observing the heatmaps for DICE in Figure 18, it is clear that participants' visual attention focused mainly on the first and second counterfactuals. Thus, DICE's explanation representation could include only two counterfactuals to reduce information overload.

Additionally, participants' fixation duration as a percentage was analyzed for LIME and SHAP by comparing top vs. bottom influencing attributes (AOI2.1 vs. AOI 2.2) and positive vs. negative influencing attributes (AOI2.3 vs. AOI2.4). After verifying the necessary assumptions, a two-way repeated measures ANOVA analysis was conducted with fixation duration as the dependent variable and explanation representation and loan decision as within factors.

Regarding the comparison of top vs. bottom attributes, the results show that the interaction effect between explanation representation and loan decision is not significant ($p = 0.120$). Moreover, there is a statistically significant higher fixation duration on top attributes for LIME than SHAP ($F(1,18) = 4.581$, $p = 0.046$) but no significant difference between rejected and approved loan applications ($p =$

0.887). Figure 23 shows the fixation duration for LIME and SHAP on top attributes as a percentage. It is possible to observe these differences in visual attention in the heatmaps in Figure 18. Participants' visual attention seems more evenly distributed between the top and bottom influencing attributes for SHAP than LIME.

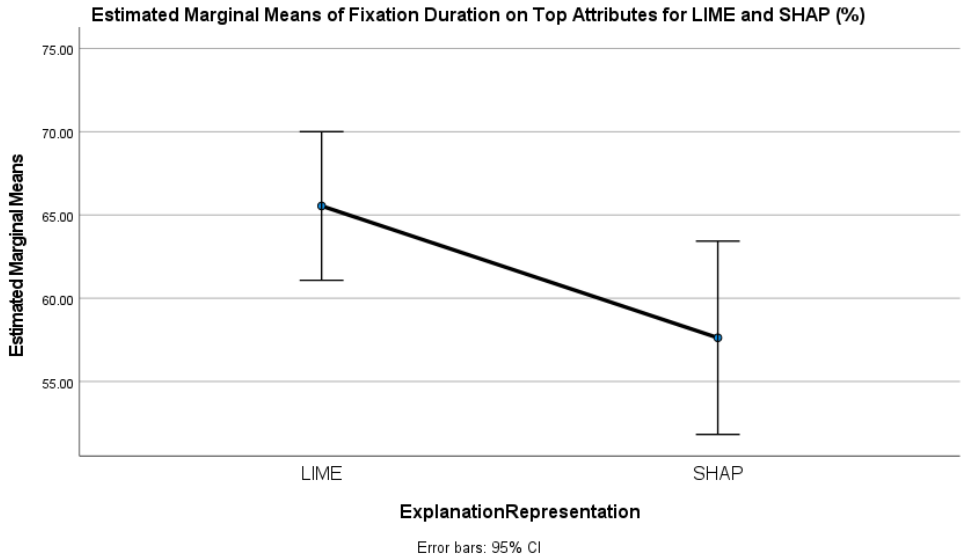


Figure 23: Percentage of fixation duration for top influencing attributes for LIME and SHAP.

The results of the comparison between positive and negative attributes reveal that the interaction effect between explanation representation and loan decision is significant ($F(1,18) = 46.216$, $p < 0.001$). Paired sample tests were conducted to further investigate the main effects of each within factor. For explanation representation, there is a significant difference between LIME rejected and SHAP rejected ($t(18) = -8.663$, $p < 0.001$) but not between LIME approved and SHAP approved ($t(18) = -1.175$, $p = 0.255$). Meanwhile, for loan decisions, there are significant differences between LIME rejected and LIME approved ($t(18) = -2.672$, $p = 0.008$), as well as between SHAP rejected and SHAP approved ($t(18) = -6.543$, $p < 0.001$). Figure 24 shows the interaction effect of explanation representation and loan decision on fixation duration on the positive attributes for LIME and SHAP. This analysis reveals that the difference in positive attributes between both explanation types is more prominent for rejected loan applications.

It is possible to observe these differences in visual attention in the heatmaps in Figure 18. Participants' visual attention seems more evenly distributed between the top and bottom influencing attributes for SHAP than LIME. The heatmaps also reveal fewer regions with a concentration of participants' visual attention on positive attributes for LIME. When further analyzing the distribution of positive influencing attributes, it was observed that the average for LIME rejected is 1.00, LIME approved 2.25, SHAP rejected 1.50, and SHAP approved 3.00. The dataset class imbalance explains these distributions of positive influencing attributes. In SHAP, this is reflected with a low base probability of 0.35, representing the percentage of rejected loan applications in the dataset. Therefore, there usually are more

negative attributes that increase the rejection probability. Nevertheless, for LIME, this class imbalance is present only in the intercept of the local linear regression, but it is not shown as part of LIME's explanation. Consequently, for certain approved loan applications, LIME's explanations can contain a majority of negative influencing attributes, which could be counterintuitive to end-users (see Figure 8).

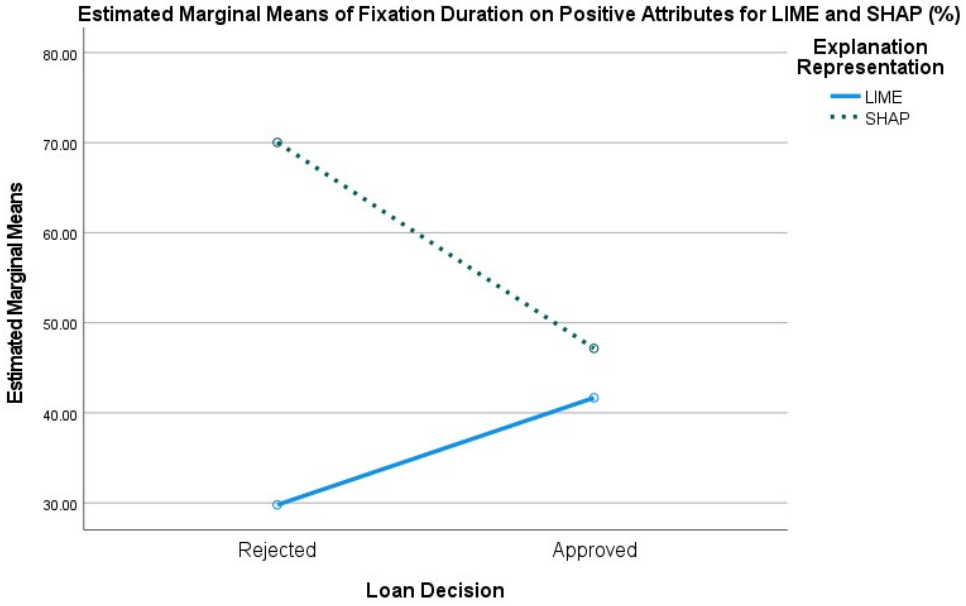


Figure 24: Interaction effect of explanation representation and loan decision on fixation duration on positive attributes for LIME and SHAP (error bars not included for readability).

3.5.2.4 Summary of Findings

The eye-tracking analyses reveal lower visual attention on Anchors than on DICE, LIME, and SHAP, which can be explained by the lower number of rules presented in Anchors' explanation representations. An interpretation of this finding is that Anchors provides simpler explanations that might require lower mental effort to be processed. Nevertheless, it is unclear if this simplicity can be translated into a better understanding for end-users of how the AI system makes decisions. Furthermore, the analyses reveal that processing DICE's explanation representation could require a high mental effort from end-users due to the number of counterfactuals shown and the need to reference the attribute's table. A further refinement of DICE's explanation representation design could reduce the number of counterfactuals.

Moreover, for SHAP, the analyses reveal a more evenly distributed participants' visual attention across the influencing attributes than LIME. Finally, the analyses reveal that LIME can generate counterintuitive explanations for imbalanced datasets by showing more attributes influencing the class contrary to the model's prediction. These counterintuitive explanations could be challenging to understand by end-users.

3.5.3 Analysis of the Semi-structured Interviews

This section reports the findings from the semi-structured interviews performed after the experiment regarding participants' perceptions of the AI system, attributes used, and explanation representations.

Anchors	DICE	LIME	SHAP
✓ **Simple and easy to understand** ✓ **Rules are simple to follow**	✓ **Straightforward explanations** ✓ **Very simple and easy to follow** ✓ **Helpful to see how to change system's decision**	✓ **Easy to visualize and understand** ✓ **Correct amount of information** ✓ **Clearly shows how much each attribute contributed to decision** • Easy to compare attributes ✓ **More intuitive and simple than SHAP**	✓ **After explanation was understood, it was found as very helpful** ✓ **Stacked attributes provides a "great overview"** • How all attributes interact ✓ **Only explanation that provides probabilities** • More confident explanations
X **Misinterpretation of explanation:** • System doesn't consider other attributes • Very strict and specific • Too "stupid" rules. • Possible to find "loopholes" X **System was not perceived as AI**	X **Difference in perception when considering participants gender** • More positive by females X **Representation is difficult to read** • Too much detail X **No information on how much influence each attribute had** X **Just a subset of "infinite" possible counterfactuals** X **Unrealistic attributes' modifications**	X **Attributes' influence as percentage was confusing** X **Not clear how attributes' influence was calculated** X **For some approved loans, majority of attributes' influence was negative → contra intuitive** • Known problem due to class imbalance	X **Initially the explanation was confusing and it took some time to understand** X **Probabilities could be difficult to understand** X **Explanation is complex**

Figure 25: Summary of findings from the semi-structured interviews.

Overall, the analysis shows that end-users process each explanation type very differently and that some factors can influence their perceptions and preferences on explanation representations. It can also be observed that in certain situations, participants' misinterpretation of the explanation representations negatively influenced their perception. Moreover, the interview data indicates that some explanation types might be adequate for specific situations. Figure 25 presents a summary of the most relevant findings of the interviews for each explanation representation.

3.5.3.1 AI System

Regarding participants' general perception of the AI system, 40% said it was reliable. Additionally, 13% of participants indicated that explanations are helpful to understand how the system works and assist with decision-making. Some examples include, "*I felt that it could help a person that needs to make a decision*" (P4) "*I know what I have to change to get my loan approved*" (P11), and "*I think it could be reliable, … there were some parameters that confused me where I personally probably would have decided differently*" (P3). Meanwhile, P10 perceived the system's reliability as limited to some explanation representations, stating: "*only [LIME] and [SHAP] are reliable recommendations*". Four (18%) participants indicated they were unsure whether the system was reliable. P13 stated: "*I am not really sure [I can rely on it], … I need more information*". Participants' overall relatively positive perception of the system indicates the positive effects of explanations on end-users' perceptions. Nevertheless, factors that influence end-users' perceived reliability need to be considered. This need is further discussed in the following sections.

3.5.3.2 Attributes

45% of participants highlighted situations where the values of the attributes seem to have a contra-intuitive influence on the system's decision. For example, for the attribute loan history, 31% of participants mentioned that the negative influence of the value "paid back all previous loans" on the system's decisions was unexpected. This influence contrasts with participants' expectations that having a good credit history's repayment should positively influence the system's decision, as shown by the P9 statement: *"[The system] always said that it is negative, and I never understood it*". The categories distribution for this attribute was analyzed, and it was observed that there were 361 approved and 168 rejected loan applications. Thus, 68% of the loan applications with this value are approved. Nevertheless, the 168 rejected loan applications with this attribute value represent 56% of all 300 rejected applications in the dataset. As a result, the algorithm correlates the value "paid back all previous loans" for the attribute loan history with a rejection of loan applications. Researchers and practitioners need to consider the implications of providing explanations for AI systems' decisions. Explanations can reveal potential flaws or unexpected behaviors in the model, negatively influencing end-users' perceptions and adoption of these systems.

3.5.3.3 Explanation Usefulness

Moreover, 59% of participants considered that explanations' usefulness depends on the person receiving them. 27% of participants commented on the usefulness of explanations for bank employees who need to make decisions. Four (18%) participants indicated that LIME and SHAP are more helpful for decision-makers as they provide relevant information regarding each attribute's influence and relevance in relation to other attributes, as stated by P4: "*SHAP is more confident and could help people in a bank to make decisions*". Meanwhile, four (18%) participants stated that Anchors and DICE explanations are not helpful for bank employees in some scenarios. P1 stated for DICE: *"[Bank employees] probably do not care that much about how to change the profile of the [applicant], …. this explanation is not useful for [them]*". Similarly, P7 stated: "*In the case of [approved loan applications], showing how it can be rejected is rather useless*". For Anchors, P9 stated: "*For someone who has to decide, [they] would not know how much everything affects the outcome*". Furthermore, 31% of participants indicated that certain explanations could be more useful for customers applying for bank loans. Six (27%) participants stated that DICE counterfactual explanations for rejected loan applications are very useful in this regard. P6 stated that "you can see exactly what needs to change in order for it to be approved".

3.5.3.4 Anchors

59% of participants stated that they perceived Anchors' explanations as simple and easy to understand. Participants stated that Anchors' "*rules were simple to follow*" (P6) and that they helped understand "*why [a loan application] was rejected or approved*" (P7). Nevertheless, two (9%) participants stated

that despite the simple representation of Anchors' rules, interpreting them was difficult, including "*I felt that I was reading a flow chart or a code diagram*" (P4). P5 stated that for "*someone that does not have programming knowledge, it would be difficult to [interpret them]*".

An interesting finding regarding Anchor's explanations is that several participants misinterpreted them, influencing their perception. Six (27%) participants perceived Anchors' explanations as "*very strict*" (P11) and "*very specific*" (P19). Their interpretation of Anchors' explanations led them to strongly criticize that the system "*only takes [those rules] into account*" (P6) and does not consider other attributes for its decision-making process. Similarly, the interpretation of two (9%) participants led them to highlight a potentially problematic situation. P19 and P21 interpreted Anchors' rules as "*too stupid because you can easily find loopholes*" (P21). They believed that for an approved loan application with only a few rules (e.g., account balance and purpose), dramatic changes to the other attributes would not influence the loan approval (e.g., unemployment). Likewise, P10 perceived that the system providing Anchors' explanations was not even AI-based because it was only following simple rules that were programmed. This analysis illustrates why participants had lower perceived usefulness and why Anchors' explanations were the least preferred.

3.5.3.5 DICE

For DICE explanations, a clear difference in participants' perceptions was observed when considering their gender. This difference is in line with the analysis of participants' perceived usefulness and the interaction effect with their gender in Section 3.5.1. As observed in this analysis, DICE's explanations had the most considerable difference in perceived usefulness for both genders. In the interviews, it was found that females provided more positive comments for DICE's explanations than males and that males provided more critiques than females. Thus, to help identify these differences, the participants' gender is presented in an aggregated form throughout the analysis for DICE's explanations.

Eight (36%) participants, of which four were females and four were males, indicated that DICE's explanations were "*straightforward*" (P4), "*very simple and easy to follow*" (P2). Additionally, 14 (63%) participants, of which eight were females and six were males, mentioned that DICE's counterfactual explanations are useful as they could "*see exactly what [they] need to change [the system's decision]*" (P6). They liked that DICE's counterfactuals for rejected loan applications allow them to know what the applicant "*could have done better*" (P20) to get the loan approved. On the other hand, four (18%) male participants commented that DICE's explanations provide "*too much detail*" (P1) and that "*having three columns [with counterfactuals] made it too difficult to read*" (P5). The eye-tracking analysis supports these design critiques, as it was found that the fixation duration was higher for DICE than for other explanations. Thus, DICE's explanation representation design could be improved by reducing the number of counterfactuals shown.

Moreover, three male and one female participant (18%) stated that with DICE's explanations, they "*do not know what is more important for the [system]*" (P3) as there is no information on "*how much influence did the attributes have [on the decision]*" (P18). Additionally, three male and one female participant (18%) indicated that DICE's counterfactuals are "*not enough to understand how [the system] makes decisions*" (P19) because they only show a limited number of attributes' change. P21 criticized that shown counterfactuals are "*only three examples [of changes that would lead to a different decision], but there might be hundreds or thousands of other [possible] combinations*". P10 stated that some of the attributes' modifications proposed in the counterfactuals are not helpful as they proposed unrealistic modifications (e.g., change job or employment time). This issue with unrealistic modifications could be addressed using DICE's library functionality to specify feature weights to restrict the search of counterfactuals and avoid generating explanations that include immutable feature changes. It was decided not to provide such restrictions for this study as it could induce the researchers' own bias into the generation of counterfactuals.

3.5.3.6 LIME

The overall participants' perceptions of LIME explanations were positive. 40% of participants commented that they perceived LIME's explanations as "*easy to visualize and understand*" (P4). Additionally, two (9%) participants mentioned that explanations had the "*correct amount of information, not too much or too little*" (P1). Nine (40%) participants stated that LIME's explanations "*made it clear which attributes contributed to the decision*" (P12), which allowed "*an easy comparison between [the attributes]*" (P8). Despite the similarity between LIME and SHAP explanations, six (27%) participants indicated that LIME's explanations were more "*intuitive and easy to understand*" (P12) as they provided "*a better overview*" (P8).

Nevertheless, participants also highlighted some problems with LIME explanations. Four (18%) participants criticized how the attribute's influence was presented as a percentage because they interpreted it as "*not really a unit*" (P18) that "*did not mean much*" (P7) for them. They stated that LIME's explanations were missing the probability shown in SHAP as it provides a reference of how confident the decision is and how difficult it would be to change it. Moreover, five (22%) participants did not understand "*how [the percentages] were calculated*" (P6) or "*why [the attributes] affected so much or so little*" (P10). Furthermore, three (13%) participants highlighted that LIME's explanation had more attributes influencing rejection for some approved loan applications, which was very confusing. As mentioned in the analysis in Section 3.5.2.3, this issue is caused by a class imbalance in the dataset, having 70% of approved loan applications and 30% rejected. This imbalance is included in LIME's linear regression as the intercept, which considerably influences approving loan applications. Nevertheless, LIME's explanations do not display the intercept (see Figure 8).

3.5.3.7 SHAP

For SHAP, five (22%) participants indicated that they initially found the explanations confusing and that it took them some time to understand them. However, when they understood the concept, they found them very useful. Eight (36%) participants mentioned that they could clearly observe the attributes' influence and their interaction. In particular, three participants stated that having the stacked bar at the top of the explanation provided a great overview of all attributes' influence, including "*attributes are shown stacked to see how they aggregate their influence. It is almost perfect*" (P2). Additionally, seven (31%) participants stated that SHAP's explanations were the only ones that provided probabilities for the system's decision.

In contrast to the other explanations, participants consider SHAP's explanations as "*more confident*" (P4) due to their probabilistic nature. P2 stated that the base probability and the decision probabilities provided a reference to understand better the interaction of the attribute's influence in the system's decision. Nonetheless, three (13%) participants highlighted that everyone might not easily understand SHAP's explanations because they might require probability knowledge. This was the case for two participants, including P16, who stated that SHAP's explanations were "*hard to understand*," and P4, which stated, "*what do the numbers really mean?*".

3.6 Discussion

In the present work, local model-agnostic explanation representations from established XAI methods were refined following an iterative design process. Leveraging eye-tracking technology, self-reports, and interviews in the evaluation of the proposed designs helped to better understand how end-users process and evaluate local model-agnostic explanation representations from XAI methods. The following section discusses this study's findings and limitations and suggests possible directions for future research.

3.6.1 Justification

According to Swartout (1985), systems must be able to explain their decisions and justify them to users in an understandable way. Swartout argues that systems that fail to justify their decisions would not be accepted by their users. The results seem to indicate that some end-users might require more comprehensive explanations, which provide them with a reasonable justification for the system's decision. It was found that participants' preferences of explanations were influenced by their knowledge background and experience with AI systems in general and ML specifically. In particular, for participants with higher ML knowledge, there was a higher variance in satisfaction across different explanation types.

A qualitative analysis of data from semi-structured interviews further found that some participants were dissatisfied with DICE counterfactual explanations because they perceived that these explanations did not explain how the AI system had reached that decision. Counterfactual statements do not comprehensively explain the system's decisions for these participants. These counterfactual statements represent only one of the many possible scenarios that could lead to the system making an alternative decision. Additionally, for these participants, counterfactual explanations do not answer why the system had made the decision. Hereby, participants were looking for an explanation that could clarify the internal logic of the predictive model. They were interested in understanding each attribute's influence on the system decisions and how these influences interacted.

Moreover, the interaction effect of end-users' experience with AI systems on their preference for specific explanations indicates that researchers need to consider other potential end-users' characteristics when designing XAI explanations. Additional factors could influence end-users' need for comprehensive justification of AI system decisions.

3.6.2 Comparison of the Evaluated Local Model-agnostic Methods

Due to its complex design and the amount of information it provides, there were many challenges in refining SHAP's out-of-the-box explanation representation at the beginning of the iterative design process. Additionally, there were concerns that end-users might not understand and trust SHAP explanation representations due to their probabilistic nature. Nevertheless, the iterative design process evaluations reveal that SHAP explanation representations were perceived similarly to Anchors, DICE, and LIME regarding trust, understandability, and satisfaction. In this line, the data analysis of the interviews revealed that despite participants' initial challenges interpreting SHAP explanations, they considered them to be very useful. Participants indicated that the base and decision probabilities provided a reference to understand better the interaction of the attributes' influences in the systems' decisions. Eye-tracking analyses supported these findings by revealing a high concentration of participants' visual attention on regions of SHAP explanation representations that show the base and decision probabilities.

According to the interviews, the rules generated by Anchors' explanations were perceived as very simple and easy to understand. These findings align with the eye-tracking analyses that revealed lower participants' visual attention on Anchors compared to the other methods. Moreover, since Anchors' rules highlight the subset of input attributes sufficient for the model to make a particular decision, they allowed participants to generalize these rules and apply them to other instances on the dataset to understand the system's decision. Nevertheless, participants in the laboratory experiment misinterpreted Anchors' explanations. They interpreted Anchors' explanations as a set of static rules that did not consider other attributes for the system's decision. Thus, they believed the system was not AI-based and could be fooled as loan applicants would know which attributes are "not considered" by the system to

make decisions. This type of misinterpretation can strongly negatively influence end-users' perceptions and adoption of an AI system.

In line with findings in the literature (Miller, 2019), DICE's counterfactuals were found to be straightforward explanations. In the interviews, participants indicated that counterfactuals were simple and easy to follow and highlighted that seeing how to change the system's decision was very helpful. Nonetheless, they also mentioned that counterfactual explanations for approved loan applications are not very useful as they indicate modifications that would cause the loan application to be rejected. Thus, counterfactual explanations in bank loan application evaluations might be limited to explaining rejected explanations. Eye-tracking analyses revealed design flaws in DICE's explanation representations. Specifically, an analysis of participants' visual attention revealed that they need to constantly check the attribute's table as a reference to interpret the counterfactuals. Moreover, these analyses also revealed that participants' visual attention was mainly focused on the first two counterfactuals. These findings were supported by the interviews with some participants indicating that the representation was challenging to read as it had too much detail.

Overall, LIME explanations were perceived as easy to visualize and understand because they balanced the right level of detail and no information overload. Many participants preferred LIME explanations over SHAP because they found them more straightforward and intuitive. Nevertheless, eye-tracking analyses revealed that in the presence of class imbalances on the dataset, LIME could provide contra-intuitive explanations showing more attributes influencing the opposite class than the one predicted by the system (see Figure 12). In the laboratory experiment, participants were frustrated when presented with these contra-intuitive explanations. These problematic explanations could adversely affect the adoption of AI systems providing LIME explanations.

3.6.3 Limitations and Future Work

The research in this study also comes with limitations. The iterative design process and evaluations were performed only in the context of bank loan applications. This domain was selected due to its relevance as financial institutions increasingly use AI systems to evaluate loan applications, and the resulting decisions can significantly impact loan applicants. Nevertheless, providing explanations of AI decisions to end-users is a critical issue in many other domains. End-users' needs could differ for domains other than bank loan applications. Thus, future work is needed to evaluate local model-agnostic explanations in other contexts to understand these end-user needs and consider whether different designs are required.

Additionally, the iterative design process and evaluations focused on refining the explanation representations for a binary classification task on a tabular dataset. For example, the design of counterfactual representations requires that the attribute modifications proposed by the counterfactual statements are aligned with a table containing the attributes' names so they can be used as a reference.

Nevertheless, due to the extensive research in XAI, many explainability methods have been developed to provide different types of explanations according to the type of data and ML task. Further research is needed to evaluate additional model-agnostic methods across different dataset types (e.g., visual, or textual) and ML tasks (i.e., supervised, and unsupervised).

Finally, to evaluate the explanation representation designs with end-users, the explanations provided during the interaction with the AI system were generated beforehand and presented a fixed number of attributes. Providing interactive explanations that allow end-users to explore the complete details of the explanations, such as the influence of all attributes, could enable them to understand better AI systems' decisions (Meza Martínez et al., 2019). Therefore, future work is needed to evaluate how model-agnostic methods generate explanations and the implications of implementing them more dynamically, e.g., through increased interactivity.

3.7 Conclusion

In this study, comparable local model-agnostic explanation representations from well-established XAI methods were derived through an iterative design process. Furthermore, eye-tracking technology, self-reports, and interviews were used to understand how end-users process and evaluate these explanation representations. The results indicate that local model-agnostic explanations from different XAI methods can effectively establish satisfaction, trust, and understandability. Nevertheless, end-users might find some explanations more useful in specific scenarios. Moreover, the results indicate that end-users' preferences for model-agnostic explanations are influenced by their individual characteristics, such as gender and previous experience with AI systems. This study contributes to the ongoing research on improving the transparency of AI systems by explicitly emphasizing the end-user perspective on XAI.

4 Study II: Why End-users Trust and Not Trust Biased XAI Systems: A Psychological Contract Violation and Social Identity Perspective[19]

4.1 Introduction

Systems based on AI support human decision makers in various critical tasks, such as criminal justice, hiring, or healthcare diagnoses. Explaining AI systems' decision recommendations to decision makers has been repeatedly proposed as an essential mechanism to increase the quality of augmented decision-making as it allows end-users to align their knowledge with the reasoning logic of the AI system (Fügener et al., 2021; Jussupow, Spohrer, et al., 2021). Therefore, XAI methods are proposed as a mean to explicate the decisions of underlying "black-box" ML algorithms in nontechnical terms to end-users (Dodge et al., 2019; Miller, 2019). These methods can help to clarify how AI systems make decisions by revealing how certain attributes influence their decision recommendations.

However, in the context of biased AI systems existing literature proposes two opposing theoretical mechanisms that describe how explanations can affect end-users' trust. On the one hand, explanations can disclose a bias resulting in perceiving the AI system as unfair and less trustworthy (Dodge et al., 2019; Law et al., 2020). For augmented decision-making, end-users need to be aware of these biases to engage in various activities to de-bias them. Thus, identifying biases through explanations can influence end-users to decide whether to rely upon the provided decision recommendations (Teodorescu et al., 2021). As a result, explanations can negatively affect end-users' trust. On the other hand, through explanations, end-users generally trust the system more (W. Wang & Benbasat, 2007; Yang et al., 2020), increasing the likelihood that end-users will agree with the provided recommendations (Yeomans et al., 2019). Thus, explanations can increase the perceived transparency of AI systems and end-users' knowledge-based trust. In turn, this increase in perceived transparency can increase the likelihood that end-users agree with a decision recommendation, even though it might be based on biased attributes (W. Wang et al., 2018; W. Wang & Wang, 2019). Specifically, the effect of disclosing a bias can be compensated by the perceived plausibility of all considered attributes. As a result, disclosing a bias through explanations does not necessarily result in less trust in the AI system.

To understand the contingency factors of explanations' positive and negative effects on end-users' trust in biased AI systems, it is necessary to examine under which circumstances end-users perceive an AI system as biased and untrustworthy or as transparent and accurate. This study relies on the psychological

[19] This chapter is based on the following papers: (Jussupow et al., 2023; Jussupow, Meza Martínez, et al., 2021). The data analysis and experimental material are found in the following GitLab repository and RADAR archive:
https://git.scc.kit.edu/h-lab/research/2092_meza_miguel_gender-biased-XAI-experiments
https://radar.kit.edu/radar/de/dataset/kFBqZNXhLtDBOmuH

contract violation (PCV) theory (Morrison & Robinson, 1997) as a theoretical lens to investigate the underlying cognitive evaluation process of explanations in augmented decision-making. This approach has already been applied to study trust in biased recommender systems (W. Wang et al., 2018; W. Wang & Wang, 2019) and the perceived fairness of AI systems (Tomprou & Lee, 2022). Aligned with prior research, a *psychological contract* in the interaction with a system providing recommendations can be conceptualized as an implicit expectation that the system will perform accurately and neutrally (W. Wang et al., 2018; W. Wang & Wang, 2019). The PCV theory proposes that individuals respond negatively to an entity if this contract is violated and lose trust in an information system (W. Wang et al., 2018; W. Wang & Wang, 2019). End-users form similar psychological contracts with algorithms and humans (Tomprou & Lee, 2022). However, before interacting with AI systems, end-users expect AI systems to perform neutrally and nondiscriminatory, i.e., fairly, because of the underlying machine heuristic (Sundar, 2008, 2020). Thus, in the context of biased AI systems, a PCV only results after a series of cognitive evaluations: First, end-users need to experience a *perceived unmet promise* by detecting a bias in the system. Second, they must *interpret* the contract breach as a *psychological contract violation* for which other factors cannot compensate. Explanations have been shown to prevent the occurrence of a PCV, even if the system is biased (W. Wang et al., 2018; W. Wang & Wang, 2019).

However, this study aims to address two significant limitations of prior work investigating explanations for biased AI systems. First, in contrast to previous work that demonstrated the compensatory effect of explanations on the experience of a PCV for biased output of recommender systems (W. Wang et al., 2018; W. Wang & Wang, 2019), this study considers biases that are often associated with discriminating decisions against minorities, women, and people of color, which are the result of biases inherent in the training data (Buolamwini & Gebru, 2018; M. P. Kim et al., 2019). In this case, the interpretation of biases can be more ambiguous as these are often hidden in the data. Second, prior research on XAI has mainly focused on the knowledge-based impact of explanations and neglected how end-users' emotional responses influence how they evaluate explanations for AI systems' decisions (Kordzadeh & Ghasemaghaei, 2021; Starke et al., 2021). Likewise, prior research on social identity theory (Tajfel, 1982) suggests that end-users differ in their sensitivity to biases depending on their own past experiences and their social identification with the stigmatized group, i.e., their stigma consciousness (Pethig & Kroenung, 2020; Pinel, 1999). Therefore, end-users with a stronger identification with the stigmatized group are more sensitive toward these biases and respond more negatively to them. Hence, in the context of biased AI recommendations, end-users' social identity can strongly influence if they perceive an AI system as biased and experience a PCV resulting in a loss of trust. Therefore, this study aims to investigate the role of end-users' stigma consciousness as a contingency factor in their evaluation of explanations that are provided by an AI system that affects their social identity:

***RQ**: How do end-users differ in their evaluation of a biased XAI system's trustworthiness based on their level of stigma consciousness?*

To investigate the positive and negative consequences of explanations for a biased AI system, this study investigates whether end-users automatically reject the AI system if a bias is visible in the explanation or if the positive impact dominates, i.e., the explanation persuades these end-users. Hence, an experimental setup was created with the following two boundaries: (1) a simple task was selected that laypeople could perform to assess the general cognitive impact of explanations on trust. Specifically, participants were asked to decide upon granting or declining consumer loans; (2) a bias that end-users could easily identify was selected. Thus, a gender-biased AI system was selected as an example of a biased AI system. Although gender is considered as a protective attribute, gender bias has been repeatedly demonstrated as a challenge in implementing AI systems. For instance, Buolamwini and Gebru (2018) revealed that several commercial face recognition algorithms significantly differed in their classification error rate for darker-skinned women vs. lighter-skinned men. Therefore, biases inherent in the data can lead to discriminatory decision recommendations that favor men over women, i.e., cause a gender bias.

Two preregistered online experiments were conducted in this Study. In Experiment 1, participants were allocated into three experimental conditions: In the biased AI condition, participants interacted with an AI system that systematically favored male loan applicants over female applicants. These participants were provided with explanations generated by the XAI method LIME (Ribeiro et al., 2016b), which displayed gender as an influential attribute on the recommendation in the explanations. Thus, the bias was directly visible to end-users, exaggerating the effect of the bias. In the neutral AI condition, however, the provided LIME explanations did not display gender as an influential attribute of the recommendation in the explanations. Lastly, participants in the control condition received only the recommendations without explanations. Survey responses were analyzed to determine whether participants perceived the AI system to be biased, whether they experienced a PCV, and whether their trusting intentions were affected. Further, this experiment assessed how participants perceived the plausibility of each decision recommendation. In Experiment 2, the findings of Experiment 1 were expanded by influencing participants' stigma consciousness through priming. Thus, a 2x2 experiment was conducted to manipulate the AI system (neutral vs. biased AI) and priming (priming vs. no-priming).

Overall, the experiments conducted in this study demonstrate that end-users do not automatically perceive the AI system as biased, experience a PCV, and lose trust in a gender-biased AI system. Instead, they reveal that end-users differ in their evaluations based on their level of stigma consciousness. Specifically, end-users weigh the perceived plausibility of the decision recommendation against the presumed bias of the AI system by evaluating the attributes presented in an explanation. Thus, some end-users perceive the system as plausible despite the underlying gender bias and do not experience a PCV from the disclosure of the gender bias if their sensitivity toward the social bias is low.

Given these findings, this study has the following contributions: (1) the PCV theory is contextualized to evaluations of social biases inherent in the data of AI systems, (2) insights into how end-users cognitively evaluate explanations based on their social identity are provided, and (3) the current understanding of the impact of explanations on trust is expanded by examining both the positive and negative impact of explanations. Further, this study contributes to practice by providing insights that help to improve design explanations for AI systems and informing organizations to consider the consequences of biased AI systems for AI-augmented decision-making.

4.2 Research Model and Hypothesis Development

The subsequent section outlines the related literature and the theoretical foundations of PCV theory and social identity theory to develop a well-grounded research model regarding the effect of explanations for a gender-biased AI system on end-users' trust.

4.2.1 Related Work on Explanations for Biased AI Systems

An exploratory literature review was conducted to assess whether explanations can help end-users detect and evaluate biased AI systems. Established digital libraries were examined to search for relevant studies evaluating end-users' interaction with biased systems. This search focused on research that evaluated bias detection, perceived bias or fairness of biased systems, or perceived trust in biased systems. In total, 16 relevant research papers were identified. Table 8 presents the results of the literature review. This table indicates whether a paper investigates biases related to social stereotypes in society (i.e., "Social") or other types of biases (i.e., "Functional"). Additionally, this table indicates whether papers measured how end-users detect and interpret biases during their interaction with AI systems under the category *bias*, whether the paper measured perceived understandability and perceived transparency in the category *transparency*, and whether the paper measured perceived *trust*. The table further classifies the applied XAI methods into *model-specific* and *model-agnostic*. Finally, the identified papers are allocated into three groups based on the AI system's characteristics and whether explanations were included in the study.

Group A consists of research papers related to the XAI field. Those studies investigate the interaction with AI systems that provide mainly model-agnostic explanations to end-users to help them understand AI's decision recommendations. All studies in this category considered social biases. Of those studies, eight investigated gender bias (Berendt & Preibusch, 2017; Binns et al., 2018; Dodge et al., 2019; Kasinidou et al., 2021; Lakkaraju & Bastani, 2020; Law et al., 2020; Schoeffer et al., 2021; Yan et al., 2020). Studies in this category have revealed that detecting a social bias is anything but trivial (Lakkaraju & Bastani, 2020; Law et al., 2020). Nevertheless, these studies do not explicitly examine the relationship between perceived bias, transparency, and trust.

Group B contains research papers investigating the effect of explanations with non-AI-based systems. The type of explanations provided in these studies is model-specific. Therefore, the generation of these explanations cannot be adapted to other systems, limiting the possibility of replicating the system functionality and even the evaluation approach. Furthermore, while the work of W. Wang et al. (2018, 2019) has considered bias detection, transparency, and trust, the bias studied by the authors can be characterized as functional. In their study, the authors evaluated sponsorship disclosure. Hence, bias means that the recommender system favored sponsored content over not sponsored content displayed to end-users. The authors did not consider social biases. R. Wang et al. (2020) have also considered gender bias in their study. However, instead of providing detailed explanations of the algorithm's decision, the explanation solely provided the algorithm's accuracy for different groups. Therefore, those papers do not reveal how end-users examine explanations generated by XAI methods and offer only limited insights into how end-users interact with social biases.

Group C contains research papers that examine end-users' interaction with systems that are functionally or socially biased. These studies do not consider the effect of explanations but show how biases affect the perceived fairness or usefulness of the system (Shandilya et al., 2020). However, only Brauner et al. (2019) show how functional biases reflected in end-users' perceived usefulness affect trust in the system. While those studies show how end-users perceive system fairness, they lack insights into how end-users cognitively evaluate explanations.

Table 8: Results of literature review of biased AI systems.

	System characteristics			**Perceptions of end-users**		
Reference	**AI-based**	**Type of explanation**	**Type of bias**	**Bias**	**Transparency**	**Trust**
Group A: Studies focusing on explainable AI systems with experimental evaluation of user interaction						
Berendt & Preibusch, 2017	Yes	Model-agnostic	Social	Yes	No	No
Binns et al., 2018	Yes	Model-agnostic	Social	Yes	No	Yes
Dodge et al., 2019	Yes	Model-agnostic	Social	Yes	No	Yes
Hoque & Mueller, 2021	Yes	Model-agnostic	Social	No	No	Yes
Kasinidou et al., 2021	Yes	Model- specific	Social	Yes	No	Yes
Lakkaraju & Bastani, 2020	Yes	Model-agnostic	Social	No	No	Yes
Law et al., 2020	Yes	Model-agnostic	Social	Yes	No	No
Schoeffer et al., 2021	Yes	Model-agnostic	Social	Yes	No	Yes
Yan et al., 2020	Yes	Model-agnostic	Social	Yes	No	No
Group B: Studies including both explanations and bias without being based on AI						
R. Wang et al., 2020	No	Model-specific	Social	Yes	Yes	No
W. Wang et al., 2018	No	Model-specific	Functional	Yes	Yes	Yes
W. Wang & Wang, 2019	No	Model-specific	Functional	Yes	Yes	Yes
Group C: Studies investigating the influence of biases on user interaction						
Barlas et al., 2019	Yes	--	Social	Yes	No	No
Brauner et al., 2019	No	--	Functional	Yes	No	Yes
Marcinkowski et al., 2020	Yes	--	Social	Yes	No	No
Shandilya et al., 2020	Yes	--	Functional	Yes	No	No

Overall, the exploratory literature review seems to indicate that so far, no study has considered the interplay of the use of explanations generated by XAI methods, gender biases of AI systems, and the implications to end-users' perceived bias, transparency, and trust. Therefore, this study develops specific hypotheses on how end-users evaluate gender-biased AI systems using the PCV theory and prior research on explanations.

4.2.2 Impact of Explanations on Trust in a Biased AI System

Novel methods of XAI have been introduced to illustrate which attributes significantly influence a particular decision recommendation of an ML model. Over the last decades, explanations have been repeatedly utilized to increase trust in decision support and AI systems (W. Wang & Benbasat, 2007; Yang et al., 2020) because they increase perceived transparency and enable end-users to verify the provided recommendations (Dhaliwal & Benbasat, 1996; Gregor & Benbasat, 1999). Explanations for the recommendations of AI systems enable end-users to judge the quality of recommendations by validating the system's underlying reasoning process (Bussone et al., 2015; Springer et al., 2020). Thus, in general, explanations are an effective mechanism to improve adherence to the recommendations of AI systems. They can therefore increase the overall decision-making performance if these systems provide accurate recommendations. In line with this research, this study proposes that explanations increase end-users' knowledge-based trust because the system is perceived to be more transparent (W. Wang & Benbasat, 2007; W. Wang & Wang, 2019). In contrast, if no explanations are provided, end-users cannot verify the plausibility of the system's recommendation. Therefore, this study proposes that explanations generally increase the perceived transparency of an AI system, even if the AI system is biased. Thus, this study hypothesizes:

H1: Explanations for a gender-biased AI system increase end-users' trust in the AI system. This effect is mediated through perceived transparency.

However, in the context of biased recommendations, there is conflicting evidence regarding the effect of explanations on augmented decision-making and trust. Specifically, disclosing a bias can result in end-users perceiving the system as unfair and disagreeing with it (Dodge et al., 2019; Law et al., 2020).

Literature on the cognitive evaluation of explanations from XAI methods indicates that end-users evaluate these explanations based on their own mental model. Thus, end-users seek to achieve coherence between their mental model of which attributes should influence a particular decision and the attributes displayed in the explanation for the AI system's decision (Miller, 2019; Thagard, 1989). Nevertheless, explanations can have different effects on perceived fairness and trustworthiness depending on end-users' personal concepts of fairness or if they are personally affected by specific decisions (R. Wang et al., 2020). Therefore, not all end-users respond in the same way to displayed biases.

In addition, end-users are not always accurate in evaluating explanations as these can persuade decision makers. For example, Bussone et al. (2015) demonstrated that providing explanations can increase trust and cause overreliance on decision support in the context of clinical decision support systems. Similarly, Nourani et al. (2021) revealed that explanations can anchor end-users. Moreover, Poursabzi-Sangdeh et al. (2021) showed that end-users were less likely to detect an AI system's mistakes when explanations for its decisions were provided. Moreover, Dodge et al. (2019) revealed that certain types of explanations were more effective in exposing case-specific bias issues while others were less effective, as end-users were not always successful in detecting a bias. Additionally, the attributes displayed in the explanation can influence how end-users assess a biased decision outcome. For instance, Law et al. (2020) demonstrated that end-users are less likely to consider additional attributes not displayed in explanations, despite carefully auditing a biased AI system. Lakkaraju and Bastani (2020) showed that end-users react differently to biased AI systems depending on whether the provided explanations reveal discriminating attributes (e.g., race or sex) or if these attributes are hidden. In particular, their study outlined that end-users can be misled to trust a biased AI system if biases are not directly displayed through discriminating attributes in the explanation but are only implied through correlated attributes (e.g., zip code). Other studies have suggested that explaining a bias to decision makers can compensate for the adverse effects of biases on trust by making the bias more acceptable to decision makers (Erlei et al., 2020; W. Wang et al., 2018; W. Wang & Wang, 2019). Overall, these findings suggest that providing explanations for a biased AI system can positively and negatively affect the system's trustworthiness.

4.2.3 Impact of a Gender-Biased AI System on Trust: A PCV Theory Perspective

To theorize on the negative influence of explanations for biased AI systems and the contingency factors, this study takes a theoretical perspective that describes the underlying cognitive evaluation process accounting for different trade-offs. Using the psychological contract violation theory (Morrison & Robinson, 1997; Robinson, 1996), this study theorizes on the subjective evaluation steps occurring before end-users decide not to trust a gender-biased AI system. PCV theory proposes that end-users form an implicit psychological contract with an entity. This contract is defined as *expectations about reciprocal obligations* (Morrison & Robinson, 1997; Robinson, 1996). Research has reported that end-users form a psychological contract and expect these systems to perform the task in good faith, i.e., neutrally and accurately, when they delegate a task in the context to a system that provides recommendations (Morrison & Robinson, 1997; Pavlou & Gefen, 2005; W. Wang & Wang, 2019). In the context of AI-augmented decision-making systems, studies indicate that end-users perceive AI systems as fair before interacting with them, as they believe that the systems' decisions are objective (Araujo et al., 2020; Sundar, 2020). Thus, end-users expect AI systems to make decisions based on objective data-driven criteria that are equal for everyone (Martin, 2019; Pethig & Kroenung, 2020).

Hence, when end-users detect a bias, they can perceive that the AI system has violated a psychological contract. Literature indicates that if a psychological contract is violated, end-users feel deceived or betrayed (Morrison & Robinson, 1997) and lose trust in the entity (Coyle-Shapiro et al., 2019; W. Wang et al., 2018). PCV theory proposes that multiple steps occur before end-users feel betrayed and experience a PCV (Morrison & Robinson, 1997). First, end-users must experience a *perceived contract breach*, i.e., they must detect an incongruence between the obligation and the entity's performance. PCV theory proposes that the salience and vigilance of end-users serve as essential moderators before they can detect that the promise was not met. Second, end-users' interpretation of the perceived unmet promise results in a *contract violation*. In this assessment, end-users' prior beliefs about the system and its neutrality (W. Wang & Wang, 2019) play a significant role in deciding whether the system has violated a psychological contract or whether other factors can compensate for the contract breach. If no factors can compensate for the contract breach (i.e., perceived system's accuracy), end-users experience a PCV, which reduces their trust in the AI system (W. Wang et al., 2018; W. Wang & Wang, 2019).

PCV theory has previously been used to study the effect of providing explanations in the context of biased recommender systems that provide sponsored content (W. Wang et al., 2018; W. Wang & Wang, 2019). While the underlying theoretical framework and the implications on trust are consistent between this study and the work of W. Wang and Wang (2019), the direction of the hypothesis in this study differs in how explanations affect PCV. In their work, W. Wang et al. (2018) and W. Wang and Wang (2019) demonstrate that end-users form a psychological contract with recommender systems by assuming that those systems are accurate and unbiased. For instance, if sponsored content is displayed, it violates the assumption that the recommender system is neutral. Nevertheless, they propose that the negative effect of such PCV on trust can be compensated through explanations because the perceived transparency can mitigate the experience of betrayal arising from the sponsored content. W. Wang and Wang (2019) propose that increasing transparency through explanations reduces PCV because it encourages end-users to "fully accept the sponsorship practice by the biased RA" (p. 510). In this case, end-users can detect the bias without any explanation in the case of sponsored content.

However, in the context of a gender-biased AI system, this study hypothesizes a different relationship between explanations and PCV. This study assumes that end-users are unable to detect gender bias inherent in the data without an explanation. While explanations enable end-users to detect gender bias, end-users still need to evaluate whether the gender bias is indeed a violation of the psychological contract. To do so, end-users must consider in detail how the recommendation was derived, and which attributes influenced it. In this evaluation process, end-users need to weigh the relative importance of the gender bias against other considered attributes and assess whether the provided attributes can compensate for the gender bias. Hence, this study hypothesizes that explanations for a gender-biased AI system's decisions reduce trust in the AI system only if end-users perceive the AI system to be biased and experience a PCV:

H2: Explanations for a gender-biased AI system decrease end-users' trust in the AI system. This effect is mediated through perceived bias of the AI system and experienced PCV from the perceived bias.

4.2.4 Stigma Consciousness

Using the PCV theory, this study theorizes about the evaluation process that results in less trust in gender-biased AI systems. In particular, it outlines that it is necessary to consider end-users' subjective evaluations and beliefs regarding the relative importance of the bias vis-à-vis the perceived plausibility and accuracy of the other attributes displayed in explanations. However, PCV theory alone is insufficient to distinguish between end-users who weigh gender bias as a severe contract violation and those who assess it as less critical. To fill this gap, this study incorporates the social identity theory (Tajfel, 1982) as it is the most prominent theory on individuals' evaluation of discrimination. This theory proposes that individuals identify with specific social groups that form their identity. If one social group is confronted with negative evaluations and stereotypes, individuals who identify with this social group are threatened in terms of their identity (Spencer et al., 2016; Steele & Aronson, 1995).

Multiple studies have revealed that women, in particular, are subject to issues of bias and societal inequality (Dastin, 2018). However, individuals differ in terms of how strongly they associate the threatened group with their own identity and, as a result, how strongly they react to discrimination. This *stigma consciousness* is based on individuals' identification with the stereotyped group, prior experiences, demographic backgrounds, and personality (Pinel, 1999). Stigma consciousness has been demonstrated to reduce usage intentions of a new technology parallel to perceived usefulness and ease of use (Pethig & Kroenung, 2020). Further, stigma consciousness has been shown to influence the perceived fairness of AI systems (Pethig & Kroenung, 2020). For example, women perceive an algorithm for hiring decisions as less biased than a male recruiter, as they suspect a male recruiter to be more biased. Specifically, stigma consciousness determines how strong individuals respond to contextual cues that signal a potential devaluation of their identity (Steele et al., 2002). Thus, it shapes how individuals process information as individuals with prior experience of being confronted with stereotypes respond more severely to cues of potential discrimination than individuals who have not experienced stigmatization or do not identify with the stigmatized group.

Hence, this study proposes that end-users differ in how they evaluate explanations based on their level of stigma consciousness. Following the decision process outlined by PCV theory, this study hypothesizes that stigma consciousness provides a moderating effect in two steps. First, end-users who are more aware of discrimination against women are also more salient in detecting gender biases and perceive the system as more biased than end-users with low stigma consciousness. Second, end-users need not only to be salient regarding the detection of a potential psychological contract breach but also need to interpret it as an actual betrayal. This interpretation process is highly subjective, as some end-users might perceive a gender bias as a contract breach while others may not. Hence, this study proposes

that end-users who are more keenly aware of gender stereotypes in society, i.e., those with a high level of stigma consciousness, are more likely to interpret a gender bias as a PCV than end-users with a low level of stigma consciousness. Therefore, high levels of stigma consciousness can be expected to increase the likelihood end-users interpret a gender bias in the AI system as a PCV. Overall, the following moderating effects of stigma consciousness are hypothesized:

H3: Stigma consciousness moderates the negative effect of providing explanations for a biased AI system on trust.

H3a: End-users with higher stigma consciousness perceive the AI system that displays a gender bias in the explanations as more biased than end-users with lower stigma consciousness.

H3b: End-users with higher stigma consciousness experience a stronger PCV from a gender-biased AI system than end-users with lower stigma consciousness.

H3c (pre-registered as exploratory): End-users with higher stigma consciousness differ in how plausible they perceive specific decision recommendations from end-users with lower stigma consciousness.

Figure 26 illustrates the overall research model tested in two experiments. For both Experiment 1[20] and Experiment 2,[21] the hypotheses and data analysis were preregistered before data collection. Deviations from preregistration are elaborated in Appendix B1. H3c is not part of the research model.

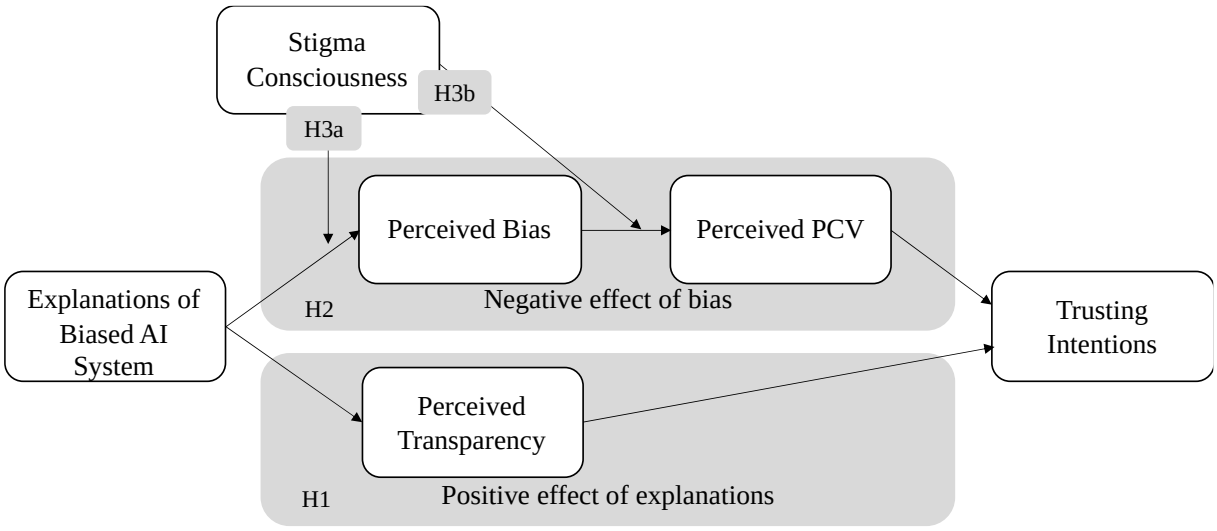


Figure 26: Overview of the overall research model.

[20] Experiment 1: https://osf.io/rht43/?view_only=9cc9a4a2884944629451b272f6f22ce0
[21] Experiment 2: https://osf.io/4v2j9/?view_only=890aafaa5b834d7a98cecd599bf5a950

4.3 Experiment 1 – Testing the Explanations' Positive and Negative Effects on Trust

A task that laypeople can perform without training was selected to test explanations' positive and negative effects on trust. Thus, this experiment investigated how end-users evaluate the recommendations of a loan forecasting AI system that recommends accepting or rejecting consumer loan applications based on a set of attributes representing the details of the loan application and the applicant's financial and personal information. Such a task has been frequently applied in ML contexts (Berendt & Preibusch, 2017). For Experiment 1, three experimental groups were utilized. In these groups, the AI system either displayed gender as an influential attribute favoring male over female applicants in its explanations (biased AI), did not consider gender as an influential attribute (neutral AI), or provided only a recommendation without any explanation (control).

4.3.1 Method

4.3.1.1 Development of the Neutral and Gender-biased AI System

An iterative process was followed to generate a neutral and gender-biased AI system by manipulating the underlying data. First, a publicly available, open-source bank loan dataset was selected (Dua & Graff, 2017) that classifies loan applications as good or bad loan risks. The dataset contains information for 1,000 loan applications, each represented by 20 attributes. Several modifications were performed to the original dataset: the names and descriptions of attributes were modified to improve comprehensibility, the attribute "personal status and sex" was simplified to reflect only "gender", and the attribute "foreign worker" was removed to avoid associations of a different bias. Second, to create the biased AI, the percentage of rejected women was increased in the dataset by changing the gender from male to female in 15 randomly selected rejected loan applications. This process was conducted iteratively to ensure that gender was sufficiently influential in the prediction but not too dominant. It also allowed for the selection of loan applications for which gender was not among the most influential attributes.

After that, two neural network classifiers were trained using the popular Python deep-learning library Keras (Chollet, 2015). Moreover, the architecture was adjusted to train the classifiers until accuracy in the neutral and gender-biased classifiers above 0.80 was achieved. The final architecture of the neural network consisted of an input layer with 66 neurons, three hidden layers with 50 neurons each, and an output layer with a "SoftMax" activation function (Bridle, 1989). The final accuracy of the neutral classifier was 0.802, while the accuracy of the biased classifier was 0.806.

Further, LIME explanations were extracted from the local linear surrogate model, where the influence of each attribute and its value on the classifier's prediction for a given instance is calculated. In these

explanations, the five most influential attributes contributing to the recommendation of each classifier were displayed. The original representation of LIME explanations was modified by using a bar plot with the attributes' names and values outside the plot, the representation of the influence of each attribute as a percentage of total influence, and a color code to display whether an attribute had a positive or negative influence on the recommendation. This new representation was evaluated during the iterative design process presented in 3.4 to ensure usability and understandability. Figure 27 presents examples of two LIME explanations in the experiment.

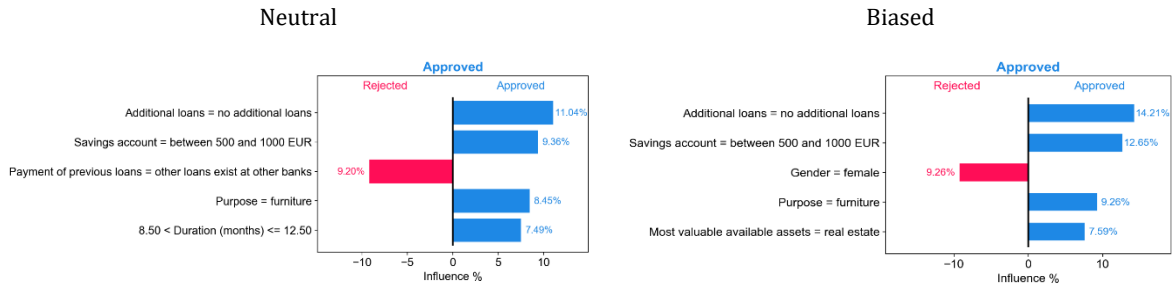


Figure 27: Example of LIME explanations for a decision recommendation of the neutral and gender-biased AI system.

4.3.1.2 Selection of Loan Applications for the Experiments

For both experiments, eight loan applications were selected in which both models provided the same decision recommendation. The actual gender of the loan applicants (male and female) and the direction of the decision recommendation (approve and reject) were balanced. The setup of the biased loan applications was pretested in an online pretest. As a result, each participant in the biased AI groups evaluated four loan applications with a gender bias visible in the explanations (female approved, female rejected, male approved, male rejected) and four loan applications with a gender bias not visible in the explanations (female approved, female rejected, male approved, male rejected). Participants in the neutral AI groups always received the same set of loan applications as the participants in the biased AI groups. However, the generated LIME explanations did not display gender as an important attribute for this group. Participants in the control group also received the same set of loan applications and decision recommendations without explanations.

4.3.1.3 Experimental Procedure and Measures

The experimental procedure consisted of four phases and was approved by the ethics committee of the University of Mannheim. Participants were recruited from the United States using the platform Prolific.

First, participants were randomly assigned to one experimental group. All participants were informed that the experiment was an evaluation experiment for a newly developed algorithm and that this experiment aimed to collect human opinions to improve the algorithm. Furthermore, participants received a detailed description of the AI system, its accuracy rate (80%), and a list of attributes the

system considered for decision recommendations. Participants in the neutral and biased AI groups also received an example of a LIME explanation with a description to help them interpret it. Comprehension checks were also included. The control group in Experiment 1 received one comprehension check, and the neutral and biased AI groups received two additional comprehension checks about the content of the sample explanation.

Second, participants assessed the eight selected loan applications. The order of those applications was randomized across all participants to avoid order effects. Participants were asked to indicate on a 7-point scale from "strongly disagree" to "strongly agree" whether they thought the algorithm's decision recommendation was plausible. Then, participants were asked to explain their decision in an open text field.

Third, after rating all loan applications, participants were asked to respond to multiple survey items about their overall impression of the loan forecasting AI system and their demographic background. Different measures were included to assess the hypotheses and control variables. Participants were asked to indicate whether they perceived the system as biased (Wallace et al., 2020), transparent (W. Wang & Benbasat, 2016), and how strongly they perceived that the system had violated a psychological contract (Pavlou & Gefen, 2005; W. Wang et al., 2018; W. Wang & Wang, 2019). Participants were also asked about their trusting intentions (McKnight et al., 2002) by inquiring whether they would recommend that bank employees use this system.

Lastly, participants' stigma consciousness (Pethig & Kroenung, 2020; Pinel, 1999) and demographic information such as gender, age, and education level were assessed. Participants were also asked about their familiarity with loans based on three items (adapted from Gefen, 2000). Furthermore, participants were requested to indicate their knowledge about algorithms and ML by selecting from four options ranging from "No knowledge" to "A lot of knowledge" (Cheng et al., 2019) and asked about participants' disposition to trust (Gefen, 2000). Except where otherwise indicated, all measurement scales are 7-point scales ranging from "strongly disagree" to "strongly agree". All measurement scales can be found in Appendix B2. After payment, participants were debriefed about the experiment's actual purpose in a separate message.

4.3.2 Results

4.3.2.1 Participants

In total, 362 participants were recruited to ensure a sufficient statistical power of 0.80 based on a power analysis conducted before the data collection. Each participant was paid £2.80 (£8.40/hr.) for their participation in the experiment, which had an average completion time of 21 minutes. Data were excluded for nine participants who failed more than one comprehension check, nine with low-quality responses in the open text field, and five because of data quality. The final sample included 339

participants, of which 109 were in the neutral AI group, 108 in the biased AI group, and 122 in the control group. Overall, the groups did not differ in terms of control variables, gender, or education. For the analyses, the variable gender was used in binary form. Demographic information and the values of the control variables can be found in Appendix B3. The neutral and biased AI groups did not differ in stigma consciousness ($p = 0.62$). Nor order effects were found for the eight loan applications.

4.3.2.2 Manipulation Check

Two independent coders assessed how often participants mentioned a gender bias in the open text field. The intercoder reliability between the coders was 99.96%. In the biased AI group, gender bias is mentioned in 150 out of the 432 cases (34.8%) for those loan applications in which a gender bias was visible. Participants in the neutral and control conditions did not mention gender bias. These results indicate that the manipulation in the biased AI group was successful.

4.3.2.3 Descriptive Statistics

Table 9 displays the descriptive statistics of all scales included in the measurement model except for the control variables for Experiment 1. The reliability and validity of these constructs were examined with CFA to ensure good model properties (see Appendix B4). Based on this approach, one item of perceived transparency was excluded. The CFA displays good psychometric quality as the comparative fit index (CFI) was at 0.98, the goodness of fit (GFI) was at 0.93, and the adjusted goodness of fit (AGFI) was at 0.91. The square root of the average variance extracted (AVE) for each variable is higher than any correlation with other variables (Fornell & Larcker, 1981). Finally, based on Harman's single-factor test (Harman, 1976), common method bias is not a concern.

Table 9: Descriptive statistics for Experiment 1.

	Group Means (SD)			**CR**	**Correlations & Sqrt AVE in parenthesis**				
Variable	**Control**	**Neutral AI**	**Biased AI**		**1**	**2**	**3**	**4**	**5**
Perceived bias	3.48	3.27	4.66	0.94	(0.92)				
	(1.54)	(1.53)	(1.57)						
Stigma consciousness	3.84	4.11	4.41	0.79	0.36	(0.77)			
	(1.71)	(1.74)	(1.71)						
Perceived PCV	3.30	2.97	3.51	0.91	0.64	0.33	(0.88)		
	(1.48)	(1.42)	(1.58)						
Perceived transparency	4.06	5.37	5.12	0.96	-0.4	-0.19	-0.6	(0.92)	
	(1.76)	(1.44)	(1.44)						
Trusting intentions	3.97	4.22	3.28	0.96	-0.68	-0.4	-0.79	0.6	(0.94)
	(1.65)	(1.81)	(1.65)						

4.3.2.4 Group Differences

An ANCOVA analysis was conducted with trusting intentions as the dependent variable and task familiarity, algorithm knowledge, ML knowledge, disposition to trust, age, gender, and education as covariates (Appendix B5). The experimental manipulation resulted in a difference in trust ($F(2) = 9.99$, $p < 0.001$). The Bonferroni post hoc analysis suggested significant group differences in trusting intentions between the neutral and the biased AI groups ($p < 0.001$) and between the biased AI and control groups ($p < 0.05$). The neutral AI and control groups did not significantly differ in trusting intentions ($p = 0.13$). Figure 28 displays the estimated marginal means of trusting intentions for Experiment 1 across all experimental conditions.

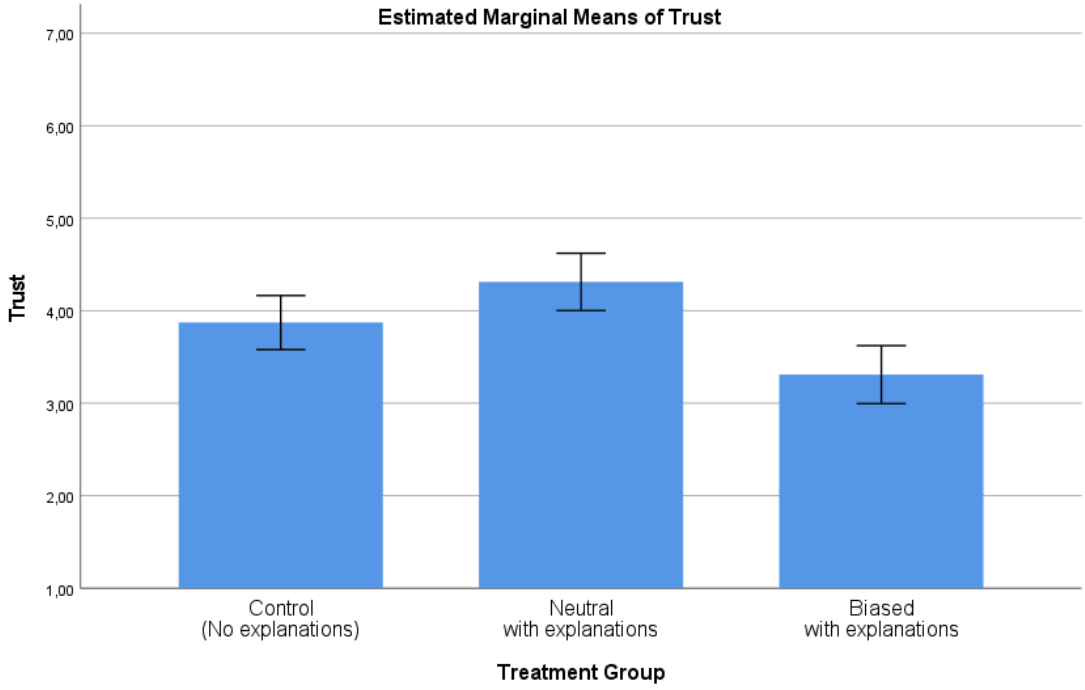


Figure 28: Estimated marginal means of trusting intentions for each group in Experiment 1.

Furthermore, an ANCOVA analysis was conducted with perceived bias as the dependent variable, and all covariates mentioned above. As expected, the groups differed significantly in perceived bias ($F(2) = 25.36$, $p < 0.001$). The post hoc analysis showed that the AI system was perceived as significantly more biased in the biased AI group than in the neutral AI group ($p < 0.001$) and the control group ($p < 0.001$). There was no difference in perceived bias between the neutral AI and control group ($p = 0.32$).

Lastly, an ANCOVA analysis was conducted with the mentioned covariates for perceived transparency as the dependent variable. Group differences were also significant for perceived transparency ($F(2) = 31.38$, $p < 0.001$). The post hoc analysis indicated that participants perceived the AI system as more transparent in the neutral AI group compared to the control group ($p < 0.001$) and in the biased AI group compared to the control group ($p < 0.001$). As expected, there were no differences in perceived

transparency between the neutral and the biased AI groups (p = 0.19). Overall, these findings indicate that the two manipulations—interacting with a gender-biased AI system and being exposed to explanations—had the intended effects on participants.

4.3.2.5 Mediation Through Perceived Bias and Experienced PCV

To test the proposed model, a moderated mediation analysis was conducted according to Hayes's (2017) procedure, using 5,000 bootstrapped samples with a 95% confidence interval (CI). The model calculated the confidence intervals (CI) of the lower (LLCI) and upper bounds (ULCI). Effects were considered significant if the CI did not include zero. The model was configured manually to account for the two pathways: (1) a moderated-mediation (perceived bias and PCV) pathway with stigma consciousness as the moderator and (2) a second mediation pathway (perceived transparency). Two moderated-mediation analyses were conducted: one treating the three groups as a continuum and a second comparing the three groups against each other using the Helmert contrast coding scheme. The latter allowed analyzing the effect of providing explanations versus not providing explanations by comparing the control group against a combined neutral and biased AI group (X1) and the effect of gender bias in the AI system by comparing the neutral AI against the biased AI group (X2). As the results of both analyses were similar, the more detailed analysis with the Helmert coding is reported in this experiment.

Supporting H1, perceived transparency positively influenced trusting intentions ($\beta = 0.30$, $t = 6.51$, $p < 0.001$). As hypothesized, providing explanations significantly increased perceived transparency (X1, $\beta = 1.30$, $t = 7.69$, $p < 0.001$). Interestingly, the neutral AI group perceived the AI system as slightly more transparent than the biased AI group (X2, $\beta = -0.38$, $t = -1.86$, $p = 0.06$). Transparency mediated the effect of explanations (X1) on trusting intentions; the bootstrapped indirect effect was significant (X1, indirect effect = 0.39, Boot95% CI = 0.24 ~ 0.57), while the effect for the neutral versus biased AI groups was barely significant (X2, indirect effect = -0.11, Boot95% CI = -0.24 ~ -0.01). Supporting H2, a higher perceived bias resulted in a higher perceived PCV ($\beta = 0.43$, $t = 4.35$, $p < 0.001$), and the perceived PCV reduced trusting intentions ($\beta = -0.66$, $t = -14.27$, $p < 0.001$).

4.3.2.6 Moderating Effect of Stigma Consciousness

Aligned with H3, the influence of the group on perceived bias was contingent upon the level of stigma consciousness; the direct effect of explanations (X1) ($\beta = -0.23$, $t = -0.54$, $p = 0.59$) and gender bias in the AI system (X2) ($\beta = 0.05$, $t = 0.10$, $p = 0.92$) on perceived bias were not significant. However, the interaction effect between the gender bias in the AI system and stigma consciousness (X2) was significant ($\beta = 0.30$, $t = 2.61$, $p < 0.01$), supporting and expanding H3a. When comparing the neutral and biased AI groups (X2), participants with low stigma consciousness ($\beta = 0.77$, $p = 0.01$, CI = 0.18 ~ 1.35) perceived the AI system as less biased than participants with a moderate ($\beta = 1.29$, $p < 0.001$, CI

= 0.89 ~ 1.68) and high stigma consciousness (β = 1.80, p < 0.001, CI = 1.28 ~ 2.34). Figure 29 illustrates the moderating effect of stigma consciousness on perceived bias for all three groups.

Contrary to the presumptions in H3b, stigma consciousness had no moderating effect on the relationship between perceived bias and PCV (β = 0.03, t = 1.20, p = 0.23). As a result, for the neutral versus biased AI groups (X2), the indirect effect of the two groups on trusting intentions through perceived bias and PCV increased significantly as the level of stigma consciousness increased. With low stigma consciousness, the effect was low (indirect effect = -0.24, Boot95% CI = -0.49 ~ -0.04), but the effect had more influence with medium (indirect effect = -0.45, Boot95% CI = -0.65 ~ -0.29) and high levels of stigma consciousness (indirect effect = -0.69, Boot95% CI = -0.98 ~ -0.43). Lastly, after considering both mediation pathways, additional variance in trusting intentions can be explained by remaining strong relative direct effects (X1: β = -0.56, p < 0.001, 95% CI = -0.81 ~ -0.30; X2: β = -0.48, p < 0.001, 95% CI = -0.76 ~ -0.20). Therefore, the findings indicate that additional variables contribute to the decrease in trust in the interaction with a gender-biased AI system.

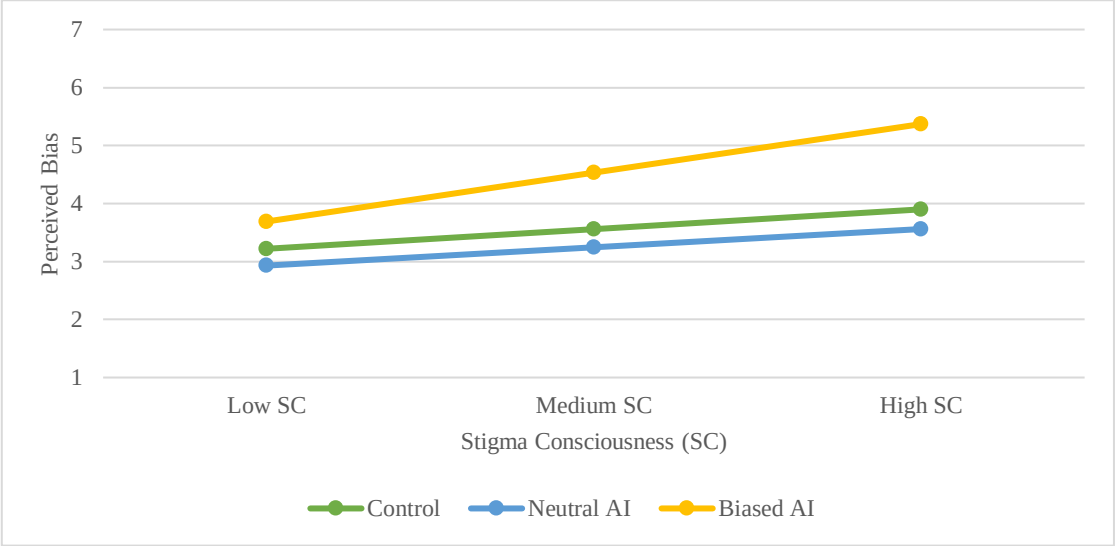


Figure 29: Moderation effect of stigma consciousness (SC) on perceived bias across the three experimental groups.

To assess the exploratory H3c, this experiment considered how participants rated the plausibility of the decision recommendation for each loan application. The biased AI group was split into two separate groups via a median-split of participants' stigma consciousness (Median = 4.33, N(low) = 51, N(high) = 57). Then, an exploratory repeated-measurement ANOVA analysis was conducted with four groups: control, neutral AI, biased AI with low stigma consciousness (biased AI/low SC), and biased AI with high stigma consciousness (biased AI/high SC). Because the sphericity assumption was violated ($\chi 2(27)$ = 235.18, p < 0.001), the Greenhouse-Geisser correction (ε = 0.83) was used. The results reveal that participants' plausibility significantly differed within subjects across the eight loan applications (F(5.84) = 68.44, p < 0.001). Because the groups strongly differed in their variance, a Games-Howell post hoc test was conducted. The biased AI/low SC group had a significantly higher rating of plausibility than

the biased AI/high SC group (95% CI = 0.06 ~ 0.83, p < 0.05) and a similar rating as the neutral AI group (95% CI = -0.25 ~ 0.47, p = 0.85). Further, the biased AI/high SC group had a significantly lower rating for the plausibility of explanations than the neutral AI group (95% CI = -0.66 ~ 0.01, p < 0.05). Figure 30 displays how the plausibility of each explanation differs between the two biased AI subgroups. As assumed, the effect of SC was especially profound in those loan applications where a gender bias was visible to participants but not in the cases in which the gender bias was hidden. These findings suggest that participants with low stigma consciousness do not respond differently to biased systems than participants interacting with neural AI systems.

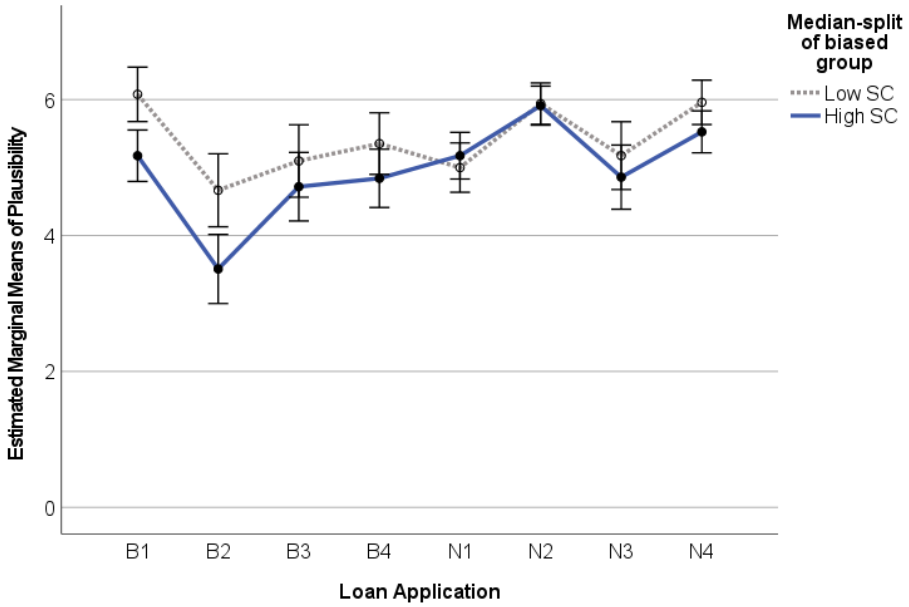


Figure 30: Plausibility of each explanation across all loan applications in the biased AI groups with median split for stigma consciousness (SC).

Notes. B = Bias visible in the explanation, N = Bias not visible in the explanation, 1 = Female, approved, 2 = Male, approved, 3 = Female, rejected, 4 = Male, rejected

4.3.2.7 Robustness Analyses

A robustness analysis was conducted by running the same analysis with the complete dataset to test if excluding participants because of data quality significantly altered the findings. The overall direction of findings was consistent with the reporting findings, as trust was the lowest for participants who interacted with a biased AI system. In the moderated-mediation model, there was no significant indirect effect of bias in the AI system through perceived bias and experienced PCV for participants with low stigma consciousness (indirect effect = -0.19, Boot 95% CI = -0.41 ~ 0.01). However, the mediation through perceived bias and PCV was significant for participants with moderate and high levels of stigma consciousness. Hence, with the complete data set, the moderating effect of stigma consciousness was stronger.

4.3.2.8 Summary of Findings of Experiment 1

The results of Experiment 1 support hypotheses H1 and H2 that explanations for a gender-biased AI system can positively and negatively affect trusting intentions. Further, aligned with H2, the negative effect of disclosing a bias through explanations on trust was mediated through the perceived bias and experienced PCV. Moreover, Experiment 1 shows that participants' stigma consciousness has a moderating effect (H3). It affects if participants perceive the system as biased, supporting H3a. However, the findings indicate that the general level of stigma consciousness does not affect if participants interpret the bias as a PCV, contrary to H3b. Further, stigma consciousness influences how participants evaluate the plausibility of the decision recommendation when the gender bias is displayed in explanations but not when it is hidden (H3c).

4.4 Experiment 2: Priming Stigma Consciousness

Experiment 2 aims to expand the findings of Experiment 1 on the moderating effect of stigma consciousness by experimentally manipulating the salience of stigma consciousness through priming. To achieve this, the same neutral and gender-biased AI systems used in Experiment 1 were also used in Experiment 2.

4.4.1 Method

4.4.1.1 Selection of Loan Applications

For Experiment 2, a different set of loan applications was selected than for Experiment 1 to increase participants' critical assessment of the explanations by providing participants with boundary decisions of the AI system. In these cases, the provided explanations consisted of three attributes favoring the decision recommendation and two opposing it. Further, loan applications with similar attributes that contributed to the decision were selected using a cluster analysis. However, there were only seven loan applications with boundary decisions in the selected cluster. Therefore, for the category "N4: male rejected without displayed bias", a loan application from another cluster was selected.

4.4.1.2 Experimental Procedure and Measures of Experiment 2

Experiment 2 differed from Experiment 1 in that following informed consent, participants were allocated to the priming or no-priming condition and the neutral or biased AI condition. Participants in the priming condition received priming of gender stereotypes following the procedure of Pinel (2004) before interacting with the AI. They were asked to rate six statements on gender stigmatization with "Yes" or "No". Two statements were taken from Pinel (2004), and four additional statements were developed based on prior research on gender stereotypes— e.g., "Women who are high achievers are

neglecting their family obligations". Pinel's full 10-item measure of stigma consciousness (2004) was included after the priming to further increase stigma consciousness.

Additionally, participants were asked to indicate their gender before interacting with the AI system. This approach has been used by prior research on psychology to increase the salience of gender stereotypes (Oyserman et al., 2007; Pinel, 2004). To align with measurement properties, the reported analyses were conducted with the three-item version of the stigma consciousness scale with the same items as in Experiment 1. Then, participants followed the procedure outlined for Experiment 1. Participants in the no-priming condition were asked to report their stigma consciousness and gender after interacting with the AI system, as in Experiment 1.

4.4.2 Results

4.4.2.1 Participants

Aiming for a power level of 0.80, 273 participants participated in the online survey distributed via Prolific. Data were excluded for three participants due to their response time, 18 because they failed two or more comprehension checks, and 15 because of low quality in the open text field. In addition to these preregistered exclusion criteria, data from 11 participants were excluded based on their responses to the priming questions that revealed few encounters with gender discrimination. The final sample consisted of 226 participants, with 62 participants in the neutral AI/no-priming group, 50 in the neutral AI/priming group, 60 in the biased AI/no-priming group, and 54 in the biased AI/priming group. Appendix B3 displays participants' demographics and responses to controls.

4.4.2.2 Manipulation Check

A t-test analysis showed that priming affects SC ($p < 0.05$), as the priming groups had a higher level of SC ($M = 4.83$, $SD = 1.44$) than the no-priming groups ($M = 4.40$, $SD = 1.65$). Furthermore, as in Experiment 1, two independent coders assessed how often participants mentioned a gender bias in their qualitative responses in the open text field (bias detection). As expected, the groups differed significantly in bias detection ($F(3) = 102.286$, $p < 0.001$). The post hoc analysis demonstrated that gender bias was more frequently detected in the biased/priming group than in the biased AI/no-priming group (95% CI= $0.06 \sim 0.16$, $p < 0.001$). These results show that the experimental manipulation was effective.

4.4.2.3 Descriptive Statistics

Table 10 presents the descriptive statistics, composite reliability, correlations, and square root of the AVE of Experiment 2. The reliability indicators measured by Cronbach's α for all items were above the minimum threshold of 0.70. The CFA displays good psychometric quality, as the CFI was at 0.99, the GFI at 0.95, and the AGFI at 0.91 (see Appendix B6). The measurement properties are similar to Experiment 1.

Table 10: Descriptive statistics for Experiment **2**.

	Group Means (SD)				**CR**	**Correlations & Sqrt AVE in parenthesis**		
Variable	**Neutral AI/ no-priming**	**Neutral AI/ priming**	**Biased AI/ no-priming**	**Biased AI/ priming**		**1**	**2**	**3**
Perceived bias	3.33	4.04	4.81	4.87	0.96	(0.95)		
	(1.54)	(1.65)	(1.59)	(1.63)				
Perceived PCV	3.33	3.57	3.56	4.33	0.94	0.61	(0.92)	
	(1.48)	(1.64)	(1.48)	(1.52)				
Trusting intentions	3.67	3.76	3.26	2.51	0.96	-0.61	-0.84	(0.94)
	(1.76)	(1.75)	(1.75)	(1.47)				

4.4.2.4 Mediation Through Perceived Bias and Experienced PCV

An assessment was performed on gender bias in the AI system impacted trusting intentions through perceived bias and PCV (H2) with a mediation analysis following Hayes' procedure (Hayes, 2017), including all covariates. Aligned with the results in Experiment 1, the negative impact of gender bias in the AI system on trusting intentions was mediated through perceived bias and PCV (indirect effect = -0.55, BootLLCI = -0.79 ~ -0.33). Further, as hypothesized in H2, perceived bias increased PCV (β = 0.55, t = 11.33, $p < 0.001$), and PCV reduced trusting intentions (β = -0.87, t = -19.30, $p < 0.001$). However, a significant direct effect remained (effect = -0.35, LLCI = -0.62 ~ -0.08), indicating that other factors influenced the effect of gender bias in the AI system on trusting intentions.

4.4.2.5 Effect of Stigma Consciousness Priming

An ANCOVA analysis was conducted to assess the influence of gender bias in the AI system and priming on trusting intentions. The results show a significant main effect of gender bias in the AI system ($F(1) = 13.09$, $p < 0.001$). Both the effect of priming ($p = 0.11$) and the interaction effect between gender bias and priming ($p = 0.14$) were not significant. Figure 31 displays the interaction effects, while Appendix B7 contains all numerical results.

To compare the four treatment groups, a second ANCOVA analysis was conducted on trusting intentions with a pairwise comparison of the treatment groups. The findings indicate that participants in the biased AI/priming group trusted the AI system less than participants in the neutral AI/no-priming group (95% CI = -1.98 ~ -0.34) and the neutral AI/priming group (95% CI = -2.03 ~ -0.27). Participants in the biased AI/no-priming group did not differ from other groups in terms of trusting intentions, and their trust levels were similar to those in the neutral AI groups. However, it is important to notice that participants in the neutral AI groups in Experiment 2 trusted the AI system less than participants in the neutral AI group in Experiment 1 (M(Experiment 1) = 4.23 SD = 1.82 vs. M(Experiment 2) = 3.71 SD = 1.75). The implications of this difference are discussed in the results and discussion summaries sections.

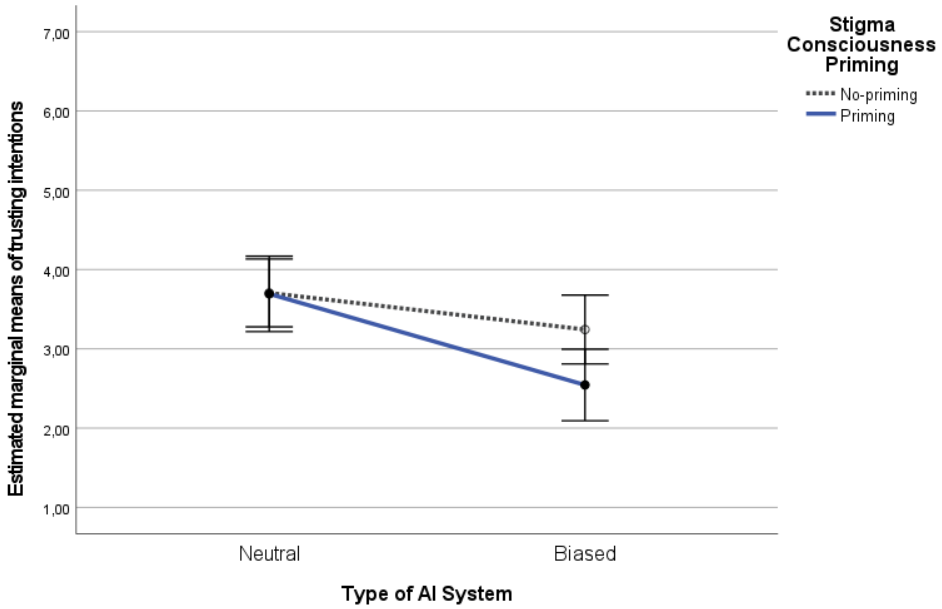


Figure 31: Results of ANCOVA analysis: interaction between gender bias in the AI system (neutral vs. biased AI) and priming (no-priming vs. priming).

To understand the moderating role of the stigma consciousness priming for the biased AI groups (N = 113), mediation analyses following the procedure of Hayes were conducted (Hayes, 2017). this experiment investigated the effect of priming on trusting intentions mediated through perceived bias and PCV, mediated only through PCV, and the direct effect. This analysis slightly deviated from the preregistration (see Appendix B1 for details). The analysis reveals that priming of stigma consciousness did not affect the perceived bias of the biased AI system ($p = 0.83$), contrary to H3a. However, the effect of priming was fully mediated by PCV alone (indirect effect = -0.37, Boot95% CI = -0.56 ~ -0.17), as priming significantly increased PCV ($\beta = 0.43$, $t = 3.69$, $p < 0.001$), extending H3b. There was no remaining direct effect of priming ($p = 0.92$).

As in Experiment 1, a repeated-measurement ANOVA analysis was conducted with the plausibility rating of each loan application as the dependent variable to test exploratory H3c (see Figure 32). Because the sphericity assumption was violated ($\chi2(27) = 109.14$, $p < 0.001$), a Greenhouse-Geisser correction ($\varepsilon = 0.85$) was used. The results show that the experimental condition significantly influenced the plausibility of explanations ($F(5.92) = 2.57$, $p < 0.05$). The biased AI groups differed slightly in their plausibility evaluation of explanations that displayed a gender bias but not for explanations that did not display this bias.

A closer analysis indicates that the difference between the biased AI groups was significant in cases B1 and B3, where female loan applicants were negatively affected by the gender bias. In these cases, participants in the biased AI/priming group rated the explanation's plausibility lower than in all other cases. Interestingly, priming also increased the sensitivity to discrimination for the neutral AI group.

Participants in the neutral AI/priming group evaluated N3 as significantly less plausible than the participants in all other groups.

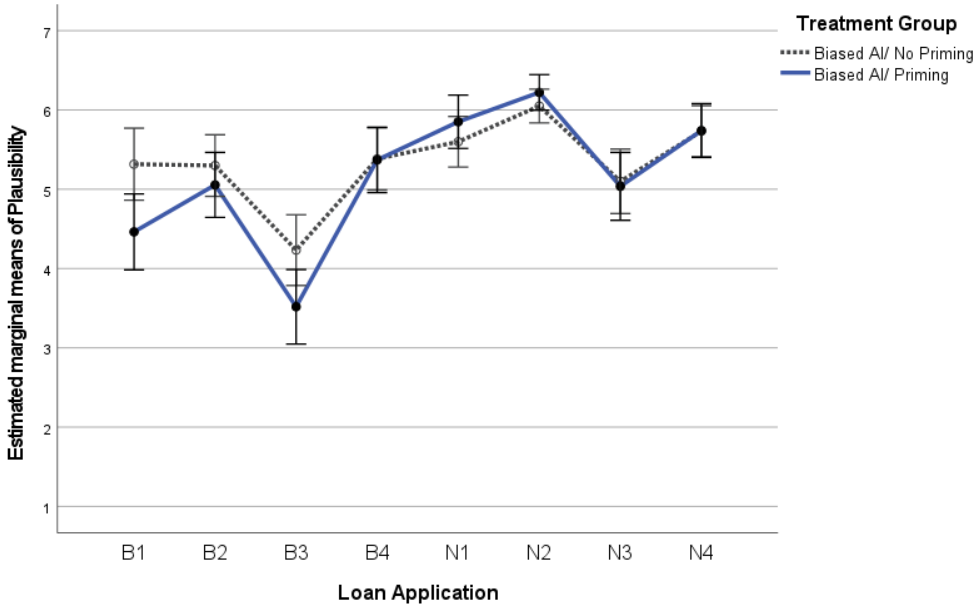


Figure 32: Plausibility of each explanation across all loan applications in the biased AI group for the two priming conditions.

Notes. B = Bias visible in the explanation, N = Bias not visible in the explanation, 1 = Female, approved, 2 = Male, approved, 3 = Female, rejected, 4 = Male, rejected

4.4.2.6 Robustness analyses

The findings were replicated with the complete data set. Trust was lower in the biased AI/ priming group than in the neutral AI/ priming and neutral AI/ no-priming groups. Also, the mediation analyses provided the same significant relationships. However, the effect size of priming is relatively small. Thus, it was essential to exclude participants with whom the priming did not work, as with those in the data set, there was no significant effect of priming on stigma consciousness. The significant results remained if non-binary participants (N= 8) were analyzed separately.

4.4.2.7 Summary of Findings of Experiment 2

Experiment 2 confirms the negative effect of providing explanations for a biased AI system on trust, mediated through perceived bias and experienced PCV (H2). Further, Experiment 2 expands insights into the moderating role of stigma consciousness (H3). Specifically, participants in the neutral AI/ no-priming group had a similar level of trust as those in the biased AI/ no-priming group. Further, priming of stigma consciousness affected if participants perceived the AI system as biased (H3a), however, in a different way than expected. Notably, there was no additional effect of priming on perceived bias if the system was biased. Instead, priming increased the perceived bias in the interaction with the neutral AI system. Moreover, stigma consciousness priming increased the likelihood of experiencing a PCV (H3b), expanding the findings of Experiment 1. Hence, priming reduced the threshold when participants

interpreted the bias as PCV. Lastly, similar to Experiment 1, priming of stigma consciousness did affect how participants evaluated the plausibility of each explanation (H3c); however, only in those cases in which the gender bias was explicitly displayed to participants and in which females faced negative consequences by the bias.

4.5 Discussion

4.5.1 Summary of Findings

This study demonstrates that end-users' social identity, i.e., their stigma consciousness, is an important contingency factor that influences if explanations positively or negatively impact their trust. Participants with a high awareness of gender discrimination (i.e., a high level of stigma consciousness) and participants primed in this direction agreed less with the AI system's decisions if the bias was displayed in the explanations. As a result, they experienced a stronger PCV and trusted the AI system less. In contrast, participants with low stigma consciousness perceive the AI system as equally unbiased and trustworthy as participants who interact with a neutral AI system. Further, if explanations do not display the bias to participants, they perceive the recommendations as equally plausible regardless of their level of stigma consciousness. Overall, this study reveals that providing explanations does not entirely prevent the negative consequences of biases in AI systems and that end-users' beliefs influence their evaluation of explanations.

4.5.2 Theoretical Contributions

This study provides three significant theoretical contributions: (1) contextualization of the PCV theory, (2) insights into how end-users cognitively evaluate explanations, and (3) an extension of the current understanding of the impact of explanations on trust.

First, the PCV theory was contextualized to AI-augmented decision-making with a gender-biased XAI system. Moreover, theoretical insights were developed on contingency factors that determine whether end-users experience a PCV from the decision recommendations of a gender-biased AI system. Prior work has demonstrated that end-users develop a psychological contract with a biased recommender agent (see W. Wang et al., 2018; W. Wang & Wang, 2019). Specifically, W. Wang and Wang (2019) showed that experiencing a PCV is contingent upon the general prior knowledge about sponsorship practices and that disclosing the bias can help reduce the experienced PCV. However, this study reveals a different, conflicting effect of explanations on PCV, as it shows that providing explanations can actually increase the experienced PCV.

Based on this conflicting effect, two theoretical conclusions are derived: On the one hand, it is necessary to theorize on the type of bias that results in a PCV, specifically, whether the bias is data or outcome-

based. While W. Wang and Wang (2019) considered sponsored content a biased outcome end-users can easily detect without explanations, this study investigated a bias inherent to the data used to train an AI system, which end-users can only detect by critically assessing the attributes displayed in the explanation. Furthermore, in contrast to the work of W. Wang et al. (2018, 2019), a social bias was considered, which relates to end-users' subjective evaluation of gender discrimination and social inequality. Future research should determine how different types of biases affect PCV. Specifically, future research should evaluate whether end-users experience more PCV from specific biases more heavily discussed in society (such as gender or race) or less prominent biases (such as nationality or age).

On the other hand, it is necessary to theorize about the role of end-users' social identity regarding their PCV experience. In particular, this study shows that end-users' stigma consciousness affects their interpretation of whether an AI system is biased and the threshold at which they perceive a psychological contract violation. However, more research is needed to understand what affects the threshold of experiencing a PCV from a biased AI system (H3b). This study shows that the effect of priming in the gender-biased AI system (Experiment 2) is different from the effect of general stigma consciousness if comparing the neutral versus the biased AI system (Experiment 1). Therefore, future research should consider in more detail the threshold needed to experience a PCV and consider additional factors related to how end-users differ in this evaluation process, such as the degree to which they are personally affected by or benefit from the discriminating AI system.

Second, this study contributes to research on cognitive evaluations of explanations. Specifically, studies in AI research propose that end-users' mental model influences whether they seek systems' explanations and how they evaluate them (Miller, 2019). Therefore, the underlying explanation evaluation can be influenced by end-users' personal biases (Miller, 2019), and explanations can persuade end-users to change their mental models (Bussone et al., 2015; Nourani et al., 2021; Poursabzi-Sangdeh et al., 2021). This study empirically demonstrates that end-users' evaluation of explanations provided by a gender-biased AI system is influenced by their stigma consciousness. Hence, end-users seek consistency between their own attitudes and the attributes displayed in explanations. For end-users that place low importance on gender discrimination (i.e., with a low stigma consciousness in Experiment 1 and no-priming in Experiment 2), their perceived plausibility is not significantly different between the decision recommendations displaying a bias and those not displaying it.

Moreover, in line with the theory on explanatory coherence (Thagard, 1989), the experiments in this study indicate that end-users' evaluation process can be influenced by the overall consistency of attributes displayed to them. In Experiment 2, participants were provided with explanations for boundary decisions, which displayed three attributes favoring the decision recommendation and two opposing it. These explanations resulted in lower end-users' trust in the neutral AI system compared to Experiment

1. Interestingly, this change in the experimental setting resulted in no difference between the neutral AI and the biased AI/no-priming group. The biased AI/priming group had the lowest trust in the AI system. These results suggest that end-users' evaluation of explanations is highly subjective and can easily be influenced by the attributes displayed in those explanations, the context (priming), and how much general importance end-users place on specific attributes. Hence, end-users evaluate not only the internal consistency of attributes displayed in explanations but also the consistency between their underlying assumptions about the importance of attributes and the displayed attributes. Therefore, future research should take a motivated reasoning perspective (Kunda, 1990) to theorize further how end-users differ in their evaluation of explanations. Specifically, some end-users might be motivated to accurately evaluate AI systems' explanations. In contrast, others might tend to a more biased evaluation seeking consistency between their attitudes and the provided explanations.

Third, this study contributes to theories on trust formation, particularly regarding the role of explanations on trust. Prior research has mainly considered the positive impact of explanations on trust by focusing on neutral AI systems (Dhaliwal & Benbasat, 1996; Gregor & Benbasat, 1999). Specifically, explanations increase knowledge-based trust because they enable the validation of the provided recommendations. Therefore, explanations have been considered as a design mechanism to foster more systematic and deliberate reasoning processes. However, the findings in this study demonstrate that in the context of biased AI systems, explanations have two opposing effects on trust: On the one hand, providing explanations can increase trust through fostering perceived transparency; on the other hand, providing explanations can decrease trust through increasing the likelihood that end-users will experience a PCV, which can be considered to be a highly emotional response. In the experiments of this study, participants' responses in the open text field indicate that some end-users with high stigma consciousness reacted very emotionally when detecting a gender bias. They provided comments reflecting outrage and anger toward the discriminating recommendations and the AI system. Considering all this, this study's findings indicate that more contextualized research is needed to understand how explanations affect trust and evaluate whether they can even damage trust in the AI system by exposing potential biases.

4.5.3 Practical Contributions

Overall, this study has two significant practical contributions. First, this study reveals that it is necessary to consider human cognitive processes in implementing AI systems. One key idea is to establish humans-in-the-loop to monitor and correct AI systems' mistakes through human knowledge and human decision-making rules. Especially in situations where the AI makes predictions under high levels of uncertainty, it is crucial that end-users critically assess the provided explanations in detail to correct the AI system's judgment. However, recent findings show that involving humans does not necessarily result in more accurate decisions because end-users may misjudge whether to delegate a specific decision to an AI

system (Fügener et al., 2021) and may not be able to correctly assess whether a provided recommendation that conflicts with their own judgment is correct or incorrect (Jussupow et al., 2021). This study provides additional experimental evidence for such challenges of human-AI collaboration by demonstrating that it is necessary to carefully consider end-users stigma consciousness to avoid the negative consequences of biases in AI systems. To prevent biased AI systems from making discriminating decisions, practitioners should increase awareness of how the personal attitudes of end-users can shape interactions with AI systems and carefully reflect on how to combine the abilities and weaknesses of automatic bias detection tools with human knowledge.

Second, this study indicates that practitioners need to consider the implicit attitudes of end-users when designing XAI systems, as it is essential to consider the amount of information disclosed by explanations. This study demonstrates that in the context of a gender-biased AI system, explanations alone are not sufficient for end-users to identify biases in an AI system and that, in fact, explanations can have adverse effects that prevent end-users from detecting these biases. Therefore, practitioners should consider developing adaptive XAI systems that provide context-aware explanations (Adadi & Berrada, 2018). Adaptive explanations based on end-users' implicit attitudes could help increase end-users' awareness of biases that can occur in a given context. Such adaptive XAI systems would improve human-AI collaboration and reduce the negative impact of biases in AI systems.

4.5.4 Limitations and Future Research

This study has some limitations that offer further opportunities for future research. First, this study considered the decision-making process of laypeople in an experimental loan decision-making task. The research goal of this study was to investigate general cognitive mechanisms regarding how end-users differ in their evaluation of gender-biased AI systems rather than examining how loan managers make loan forecasting decisions. Thus, the experimental context must be considered as a boundary of this study's findings and interpretation. It is important to note that prior research has shown that experts differ in their evaluation of explanations from novices (Arnold et al., 2006). Moreover, experts tend also to compare AI systems' decision recommendations and reasoning against their own knowledge and judgment. If those two assessments diverge, experts will likely follow their own judgment. In contrast, end-users with less expertise tend to be over-reliant on AI systems' recommendations and are likely to follow those recommendations if they diverge from their own judgment (e.g., Jussupow et al., 2021). Nevertheless, it is not unlikely that stigma consciousness would similarly affect evaluations of the AI system in AI-augmented decisions made by human resources, medicine, and finance experts.

Second, Experiment 2 provides another interesting insight that could stir future research: Participants who were primed and interacted with the neutral AI system were more sensitive to a potential bias than participants who were not primed. This behavior is not fully justified by the experiment design and collected data because participants perceived the neutral AI system as biased even though it was not.

However, this study does not provide enough empirical data to test for this oversensitivity to AI biases and assess its potential consequences. Future research could investigate in more detail whether and how often end-users perceive AI systems as biased even when they are not, which causes them to disagree with them.

4.5.5 Conclusion

This study shows that explanations in a biased AI system have two conflicting effects on trust. On the one hand, they increase trust through perceived transparency. On the other hand, they reduce trust as end-users perceive the system as biased and experience a violation of a psychological contract. In two experiments, this study demonstrates that stigma consciousness is an important moderator that influences how explanations are cognitively evaluated, whether end-users perceive an AI system as biased, and the threshold when they experience a PCV.

5 Study III: Designing Interactive XAI Systems for End-users[22]

5.1 Introduction

Due to the high performance that artificial intelligence (AI) systems have achieved across a wide range of applications, they are increasingly being deployed in high-stake domains with the expectation of improving decision-making quality and efficiency (Binns et al., 2018). Nevertheless, despite their success, it has been shown that AI systems are prone to replicate biases, which results in unfair decisions that can have considerable consequences for individuals (Pfeuffer et al., 2023). While humans can also fail in their judgment and provide biased decisions, asking them for a rationale and holding them accountable is possible (Binns et al., 2018). In contrast, AI systems building on machine learning (ML) models can be extraordinarily complex and difficult to understand or audit (Dodge et al., 2019). The effectiveness of AI systems is limited by their ability to explain their decisions to human users (D. Wang et al., 2019). In particular, for AI systems deployed in high-stake applications, end-users who can be potentially affected by AI systems' decisions must understand the logic behind these decisions to trust and accept them (Fernández-Loría et al., 2022). As a result, regulations such as the European Union General Data Protection Regulation (GDPR) have been implemented to ensure users' "right to explanations" of all decisions made or supported by AI systems (Goodman & Flaxman, 2017).

To address AI systems' lack of transparency, many researchers and practitioners have resorted to the field of explainable artificial intelligence (XAI). Research in XAI aims to support human understanding and trust in AI systems by developing models that explain AI decisions to end-users in non-technical terms (Diakopoulos et al., 2017). In recent years, extensive research in XAI has developed many innovative explainability methods to provide explanations of AI systems (Carvalho et al., 2019). However, most of these XAI methods have focused on providing static explanations, such as highlighting relevant features through static visualizations (Liu et al., 2021). However, providing static explanations represents a one-way communication from AI systems to users that limits the amount of information conveyed to them, which can, in turn, result in an insufficient understanding of how these systems make decisions (Cheng et al., 2019; Liu et al., 2021).

Therefore, researchers have argued for enhancing explainable AI systems by allowing users to explore explanations interactively (Lombrozo, 2006; Miller, 2019). Recently, there have been efforts to explore how interactive XAI systems should be designed to improve their transparency and end-users'

[22] This chapter is based on the following paper (Meza Martínez & Maedche, 2023). The data analysis and experimental material are found in the following GitLab repository and RADAR archive:
https://git.scc.kit.edu/h-lab/research/2095_meza_miguel_interactive_xai
https://radar.kit.edu/radar/de/dataset/FnqSoDiRXtHLqXqP

understanding of these systems' decisions (Liu et al., 2021). However, most studies have rather focused on developing interactive XAI systems for data scientists or domain experts (e.g., Hohman *et al.*, 2019; Spinner *et al.*, 2020). The explanations provided by interactive XAI systems are often too complex for end-users, making them very challenging to understand (Cheng et al., 2019; Miller, 2019). Therefore, it is necessary to design interactive XAI systems that provide explanations for end-users that support their understanding of AI systems' decisions. As a result, this study aims to address the following research question:

***RQ**: How to design interactive explainable artificial intelligence (XAI) systems to help end-users to better understand AI systems' decisions?*

This study relied on the design science research (DSR) methodology and existing design knowledge in literature to design and develop an interactive XAI system prototype to address this research question. This prototype shows explanations based on SHAP, an XAI method that provides feature attribution scores using Shapley values from game theory (S. M. Lundberg et al., 2017). This study presents the result of Cycle 1 of the DSR project, with an initial evaluation of the prototype and interviews with end-users. This study contributes to design knowledge for XAI systems by demonstrating how interactive explanations can give end-users more control over the information they receive and help them better understand how AI systems make decisions. Moreover, this study offers practitioners guidelines for designing and developing interactive XAI systems for end-users, as well as a GitHub repository with the implementation of the interactive XAI prototype and the software architecture design. [23]

5.2 Related Work

Contemporary AI systems can be extraordinarily complex and difficult to understand (Dodge et al., 2019; Janiesch et al., 2021). They are designed to process large amounts of data to perform a complex optimization process for a specific performance measure. As a result, AI systems are often considered "black boxes" where only their output is available to users (Rudin, 2019). This lack of transparency leaves the inner working mechanisms behind their decisions unclear to users. The inability of AI systems to explain their decisions to users is a critical limitation to their adoption and effectiveness (D. Wang et al., 2019). In particular, it is very challenging for AI systems deployed in high-stake applications to scrutinize them and identify potential biases that can have considerable consequences for individuals (Binns et al., 2018).

Providing explanations of AI systems' decisions has been proposed as a helpful means to increase the transparency of AI systems and enable users to understand the reasons behind these systems' decisions (Binns et al., 2018; D. Wang et al., 2019). Multiple studies have shown that providing explanations

[23] https://github.com/miguelmezamartinez/interactive_xai_system

improves users' trust (W. Wang & Benbasat, 2007; Yang et al., 2020). Moreover, prior work has found that explanations can increase the likelihood that users agree with AI systems' decisions (Liu et al., 2021; Yeomans et al., 2019). As a result, explainability is becoming a critical component of trustworthy AI systems (AI HLEG, 2020). According to Ribera and Lapedriza (2019), the requirements for explanations depend on the target audience. They argue that it is necessary to identify the target users, their goals, background, and relationship to the system to design adequate explanations that ensure proper understanding.

In this light, there has been a recent surge of interest in XAI among scholars and practitioners seeking to increase the transparency of AI systems (Miller, 2019). As a result of the extensive research performed in different communities over the last few years, many innovative XAI methods have been developed. For instance, some methods extract easily interpretable rules from the predictive model and present them to users as an explanation of the model's decision (e.g., Jian et al., 2000) Alternatively, others highlight regions of an image to indicate which pixels were influential in the model's prediction (e.g., J. Zhou et al., 2017). Several studies have surveyed the literature to provide a detailed overview of XAI by presenting the different developed methods (Adadi & Berrada, 2018; Carvalho et al., 2019; Guidotti, Monreale, Ruggieri, Turini, et al., 2018).

So far, most XAI methods provide static explanations that only reveal pre-defined information about AI systems' decisions (Liu et al., 2021; Ribera & Lapedriza, 2019). For example, some methods highlight sections of an input text to indicate the importance of certain features towards the system's prediction (e.g., Bansal et al., 2021) while others present a set of influential training examples (e.g., Koh & Liang, 2017). Alternatively, other methods rely on static visualizations to show each feature's influence on the systems' decision (e.g., Ribeiro, 2016). These static explanations represent a one-way communication from AI systems to end-users, which can limit the information conveyed and may result in an insufficient understanding of how these systems make decisions (Cheng et al., 2019; Liu et al., 2021). Specifically, studies have found that end-users perceive static explanations as not transparent enough as they do not allow them to investigate further the factors that influence a given decision (Sun & Sundar, 2022).

Interactivity has been identified in the literature as an essential component of XAI systems that can help address the challenges posed by static explanations (Lombrozo, 2006; Miller, 2019). Providing interactive explanations allows users to explore the system's behavior, giving them more control over the information they receive and a sense of agency that can promote trust in AI systems (Sun & Sundar, 2022). Recently, there have been efforts to start exploring how to design interactive XAI systems. In particular, some studies have focused on incorporating research in information visualization as it excels at knowledge communication due to the extensive work investigating how to transform abstract data into meaningful representations over hundreds of years (Friendly, 2008). For instance, Hohman *et al.* (2019) developed an interactive XAI system for data scientists that allows them to explore the factors

influencing the decision of an individual instance or a group of instances and to search and compare the decision for similar instances. Their system relies on generalized additive models (GAMs) (Friedman, 2001) to generate explanations represented by interactive plots for each feature that data scientists can explore to observe the feature's impact on the system prediction. Meanwhile, Spinner et al. (2020) developed an interactive XAI system to support users in developing and debugging ML models. Their system allows users to explore visual explanations from multiple XAI methods to support model understanding, diagnosis, and refinement.

However, there have been critiques that interactive XAI systems developed so far are designed for users with a solid understanding of statistical and ML concepts (Cheng et al., 2019; Miller, 2019). For example, some of these approaches rely on diagrams such as scatter plots, area under the curve (AUC), or precision-recall graphs, which are known to be hard to understand for end-users (e.g., Amershi et al., 2015; Cabrera et al., 2019). While data scientists are familiar with these concepts, end-users often do not have the necessary knowledge to understand these interactive XAI systems' explanations. Therefore, researchers have called for designing interactive XAI systems that consider end-users' needs to support their understanding of AI systems' decisions (Cheng et al., 2019; Liu et al., 2021; Miller, 2019).

5.3 Research Method

To design an interactive XAI system for end-users, this study followed the DSR methodology by Peffers et al. (2007). This methodology allowed the development of an interactive XAI system by proposing design principles, instantiating them in a prototype, and evaluating it with end-users. Figure 33 presents the overall DSR project consisting of two cycles. The focus of this paper is on the finalized Cycle 1.

This DSR project relied on previous work investigating how end-users engage with static explanations from XAI methods (e.g., Binns *et al.*, 2018; Dodge *et al.*, 2019; El Bekri, Kling and Huber, 2019; Hase and Bansal, 2020), as well as research exploring how to design interactive XAI systems (e.g., Hohman *et al.*, 2019; Spinner *et al.*, 2020). In Cycle 1 of this DSR project, the evaluation results from these studies were analyzed to comprehend how explanations help end-users understand AI systems' decisions. Furthermore, several challenges end-users face when interacting with AI systems due to their lack of explicit interactive explainability designed for them were identified. Afterward, relying on design knowledge from the XAI literature, two meta-requirements of interactive XAI systems for end-users were derived. Then, four refined design principles were proposed to address these meta-requirements, and three design features based on these principles were implemented in an interactive XAI system prototype. As a last step, an evaluation study and interviews with end-users were conducted.

As part of this DSR project presented in Figure 33, it is planned to conduct one additional cycle to further improve the design of the interactive XAI system prototype. In cycle 2, after reviewing the

evaluation results of cycle 1, the plan is to refine the design principles and develop a second prototype of the interactive XAI system. Then, this prototype will be evaluated in an experimental study to quantitatively analyze how the interactive XAI system affects end-users' understanding and trust.

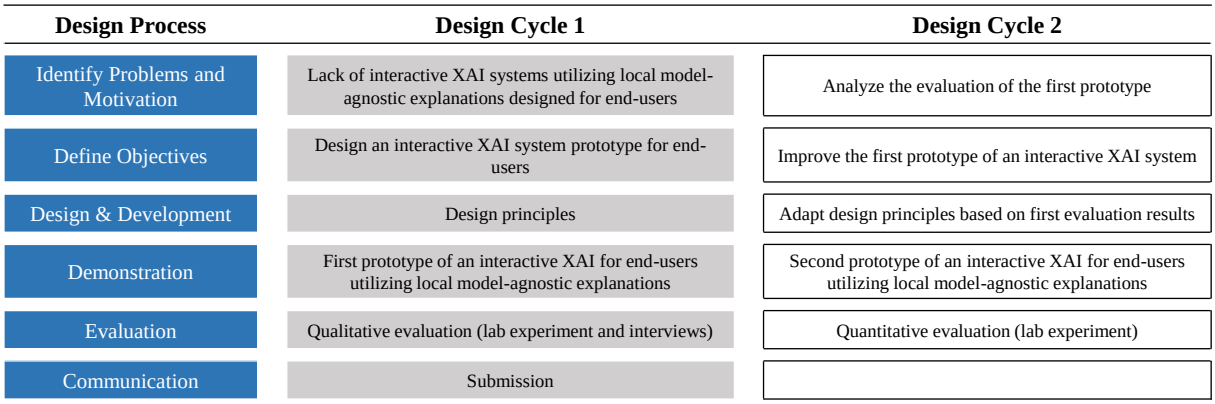


Figure 33: Overall DSR project (adopted from Peffers et al. 2007).

5.4 Conceptualization

5.4.1 Problem Awareness and Meta-Requirements

The first meta-requirement (MR1) refers to offering end-users explanations they can really understand. Even though extensive research has focused on developing XAI methods, there is no sufficient empirical evidence on whether the explanations these methods provide are understandable to end-users (Cheng et al., 2019). There is strong criticism that most of these explanations are based on researchers' and practitioners' intuition instead of a deep understanding of what end-users need (Adadi & Berrada, 2018; Miller, 2019; Ribera & Lapedriza, 2019). As a result, many of these explanations require a deep technical understanding of statistical and ML concepts (Cheng et al., 2019; Miller, 2019). Moreover, the quality of explanations generated by these methods is often evaluated using a mathematical definition of interpretability without any end-user evaluation (Doshi-Velez & Kim, 2017; D. Wang et al., 2019). Besides, most research efforts investigating how to design interactive XAI systems have instead focused on understanding the requirements these systems must fulfill to assist data scientists or domain experts (e.g., Hohman *et al.*, 2019; Spinner *et al.*, 2020; Narkar *et al.*, 2021). Therefore, it is necessary to investigate which type of explanations can be integrated into interactive XAI systems to help end-users understand how the system makes decisions.

***MR1**: An interactive XAI system should be able to provide end-users with understandable explanations that reveal in non-technical terms the reasons behind its decisions.*

The second meta-requirement (MR2) refers to the system's capacity to allow end-users to request additional information regarding its decision logic. Several studies have found that explanations are often insufficient for users to fully understand the logic behind the system's decisions (e.g., Ribeiro,

Singh and Guestrin, 2018; Kaur *et al.*, 2019; Hase and Bansal, 2020). Each XAI method relies on a different approach to provide explainability of AI systems. As a result, due to how explanations are generated, they focus on describing certain aspects of a given decision (Cheng et al., 2019). Some studies have found that end-users can perceive the system as not transparent enough due to the limited information provided by some of these explanations (Sun & Sundar, 2022). Therefore, researchers have argued for enhancing XAI systems to give end-users more control over the explainability information they receive (Krause et al., 2016; Miller, 2019).

***MR2**: An interactive XAI system should allow end-users to request additional information if explanations are insufficient to understand the decisions.*

5.4.2 Design Principles

To address the two derived meta-requirements (MRs) presented in the previous section, this study proposes design principles (DPs) for interactive XAI systems to help end-users understand AI systems' decisions. Interactive XAI systems should provide end-users with understandable explanations that reveal information about how the system makes decisions according to their needs (MR1). Regarding their scope, XAI explanations are classified as either global or local. Global explanations provide a comprehensive and holistic description of the model behavior across all instances for a given dataset (Guidotti, Monreale, Ruggieri, Turini, et al., 2018). This type of explanation is better suited for researchers or practitioners trying to improve the predictive model's performance or domain experts looking to learn from the system to improve their decision-making (Doshi-Velez & Kim, 2017; Ribera & Lapedriza, 2019). In contrast, local explanations describe how a particular system's decision was made by considering the vicinity of the instance to be explained (Molnar, 2020). Local explanations can help justify a system's decision to end-users, for whom this decision can have a personal or economic impact (Doshi-Velez & Kim, 2017; Ribera & Lapedriza, 2019).

***DP1**: Provide local explanations that reveal to end-users how a specific system's decision was made.*

The type of explanation an interactive XAI system provides is another critical factor in delivering adequate information to help end-users understand AI systems' decisions (MR1). XAI methods rely on different approaches to generate explanations, which influences the information disclosed by explanations. Research has shown that explanations from some of these XAI methods might not be sufficient for end-users to understand the reasoning behind decisions (Dodge et al., 2019; Hase & Bansal, 2020). For instance, Binns et al. (2018) found that end-users get frustrated with counterfactual explanations as they do not reveal which features had more influence on a decision. In this line, Doshi-Velez and Kim (2017) argue that explanations should provide information regarding the factors used in a decision and their relative importance.

***DP2**: Provide explanations that disclose the factors influencing each decision and their relative weights.*

How explanations are presented to end-users also plays an essential role in their cognitive process to analyze the information they contain (MR1). In XAI research, textual explanations and visual charts are the two main approaches to presenting explanations to end-users. Research has found that visual representations help end-users understand XAI explanations. For instance, Cheng *et al.* (2019) found that explanations in the form of interactive visualizations helped to improve end-users' objective comprehension of the logic behind the system's decisions. Furthermore, Szymanski et al. (2021) found that end-users prefer more visual explanations than textual explanations because these provide an easier way to obtain an overview of the factors influencing a decision.

Nonetheless, research has also found that end-users can often misinterpret visual explanations when they are too complex or lack details due to poor design (Kaur et al., 2019; Szymanski et al., 2021). Several studies have found that lengthy and complex explanations are harder to understand for end-users and can overload their cognitive abilities (Narayanan et al., 2018; Poursabzi-Sangdeh et al., 2021). To reduce the complexity of explanations, many researchers have resorted to limiting the number of factors presented to end-users by showing only the most relevant influencing a decision (e.g., Binns *et al.*, 2018; Ribeiro, Singh and Guestrin, 2018; Hase and Bansal, 2020). However, such strategies can also result in counterproductive effects as end-users would have only limited information on the system's inner workings (MR2). An alternative approach is to utilize interactive visualizations that provide an overview of the most relevant factors influencing a decision while allowing end-users to request details about the additional factors.

DP3*: Provide interactive explanation visualizations that offer an overview of the most important factors influencing a decision and allow end-users to request details regarding the additional factors.*

XAI systems should allow end-users to request additional information about its logic (MR2). Nonetheless, many of the XAI methods proposed in the literature generate only static explanations that are insufficient to help end-users understand how AI systems make decisions due to the limited information they provide (Abdul et al., 2018; Cheng et al., 2019; Ribera & Lapedriza, 2019). An interactive user interface has been proposed to empower end-users to explore and investigate how an AI system makes decisions. One strategy utilized in the literature consists of allowing users to modify the input feature values to observe how the system's decisions and explanations change accordingly (e.g., Krause, Perer and Ng, 2016; Cheng *et al.*, 2019; Hohman *et al.*, 2019). This interactive interface can enable end-users to evaluate counterfactual scenarios that reveal a causal relationship between the feature changes and the model decision (Molnar, 2020). Studies have found that interactive interfaces implementing this strategy can improve end-users' understanding of how AI systems make decisions (Cheng et al., 2019).

DP4: *Provide an interactive user interface that allows end-users to explore how changes in the input features affect AI systems' decisions.*

5.4.3 Prototype Implementation

To design and implement an interactive XAI system prototype, this study proposes design features (DFs) that represent specific system capabilities that aim to satisfy the proposed design principles (Meth et al., 2015). To instantiate the prototype, the interactive XAI system was developed for the bank loan application domain, which is commonly used in XAI research because it involves the notion of trust in AI systems (Adadi & Berrada, 2018; Aggarwal et al., 2019; Binns et al., 2018; Chakraborty et al., 2020). The following section presents the domain and describes the dataset and model used by the interactive XAI system prototype. Afterward, the DFs are described in detail to clarify how they help to address the design principles.

The bank loan application domain was selected in a scenario where an AI system predicts the decision to approve or reject loan applications. This scenario has been widely used in XAI research because end-users are familiar with the process of applying for a loan at a bank and because it allows researchers to investigate the notion of trust in the system (Binns et al., 2018; Chakraborty et al., 2020; Gurumoorthy et al., 2019). Moreover, this is considered a high-stake domain, where the decisions made by an AI system can significantly impact loan applicants (Binns et al., 2018). To train the predictive model, a publicly available, open-source dataset was used with 1,000 instances of bank loan applications and their corresponding decision (700 approved and 300 rejected). Each loan application is represented by 20 features describing the details of the loan application and the applicant's financial and personal information. The original dataset was modified by adjusting the features' names and descriptions and removing the two sensitive features, "personal status and sex" and "foreign worker". A neural network trained using the Python library Keras (Chollet, 2015) was used as the predictive model. A Synthetic Minority Oversampling Technique (SMOTE) technique was incorporated to address the class imbalance in the training data (Chawla et al., 2002). Categorical features were one-code encoded, and continuous features were min-max scaled. A grid parameter search was performed to find the best hyperparameters and architecture. The architecture with the highest score consisted of 2-hidden layers, each with 65 and 33 neurons. The predictive model had an accuracy of 0.77 and an f1-score of 0.83.

To satisfy DP1 and DP2, explanations were provided using the explainability method SHAP presented in Section 2.4.3 (S. M. Lundberg et al., 2017). SHAP is a method that provides local and global explanations by providing feature attribution scores using the concept of Shapley values (Shapley, 2016). Shapley values, which have a solid theoretical foundation in game theory, compute how the influence on the model's prediction is fairly distributed among the features used by the model (Molnar, 2020). According to Lundberg et al. (2017), SHAP builds on the concept of the popular method LIME (Ribeiro et al., 2016b) to build a linear regression model with Shapley values as weights, which indicate how much influence each feature had on the system's decision. SHAP explanations are contrastive because Shapley values are calculated from all the possible feature value collisions across all dataset

instances (Molnar, 2020). As a result, the prediction of one instance can be compared against the predictive model's average prediction.

Moreover, SHAP is a model-agnostic method that can generate explanations for any underlying predictive model. In contrast to model-specific and model-class-specific methods that provide explanations to only one predictive model or a specific model family (Sokol & Flach, 2020), model-agnostic methods offer great flexibility and scalability in their implementation due to their decoupling of explainability from the prediction (Ribeiro et al., 2016b). Despite model-agnostic methods' benefits, most studies investigating how to design interactive XAI systems have focused on developing and evaluating systems that utilize non-model-agnostic methods to provide explainability (e.g., Cheng *et al.*, 2019; Sevastjanova *et al.*, 2021; Guo *et al.*, 2022).

DF1*: Provide local model-agnostic explanations based on the XAI method SHAP, which relies on the concept of feature importance to explain how features influence the system's decision.*

SHAP has gained popularity in research and practice due to the unique consistency and local accuracy of the attribution values it provides. For instance, SHAP has been implemented by explainability libraries such as AIX360 (Arya et al., 2019) and InterpretML (Nori et al., 2019). Furthermore, SHAP explanations have been incorporated in several research studies investigating how to provide explainability of AI systems (e.g., Kaur *et al.*, 2019; Weerts, van Ipenburg and Pechenizkiy, 2019; Jesus *et al.*, 2021). The interpretation of the feature attribution scores provided by SHAP explanations depends on the ML task performed by the predictive model. When explaining a regression model, SHAP scores represent the contribution of each feature value to the model's predicted value compared to the average prediction value. Thus, the scores can be directly presented as an increment or decrement from the average predictive value with the same units of measure as the target variable (e.g., Bove *et al.*, 2022). In contrast, when explaining a classification model, SHAP scores represent the contribution to the average predicted class probability of the model.

Figure 34 shows an example of a SHAP explanation for the selected binary classification scenario using the visualization of SHAP's open-source library (S. Lundberg & Lee, 2016). The model's average predicted probability is represented by the "base value". The feature attribution scores are represented by arrows that increase or decrease the prediction probability for the explained instance. Adding the base value and the scores results in the model's prediction probability represented by "f(x)".

To satisfy DP3, two interactive visualizations were designed for the system prototype, which provide an overview of each feature's influence on the model decision while only showing the details of the most influential features (i.e., name and attribution score). Nevertheless, end-users can hover over the explanation elements of the visualization to observe the details of the additional features.

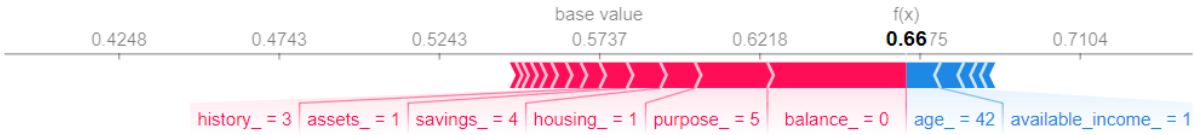


Figure 34: Example of a SHAP explanation using the original visualization.

***DF2**: Display an interactive visualization of SHAP's explanations highlighting the most influential features and allowing end-users to see details for the rest of the features.*

Figure 35 presents the two designs of the interactive visualizations integrated into the prototype. The system interface provides an overview of the loan application's details in the upper section by showing the features' values across the categories: financial information, personal information, and loan details. End-users can hover over the information icon at the top of the screen to see a detailed description of each feature. The interface also shows the system's decision recommendation and prediction probability in the top right corner. The probability is presented to end-users as a confidence level in a percentage, together with a text legend indicating one of the system's three levels of confidence on the recommendation (low, medium, or high confidence).

The "cascade" visualization on the left side of Figure 35 is an adaptation of SHAP's original visualization. In the design of this visualization, the overview provided by the stacked bars from the original visualization was maintained. Nonetheless, an individual bar for each feature was included below. Moreover, the name of the most influential features was included in the bar. SHAP's original color coding was maintained to represent features contributing to approval with blue bars and rejection with red bars. The label "Base Probability" was used to indicate the model's average predicted probability and the label "Decision Probability" to show the prediction probability of the explained instance. It was decided to show each class prediction probability instead of a complementary probability below 0.5. Thus, for approved instances, the features increasing the probability were shown in blue and decreasing in red, while for rejected instances, this was inverted (see Figure 36). The interactive visualizations allow end-users to hover over each bar to see the feature name, value, and corresponding attribution score.

The "treemap" visualization presented on the right side of Figure 35 was designed to provide a simpler visualization without the probability axis. In contrast to the cascade visualization, the treemap visualization uses boxes to represent the attribution scores. The box size representing each feature corresponds to the magnitude of their score. Moreover, this visualization utilizes the same color coding as the cascade visualization and shows the features' names and attribution scores for the most influential features. In contrast to the cascade visualization, the features influencing approval are always located on the right side of the visualization, while the features influencing rejection are on the left. End-users can hover over the boxes to see the details of the corresponding feature. The treemap visualization includes the model's average predicted probability as an additional box with the description "Baseline".

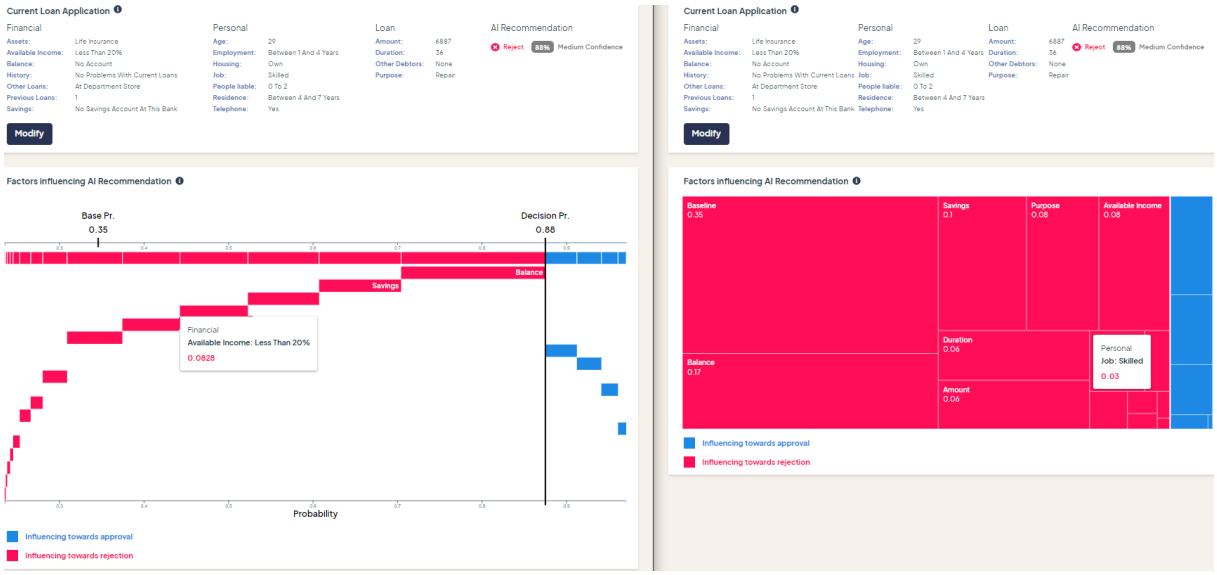


Figure 35: Design of proposed interactive visualizations for SHAP explanations. The figure shows the cascade visualization on the left and the treemap visualization on the right.

The base probability for both visualizations is shown according to the predicted class. The predictive model's average prediction value is 0.35, representing the dataset's oversampled 300 rejected instances. Thus, for rejected instances, a base probability of 0.35 is shown. In the case of approved instances, the complementary base probability of 0.65 is shown, representing the 700 approved instances.

To satisfy DP4, a "what-if" interactive functionality was instantiated that allows end-users to explore how modifications to the features' values of the instance being explained affect the system's decisions.

***DF3**: Provide an interactive user interface that allows end-users to explore "what-if" scenarios by changing the features' values and observing the system's decision and corresponding explanation visualization.*

Following DP3, the what-if functionality is disabled by default. End-users can activate it by clicking on the "Modify" button in the left part of the screen below the instance feature details (see Figure 35). When this button is clicked, all features display a caret-down icon next to the original value to indicate that modifying the values is now possible, as shown in Figure 36. Additionally, the "Modify" button is replaced by a "Reset" button designed to revert any modifications and turn off the what-if functionality. End-users can click on any feature value or its corresponding caret-down icon to open a drop-down menu that allows the modification of the original value. The drop-down menu lists valid values for categorical features and an adjustable slider for numerical features. To provide an overview of the features' values that have been modified, the interface highlights them using orange text. When there is at least one modified value, the interface displays the decision the system would make and its corresponding confidence level below the original decision on the right side of the screen. Moreover, the interface shows the "Generate New Explanation" button that provides the corresponding interactive visualization for the SHAP explanation of the modified instance, as shown in Figure 36.

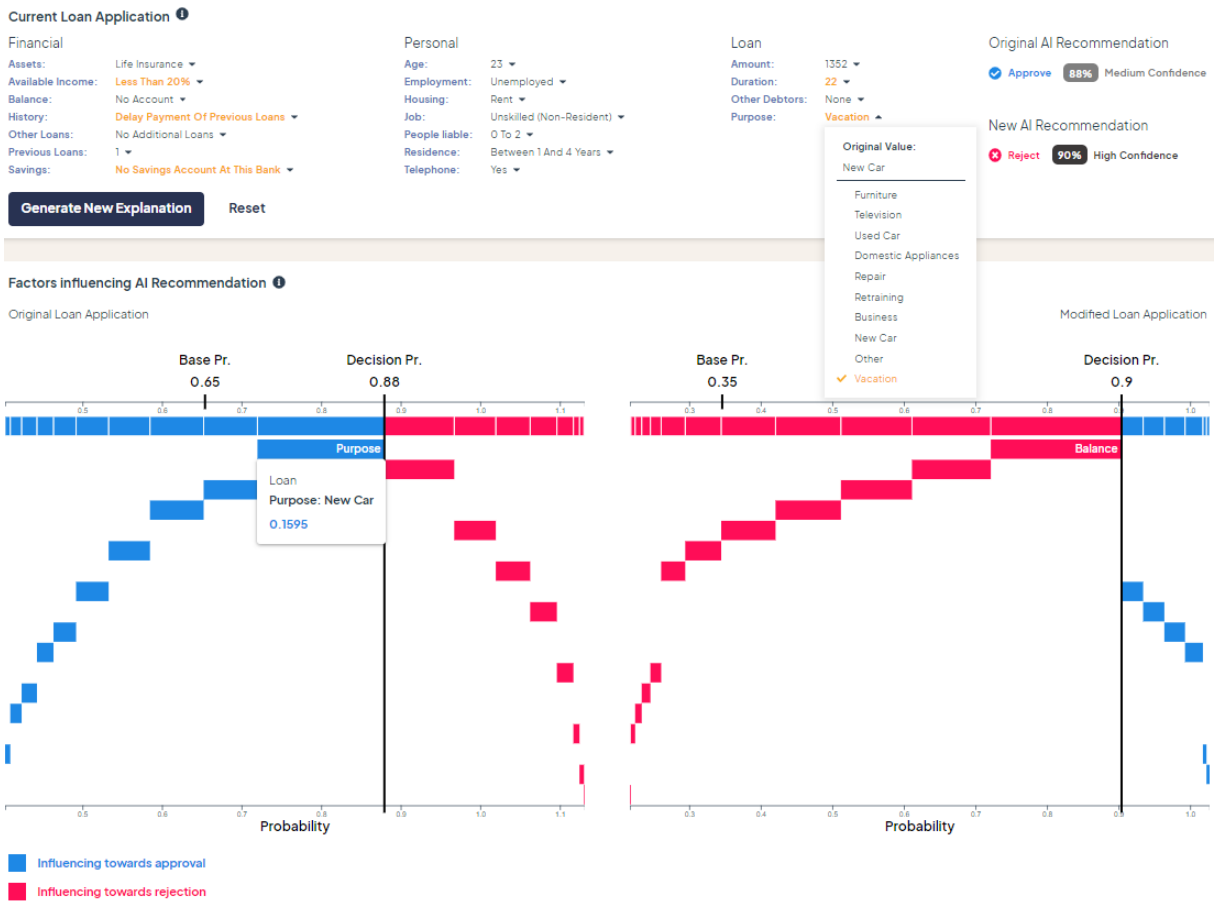


Figure 36: Design of proposed interactive visualizations for SHAP explanations. The figure shows the what-if functionality to generate explanations for the modified loan application.

5.5 Evaluation

An evaluation study and interviews were conducted to assess the design principles and the instantiated interactive XAI system prototype in Cycle 1 of the DSR project. In the evaluation study, participants interacted with one of five configurations to understand better how they perceived the different design features. The five configurations were: what-if without visualization (what-if), cascade visualization without what-if (cascade), treemap visualization without what-if (treemap), cascade visualization with what-if (cascade-what-if), and treemap visualization with what-if (treemap-what-if).

The evaluation study consisted of four phases. First, participants were randomly assigned to one configuration and were introduced to the scenario and the features used. Second, participants were presented with information describing the system's interface, the visualization, and the what-if functionality according to their corresponding configuration. Third, participants were asked to interact with the system by reviewing eight loan applications (four approved and four rejected) with the corresponding system's decision recommendation, confidence level, and design features according to their group. Participants were asked whether they would approve or reject each loan application. Fourth, participants were asked to respond to questions regarding their demographics and their evaluation of the

design features through self-reported measures. After the evaluation, semi-structured interviews were conducted with participants to discuss their perceptions of the system (see Appendix C1).

Twenty-one students were recruited from the KD2Lab panel as proxies for end-users applying for a bank loan, 11 females and ten males (https://www.kd2lab.kit.edu/). Eighteen participants were between 18 and 25, while three were between 26 and 30. Thirteen participants were coursing a bachelor's degree, seven a master's degree, and one a doctoral degree. Following Cheng *et al.* (2019), participants were asked about their familiarity with the task of credit scoring using a four-level scale (i.e., no experience, a little experience, some experience, a lot of experience). Nineteen stated they had no experience, and only two said they had little experience. Moreover, participants were asked to indicate their knowledge level of ML (Cheng et al., 2019). Seven had no previous knowledge, 11 had little knowledge, two had some knowledge, and one had a lot of knowledge.

The distribution of participants across the five groups was: what-if (4), cascade (4), cascade-what-if (4), treemap (5), and treemap-what-if (4). Each participant was paid 12.00 € for participating in the evaluation study and the interviews. The average duration for the evaluation study was 30.19 minutes (SD = 8.12) and 11.28 minutes (SD = 10.66) for the interviews. All interviews were recorded and transcribed for analysis. An open coding strategy was utilized to analyze the transcripts and extract participants' evaluations of the elements of the interactive XAI system prototype (Myers, 2002).

The analysis of the semi-structured interviews revealed that participants had positive and negative feedback about the different design features of the prototype. Moreover, participants also made some suggestions for improvements that are planned to be incorporated in Cycle 2. Table 11 presents a summary of the evaluation study across the most relevant elements of the prototype. The percentages shown in Table 11 relate to the number of participants that interacted with the corresponding elements of the prototype according to their assigned group.

Regarding their general perception of the system, a third of the participants indicated that it was fun to use the system. P1 and P11 indicated that it was fun because it was possible to see how the system works. Four participants indicated that the system was easy to use. However, eight participants raised concerns about the features used by the system to make decisions and the granularity of their categorical values. For instance, P9 mentioned that "*savings are only considered for this bank*", and P15 indicated that it would be good to "*get more customers data*". Four participants indicated that they would like to understand how features influence decisions for other instances, indicating that some participants would like to receive global explanations revealing the model behavior across all instances.

Moreover, three participants indicated that the confidence level was very helpful, allowing them to "*see when the system was not really sure*" about a particular decision (P8). Nonetheless, two participants said it was confusing that the system provided decisions with low confidence. P1 stated that "*low confidence*

[decisions] should be reviewed by a person". In this line, although many participants liked the explanation visualizations and what-if functionality, fourteen participants indicated they would like decisions to be made in a human-system collaboration. P19 stated that it would be preferable to have the "*mathematical [reasoning]*" of the system, which can be "*very precise*", in combination with the "*human element*".

Table 11: Summary of evaluation of elements from the interactive XAI system prototype with the percentage of interviewees mentioning each point.

Element	Positive Feedback	Negative Feedback	Suggested Improvement
System	• Fun to use (33.3%) • Easy to use (19.0%)	• Problems with features (38.1%)	• Global explanations (19.0%)
Confidence Level	• Shows how sure it is (14.3%)	• Some low-confidence decisions (9.5%)	• Human-system decision (66.7%)
DF1: cascade visualization	• Helps understand decisions (87.5%) • Reveal feature importance (25.0%)	• Not clear how weights are calculated (25.0%)	• Clarify how weights are calculated (25.0%)
DF2: cascade visualization	• Easy to understand (75.0%) • Satisfying design (37.5%) • Overview of features (25.0%)	• Base probability is confusing (25.0%) • Colors change depending on the decision (25.0%)	• Do not change the position of colors (25.0%)
DF1: treemap visualization	• Helps understand decisions (100%) • Reveal feature importance (11.1%)	• No information on how weights are calculated (66.7%)	• Clarify how weights are calculated (66.7%)
DF2: treemap visualization	• Easy to understand (77.8%) • Overview of features (77.8%) • Satisfying design (33.3%)	• Baseline is confusing (44.4%) • No control over baseline (11.1%)	• Show values in all boxes (11.1%)
DF3: what-if functionality	• Easy to use (75.0%) • Helps understand decisions (66.7%) • Possible to analyze alternative scenarios (50%)	• Lack of feature importance (16.7%) • Unrealistic modifications (8.3%)	• Limit modifications to some features (8.3%) • Input field beside slider (8.3%)

Regarding the evaluation of DF1, six participants interacting with the cascade visualization and all nine interacting with the treemap visualization indicated that the provided local explanations based on SHAP helped them understand how the system makes decisions. Expressly, four participants, two of each visualization, indicated that it was beneficial that the influence relevance of the features was shown. P21 stated about the cascade visualization that "*it gives you a good feeling [to know] about how important some aspects are*", while P4 said regarding the treemap that it is essential to know "what factors are influencing [decisions]". However, two participants in the cascade groups and six in the treemap groups mentioned that they would like the system to clarify how the influence weights of each factor are calculated. For instance, P20 stated, "*I wish more transparency … to see how it calculates [the weights]*".

Concerning DF2, two participants of the cascade groups and seven participants of the treemap indicated that the visualizations provided a good overview of the features' influence on the decision. For instance, P6 stated about the treemap visualization, "*It was very clear what factors were in favor and what factors were against the approval*". Similarly, for the cascade visualization, P1 stated, "*The graphic makes it quite clear what the system is doing ... without the graphic, it would be ... impossible to understand how the system works*". Moreover, three participants from the cascade group and three from the treemap group indicated that the design of the visualizations was good. For the cascade visualization, P1 stated that "*the design was really satisfying*", while P17 said it "*looked beautiful*". Meanwhile, for the treemap visualization, P20 stated that "*it is easy to understand because it uses complementary colors and has squares of different sizes*".

Nonetheless, there were also some critiques about the visualizations: Two participants of the cascade group and three of the treemap group indicated that the base probability and baseline were confusing. P4 said over the baseline that it is unclear whether "*it is a strategical decision*" and an "*influential factor that you do not have control over*". Likewise, P14 did not understand the base probability and why it was sometimes shown as 0.35 and others as 0.65. Moreover, for the cascade visualization, two participants indicated that it was confusing that the colors were inverted in the graph for approved and rejected instances. P7 said it took some time to get used to this change, while P1 said they should not change.

Regarding the evaluation of DF3, the what-if functionality was considered by nine participants as easy to use. Additionally, eight participants indicated that it was helpful to understand how the system makes decisions. For instance, P15 said that with the modifications, it was possible to see "*how to get [higher confidence]*". Moreover, six participants indicated that the what-if functionality allows them to analyze alternative scenarios. P1 said that "*the possibility to modify the criteria and see the new recommendation was useful*". P1 also mentioned that it was possible to compare the two graphics next to each other. However, two participants from the what-if group criticized that there was no information on the relevance of each feature for the decision. P3 said that "*it would be nice to know how much each component is weighted*", while P18 said that knowing how "*each variable affects the recommendation*" would be helpful. Additionally, P4 indicated that modifying certain features would not present a realistic scenario because applicants cannot change some of their personal information. To address this, P4 suggested allowing only changes to some of the features. Finally, P3 suggested allowing writing the modification values on an input field in addition to the current slider shown in the drop-down menu of the numerical features.

The evaluation revealed that the design features utilized to instantiate the proposed design principles helped participants understand how the system makes decisions. Most participants considered that both the cascade and treemap visualizations proposed to represent SHAP explanations were easy to

understand. Participants also appreciated that the provided explanations disclosed how much influence each feature had on the decisions. Moreover, participants indicated that the what-if functionality was easy to use and allowed them to analyze alternative scenarios to understand how features affect decisions.

However, participants also highlighted difficulties in understanding some aspects of the proposed interactive XAI system prototype. Regarding the visualizations proposed to represent SHAP explanations, most negative feedback was related to how the model's average prediction is displayed. Several participants indicated this concept was confusing and highlighted that they could not fully understand what this value represented despite the detailed information received in the evaluation's introductions. Regarding the what-if functionality, some participants indicated that modifying certain features would be unrealistic because loan applicants cannot modify some aspects, such as how long they have been working in a company or living in their current address.

5.6 Discussion

5.6.1 Design Challenges

In this study, an interactive XAI system prototype was designed and developed to provide explainability for end-users as part of the first cycle of a larger DSR project. The evaluation of the prototype with end-users revealed several challenges in designing interactive XAI systems. First, end-users found that modifying certain features leads to unrealistic scenarios when analyzing counterfactual scenarios through the what-if functionality. End-users felt they could not modify certain features to get approval if their loan application was rejected. Therefore, it seems that it might be helpful to restrict modifications for specific features in an interactive XAI system in certain scenarios.

Second, end-users had problems understanding the model's average prediction concept on both the cascade and the treemap visualization. For explanations of XAI methods that rely on weights from a regression model to represent each feature's influence on the decision, the intercept represents the model's average prediction across the dataset. Some XAI methods, such as LIME, do not incorporate this intercept as part of their explanation and instead only show the features' influence on each class. Nonetheless, this expected prediction value reflects the skew towards a given class in the dataset for classification tasks with imbalanced datasets such as the one used in the selected scenario. Thus, failing to disclose the average prediction value can result in contra-intuitive explanations that show more features influencing the opposite class than the one predicted by the system. In an ideal scenario, having an equal class distribution in the training dataset would result in an intercept value in the regression model that has an insignificant effect on the model's decision. However, there are many applications in which it is very challenging to achieve an equal class distribution in the training data because one class

is significantly underrepresented (e.g., positive diagnostics in the health domain). In cases where an equal class distribution can be achieved, the baseline box could be removed from the treemap visualization simplifying the explanations by focusing only on each feature's influence on the system's decision. For the cascade visualization, the base probability would still be displayed with an approximate value of 0.5 for each predicted class.

5.6.2 Limitations and Future Work

There are limitations in the work conducted in Cycle 1 of the DSR project that need to be addressed by future research. First, the proposed design principles were instantiated in an interactive XAI system prototype evaluated in the bank loan application domain. This domain was selected as a representation of high-stakes domains where the decisions of AI systems can significantly impact individuals. Nevertheless, other domains can significantly differ in important factors that can have implications on how the proposed design principles are instantiated and how they are perceived by end-users. For instance, explanations might need to be adapted to account for the risk of disclosing proprietary information in highly sensitive domains. Therefore, it is necessary to investigate how the proposed design principles can be instantiated in interactive XAI systems developed for other domains with different characteristics. Moreover, these systems should be evaluated with targeted end-users in those domains to investigate if the explanations they provide can help them understand these systems' decisions.

Second, the prototype was implemented for a classification task using a tabular dataset with relatively few features. While the design principles provide guidelines that can be easily adapted to other tasks and types of datasets, the design features instantiated in the interactive XAI prototype might need to be adapted for different conditions. For instance, when dealing with a dataset with a significantly higher number of features, it might be very challenging to present the influence of all features on the decision, as the boxes or bars representing the influence of the least relevant features might not be visible at all. One possible way to address this challenge would be to aggregate features below a certain threshold and display them together in the visualizations. End-users could then display the details of these features by utilizing a drill-down functionality. Likewise, the actual implementation of the what-if functionality might not be appropriate to allow end-users to change the inputs for text or image data. Instead, a suitable interactive user interface would need to be designed to allow input modifications for these data types.

The plan is to address these limitations in Cycle 2 of the DSR project by instantiating the proposed design principles across two application domains and tasks. Thus, two independent, interactive XAI system prototypes that provide explanations for different datasets are planned to be developed to investigate how the derived meta-requirements and proposed design principles can be generalized. Furthermore, a quantitative evaluation with a larger sample size is planned to be conducted by recruiting target end-users of the corresponding application domains.

5.6.3 Theoretical and Practical Implications

As AI systems increasingly support decision-making in high-stake applications, end-users affected by these systems' decisions must have access to explanations that help them understand the reasons behind these decisions to trust and accept them (Fernández-Loría et al., 2022). To address these requirements, research in the field of XAI has proposed models that provide explanations in non-technical terms to support end-users understanding (Diakopoulos et al., 2017). Nonetheless, despite these research efforts, it has been shown that most of the developed XAI methods provide static explanations that limit the amount of information conveyed to end-users, resulting in insufficient understanding (Cheng et al., 2019; Liu et al., 2021).

This study contributes design knowledge for interactive XAI systems by demonstrating how explanations in the form of interactive visualizations can give end-users more control over the information they receive. Interactive visualizations can transform abstract data into meaningful representations, which help to provide an overview of the most relevant factors influencing a decision to end-users. Additionally, these interactive visualizations allow end-users to explore details about additional factors not presented in an overview. As a result, these interactive explanations give end-users a sense of agency that can promote understanding and trust in AI systems (Sun & Sundar, 2022). Moreover, this study provides insights into how end-users interact with different elements of interactive XAI systems and how these elements can help them understand AI systems' decisions.

Furthermore, this study derived meta-requirements from existing knowledge in the literature and then proposed four design principles to address them. Afterward, three design features were proposed to instantiate the proposed design principles into an interactive XAI system prototype. Through these proposed design principles and design features, this study offers practical guidelines for researchers and practitioners in designing interactive XAI systems. Additionally, a GitHub open-source repository with the implementation of the system prototype and the software architecture design is provided. Therefore, researchers and practitioners can rely on this work to continue exploring how to design interactive XAI systems and investigate how they help end-users understand their decisions.

5.7 Conclusion

This study argues for the importance of designing interactive XAI systems for end-users to assist them in understanding AI systems' decisions. Relying on the DSR methodology and existing design knowledge provided by the XAI literature, the first design cycle of the DSR project was conducted, and two meta-requirements for interactive XAI systems designed for end-users were derived. To address these meta-requirements, four design principles were proposed and instantiated into an interactive XAI system prototype. An evaluation of the prototype and interviews with end-users revealed that the

proposed design features implemented in the initial prototype could help end-users understand how the system makes decisions. However, several potential improvements were also identified, which are planned to be addressed in Cycle 2 of the DSR project. This study contributes to the XAI literature by identifying design knowledge for developing interactive XAI systems to increase end-users' understanding and trust. Furthermore, with the development of the interactive XAI system prototype instantiating the design principles, this study provides researchers and practitioners with guidelines on giving end-users more control over the information they receive to help them better understand how AI systems make decisions.

6 Discussion

Designing AI systems that provide local explanations from model-agnostic XAI methods to end-users is not trivial. For end-users to comprehend the logic behind AI systems' decisions, it is essential to design explanation representations that end-users can understand. Furthermore, these explanations must allow end-users to review the underlying factors influencing decisions to ensure they are not biased. Therefore, it is necessary to design explanations that communicate the logic behind AI systems to end-users in non-technical terms. Additionally, it is necessary to understand how end-users cognitively process explanations and the contingency factors that influence whether or not they trust these XAI systems. Explanations have been shown to increase end-users' trust through perceived transparency or to decrease trust when explanations reveal biases in the system's decisions. Thus, it is necessary to consider end-users' subjective evaluation and beliefs to understand the effects of explanations on trust. Furthermore, prior research has shown that providing end-users with the static explanations that most XAI methods generate can result in an insufficient understanding of how AI systems make decisions. Therefore, this dissertation demonstrates that it is necessary to design interactive explainable interfaces that allow end-users to investigate further the factors influencing a given decision.

To address the challenges of designing XAI systems that successfully deliver local model-agnostic explanations to end-users, this dissertation investigates across three research studies how end-users evaluate local explanations from model-agnostic XAI methods. Moreover, this dissertation also investigates how these explanations need to be designed to impact end-users' trust positively. To achieve this, this dissertation was centered on the bank loan applications domain, where AI systems predict the decision to approve or reject a loan application by evaluating its risk using a set of attributes. The results of all three research studies have theoretical and practical contributions to the ongoing research on improving the transparency of AI systems through XAI methods. These contributions are discussed in the following sections. Additionally, the limitations of this dissertation and the avenues for future work for each study are outlined.

6.1 Theoretical Contributions

This dissertation contributes knowledge for designing local explanations of model-agnostic XAI methods and a deeper understanding of how end-users evaluate these explanations. Additionally, it presents design knowledge for an interactive XAI system providing local model-agnostic explanations that help end-users interactively explore the underlying factors behind the system's decisions.

Study 1 utilized a user-centered perspective to analyze the representation design of local model-agnostic explanations and understand their evaluation by end-users. Therefore, this study focuses on answering the two research questions: *How to design comparable local model-agnostic explanation*

representations from different XAI methods following a user-center approach? (**Sub-RQ1**) and how *do end-users perceive, evaluate, and visually attend to the designed local model-agnostic explanation representations from different XAI methods?* (**Sub-RQ2**). To achieve this, Study 1 first aggregated the challenges of designing local model-agnostic explanations for end-users and provided an overview of research investigating how end-users evaluate explanation representations from different model-agnostic methods.

Subsequently, Study 1 refined the representations of local explanations from four well-established model-agnostic methods (i.e., Anchors, DICE, LIME, and SHAP) following an iterative design process involving end-users. Throughout the three iterations of the design process, end-users' perceived satisfaction, understandability, and trust were measured using self-reports to investigate whether there were any differences among the evaluated explanation representations. Additionally, end-users' ability to extrapolate provided explanations to guess the system's prediction for other unseen instances was analyzed as an objective performance measure. Moreover, end-users' feedback on the explanation representations design was collected to investigate their perceptions and identify potential understanding challenges that might arise from these designs. The results of this iterative design process revealed similar levels of perceived satisfaction, trust, and understandability, indicating that explanations can, to a certain degree, help end-users understand how AI systems make decisions. Furthermore, it was observed that Anchors' explanations allowed end-users to generalize them to guess the system's prediction in other instances correctly.

After the iterative design process to investigate end-users' visual attention of explanation representations, Study 1 analyzed existing literature incorporating eye-tracking in the evaluation of explanations. Then, Study 1 conducted a laboratory experiment leveraging eye-tracking technology to evaluate how end-users visually attend to each of the refined explanation representations' designs. This evaluation was followed by interviews to understand end-users' perceptions and preferences of the explanation representations. The end-users' eye-tracking data analyses showed differences in visual attention on explanations across methods. Specifically, they revealed lower visual attention on Anchors' explanations than on the explanations of the other methods, indicating that Anchors' explanations provide simpler explanations that might require less mental effort to be processed. Moreover, the results show that processing DICE's explanations could require a high mental effort from end-users due to the amount of information they present. Regarding the interviews conducted, an analysis revealed that end-users process each explanation type very differently and that some factors can influence their perceptions and preferences.

Study 2 applied a PCV theory and social identity theory perspective to investigate the contingency factors of the opposing effects of explanations on end-users' trust in the context of biased AI systems. Therefore, this study focuses on answering the research question: *How do end-users differ in their*

evaluation of a biased XAI system's trustworthiness based on their level of stigma consciousness? (**Sub-RQ3**). To provide an answer to this question, Study 2 analyzed the two opposing theoretical mechanisms proposed in the literature that describe how explanations can affect end-users' trust. Specifically, prior research suggests that in the context of biased AI systems, explanations can either increase trust through perceived transparency or decrease trust as end-users perceive the system as biased and unfair. Moreover, Study 2 aggregated an overview of research investigating how end-users perceive and evaluate AI systems that provide biased decision recommendations.

To investigate these conflicting effects of explanations on trust, Study 2 selected a scenario in which a gender-biased AI system reveals favoritism for male over female loan applicants through explanations. Then, this study derived a research model grounded on related literature and theoretical foundations of PCV theory and social identity theory to hypothesize about explanations' effects on end-users' trust (see Figure 26). Specifically, this research model proposes that two pathways moderate explanations' effects on end-users' trust. In the first pathway, explanations increase end-users' trust through mediation from perceived transparency. On the second pathway, explanations decrease end-users' trust through mediation from perceived bias and subsequently through experienced PCV from the perceived bias. Moreover, this research model proposes that end-users' reaction to discrimination of the threatened group according to their own identity (i.e., stigma consciousness) has a moderating effect on their perceived bias and experienced PCV.

Then, Study 2 evaluated the proposed research model in two online experiments that provided explanations generated by the XAI method LIME. Experiment 1 compared groups of end-users allocated in three conditions according to the presence or absence of bias in the system and explanations: neutral AI with explanations, gender-biased AI with explanations, and AI without explanations. The results of Experiment 1 supported the hypotheses regarding the effects of explanations of end-users' trust and the moderating effect of stigma consciousness on perceived bias but not on PCV. Experiment 2 expanded the findings of Experiment 1 on the moderating effect of stigma consciousness by experimentally manipulating the salience of stigma consciousness through priming by comparing groups of end-users allocated in four conditions: neutral AI/no-priming, neutral AI/priming, biased AI/no-priming, and biased AI/priming. The results of Experiment 2 confirmed the negative effect of explanations for biased AI systems on trust through perceived bias and PCV. Additionally, these results provided insights into the moderating role of stigma consciousness. Specifically, priming reduced the threshold when participants interpreted the bias as PCV.

Overall, Study 2 contextualizes the PCV theory on gender-biased XAI systems revealing conflicting effects of explanations to those found in previous research. Specifically, in contrast to W. Wang et al. (2018, 2019) work showing that explanations reduce PCV by disclosing biases about sponsorship in recommender agents, this study reveals that explanations increase PCV by disclosing gender biases.

Furthermore, Study 2 contributes to research on cognitive evaluations of explanations by demonstrating that end-users' evaluation of explanations provided by a gender-biased AI system is influenced by their stigma consciousness. Moreover, this study demonstrates that end-users' evaluation of explanations is highly subjective and can easily be influenced by the attributes displayed in explanations, the context (priming), and end-users' attitudes towards them. Finally, Study 2 contributes to theories on trust formation, particularly regarding the role of explanations on trust. Specifically, explanations increase knowledge-based trust because they enable the validation of the provided recommendations by AI systems.

Finally, **Study 3** proposed an interactive XAI system design that allows end-users to explore explanations interactively, giving them more control over the information they receive. Therefore, this study focuses on answering the research questions: *How to design interactive explainable artificial intelligence (XAI) systems to help end-users to better understand AI systems' decisions?* (**Sub-RQ4**). To achieve this, Study 3 first aggregated the challenges of providing static explanations from model-agnostic methods and an overview of research related to developing interactive XAI systems.

Afterward, Study 3 relied on a DSR project and existing knowledge in the literature to derive two meta-requirements (MRs) of interactive XAI systems. MR1 identified that an interactive XAI system should be able to provide end-users with explanations that reveal meaningful information on the logic behind its decision in non-technical terms. Meanwhile, MR2 identified that an interactive XAI system should allow end-users to request additional information about its logic if explanations are considered insufficient to understand and trust its decisions. Then, this study proposed four design principles (DPs) to address the derived meta-requirements. DP1 proposes enabling the system to provide explanations that reveal how a specific system's decision was made. Meanwhile, DP2 recommends enabling the system to provide explanations that disclose the factors influencing each decision and their relative weights. DP3 proposes to enable the system to present an interactive visualization that provides an understandable overview of the most important factors influencing decisions and allows end-users to request details regarding additional factors. Finally, DP4 recommends providing the system with an interactive interface that allows end-users to explore how changes in the input features affect its decisions.

To instantiate the derived DPs, Study 3 proposed design features (DFs) for an interactive XAI system prototype for the bank loan application domain utilized in previous studies. To satisfy DP1 and DP2, DF1 suggests that the prototype provides local explanations from SHAP that give Shapley values as weights indicating how each feature influenced the system's decision. Moreover, to address DP3, DF2 proposes two designs of interactive visualizations of SHAP's explanations highlighting the most influential features and allowing end-users to explore details for the rest (i.e., cascade and treemap). Lastly, to address DP4, DF3 suggests providing an interactive interface that allows end-users to explore

"what-if" scenarios by changing the feature's values and observing the resulting system's decision and corresponding explanation visualization.

Subsequently, Study 3 conducted an evaluation and interviews with end-users to assess the design principles and instantiated interactive XAI system prototype. In this evaluation, participants interacted with one of five prototype configurations to understand how they perceived the different DFs: what-if without visualization, cascade visualization without what-if, treemap visualization without what-if, cascade visualization with what-if and treemap visualization with what-if. The evaluation results revealed that the DFs can help end-users understand how the system makes decisions. Specifically, end-users appreciated that both visualization designs disclosed each feature's influence on the decisions. Moreover, end-users indicated that the what-if functionality allowed them to analyze alternative scenarios to understand how features affect decisions. Additionally, the evaluation revealed challenges in communicating the ML model's average prediction to end-users.

Overall, Study 3 contributes design knowledge for XAI systems by demonstrating how interactive explanations can give end-users more control over the information they receive. Moreover, this study provides insights into how end-users interact with the different elements of interactive XAI systems and how these elements can help them understand how the system makes decisions.

In summary, the theoretical contributions of the three studies that comprise this dissertation provide meaningful knowledge to inform future research on designing XAI systems with a specific emphasis on visual representations. Specifically, these studies provide a solid theoretical foundation regarding how end-users evaluate visual representations of local explanations from model-agnostic XAI methods. The studies' findings can be leveraged to inform the design of XAI systems from different perspectives. First, the iterative design process and comparative evaluations in Study 1 provide insightful information on how researchers can increase the comparability of explanation representations while controlling for confounding factors due to the different explanation approaches. Moreover, the experimental evaluation using eye-tracking technology in Study 1 provides new knowledge about the complexity of explanation representations and the mental effort of end-users required to comprehend them. Meanwhile, the evaluation of the contingency factors influencing the effects of explanations on trust from Study 2 shows that research must extend its investigation coverage on the effects of explanations on trust by considering additional factors such as application context and end-users' characteristics. Study 2 also demonstrates that researchers need to consider the potential negative consequences that explanations can have when they reveal biases embedded in AI systems. Finally, the design research project from Study 3 informs the design of interactive interfaces on XAI systems. Specifically, Study 3 reveals that research needs to consider how end-users utilize these interactive explainable interfaces and the effects of the different interface elements on end-users' trust.

Table 12 summarizes the main theoretical contributions of this dissertation across the three studies it comprises. The research findings of this dissertation contribute to research in the domains of XAI and HCI.

Table 12: Theoretical contributions of this dissertation.

Section	Main Theoretical Contributions
Study I	• Design knowledge for the representations of local explanations from model-agnostic XAI methods. • Understanding of how end-users perceive, understand, and trust these explanation representations. • Understanding of how end-users visually attend and utilize different explanation representations and which ones they prefer.
Study II	• Contextualization of PCV theory on gender-biased XAI systems. • Insights on how end-users cognitively evaluate explanations. • Extend current understanding of the impact of explanations on trust.
Study III	• Design knowledge for XAI systems that allow end-users to explore explanations, giving them more control over the information they receive. • Understanding of how end-users interact with elements of interactive XAI system and how these elements help them understand how the system makes decisions.

6.2 Practical Contributions

In addition to the theoretical contributions mentioned above, this dissertation also contributes to practice by providing tangible artifacts, evaluations, and design guidelines. The research findings of this dissertation have a significant impact on the design and development of AI systems, especially in the efforts to achieve ethical and trustworthy AI.

Study 1 refined the representations of local explanations from the model-agnostic methods Anchors, DICE, LIME, and SHAP in an iterative design process with end-users. Throughout this iterative design process, the explanation representation designs were adapted from the proposed representations in each method's corresponding GitHub open-source libraries to improve their comparability. Specifically, the color coding, layout, amount of information, fonts, and terminology used were standardized as much as possible while maintaining their explainability approach. Therefore, the resulting explanation representation designs are a valuable resource for researchers and practitioners, as these were refined considering end-users' needs. Additionally, these refined explanation representations' evaluations provide empirical evidence on how they can help end-users understand AI systems' decisions. Study 1 also provides an open-source reference implementation in a GitHub repository of the refined explanation representations design by providing the Python code that implements the explainability methods of the four model-agnostic methods and generates the explanation representation visualization.[24]

[24] https://github.com/miguelmezamartinez/Local-Model-agnostic-Explanations-Representations

Moreover, analyses of end-users' eye-tracking data in Study 1 provide insightful information on how end-users process the explanation representations from different perspectives. For instance, they show differences in end-users' visual attention on explanation representations across methods and among regions of an explanation representation. These differences in visual attention are observable in the heatmaps provided in Study 1. Overall, the eye-tracking evaluation provides researchers and practitioners insights into how end-users utilize explanation representations, which can be used to inform their design in the future. Furthermore, the interviews in this study show that end-users process each explanation type very differently and that external factors can influence how they perceive explanations and which ones they prefer. These interviews also reveal the characteristics that end-users appreciate from each explanation type and the limitations these might have in real-world applications.

Study 2 evaluated explanations from LIME in two online experiments in the context of biased AI systems to investigate the effects of explanations on end-users' trust. As a result of these experiments, this study provides insights into the challenges of evaluating explanations of a gender-biased AI system. Specifically, Study 2 results demonstrate that it is necessary to consider how end-users' stigma consciousness affects how they evaluate explanations, providing evidence that end-users' personal attitudes can shape their interaction with AI systems. These results inform practitioners of the importance of considering end-users' characteristics and the application context for the design of explanations.

Furthermore, Study 2 demonstrates that in the context of a gender-biased AI system, explanations alone are insufficient for end-users to identify biases in the system's recommendations and that, in fact, explanations can have adverse effects that prevent end-users from detecting these biases. Therefore, this study indicates that practitioners need to consider end-users' implicit attitudes when designing XAI systems, as it is essential to consider the amount of information disclosed by explanations.

Study 3 proposed an interactive XAI system design relying on a DSR project. Specifically, this study derived meta-requirements (MRs) from existing knowledge in the literature and then proposed four design principles (DPs) to address them. Afterward, Study 3 proposed three design features (DFs) to instantiate the proposed DPs into a system prototype. The resulting system prototype provides two interactive visualizations of local explanations from SHAP that reveal the attributes' influence on the system's decisions (i.e., cascade and treemap). These proposed visualizations provide an overview of the most influential attributes and allow end-users to explore details for the rest.

Moreover, the interactive interface allows end-users to explore "what-if" scenarios by changing the feature's values and observing the resulting system's decision and corresponding explanation visualization. Therefore, through the proposed DPs and DFs, this study offers practical guidelines for researchers and practitioners in designing interactive XAI systems. Furthermore, this study provides a

GitHub repository with the implementation of the interactive XAI prototype and the software architecture design.[25]

In summary, this dissertation provides several practical contributions across the three conducted studies. The empirical evidence and insights provided in this dissertation on how end-users utilize explanations and the effects that explanations can have on end-users' trust may be used to inform future research on the field of XAI. Additionally, AI practitioners can use this evidence and insights, together with the guidelines proposed in this dissertation, to design and develop explanations and interactive explainable interfaces that satisfy end-users' needs. In this line, this dissertation provides reference implementations as open-source repositories of the refined explanation representation designs and the interactive XAI system prototype. Researchers and practitioners can adapt these reference implementations for future research, design, and development of local explanations from model-agnostic methods and interactive explainable interfaces of AI systems. Table 13 summarizes the main practical contributions of this dissertation across the three conducted studies.

Table 13: Practical contributions of this dissertation.

Section	Main Practical Contributions
Study I	• An open-source reference implementation of the refined explanation representations of local model-agnostic methods Anchors, DICE, LIME, and SHAP. • Empirical evidence on how these refined explanation representations can help end-users understand AI systems' decisions. • Insights into how end-users utilize these explanation representations through analyses of end-users' eye-tracking data and interviews.
Study II	• Insights into the challenges of evaluating explanations of biased AI systems. • Empirical evidence on how end-users' personal attitudes can shape their interaction with AI systems. • Suggestions for practitioners to consider end-users' implicit attitudes when designing XAI systems.
Study III	• Guidelines on how to design and develop interactive XAI systems. • A reference implementation of interactive XAI system prototype.

6.3 Limitations and Future Research

All three studies in this dissertation come with limitations. These limitations should be explicitly mentioned and addressed in future research to expand the research findings of this dissertation. The following section outlines these limitations along each study and discusses their implications for future work.

Study 1 investigated how to design comparable local model-agnostic explanation representations from different XAI methods and how end-users evaluate and visually attend to these designs. To decide which

[25] https://github.com/miguelmezamartinez/interactive_xai_system

XAI methods to investigate, a search was conducted in Study 1 to identify relevant methods among researchers and practitioners. It was observed that there were model-agnostic methods published in open-source libraries utilizing different programming languages. Thus, an assessment was performed to consider the availability of implementations supporting the creation of the ML predictive model. Python was selected as the programming language for the implementation due to the large availability of data science and ML libraries such as scikit-learn,[26] Keras,[27] and TensorFlow.[28] As a result, only model-agnostic methods available as open-source libraries in Python were considered for the investigations conducted in Study 1. For the selection process, this study considered these methods' relevance in the literature and among practitioners and the type of explanations they provide. On this basis, the selected local model-agnostic methods in Study 1 were Anchors, DICE, LIME, and SHAP. Nonetheless, these investigated methods represent only a proportion of all available local model-agnostic methods. Therefore, it is necessary to further investigate how end-users evaluate explanations from other local model-agnostic XAI methods following different explainability approaches. For instance, researchers could consider methods that utilize feature(s)-prediction relationships, such as ICE (Goldstein et al., 2015) or case-based explanations (Doyle et al., 2003; Hase & Bansal, 2020).

Moreover, Study 1 utilized the context of bank loan applications to conduct the iterative design process and evaluations of explanation representations with end-users. This application domain was selected as a representative instance of a high-stake domain where AI systems decisions can significantly impact individuals. Nevertheless, end-users' explainability needs could differ drastically in other application domains. For instance, factors such as end-users' goals or the criticality of AI systems' decisions could influence how end-users evaluate explanations. Thus, future work is needed to investigate whether end-users' evaluation of explanation representations from model-agnostic XAI methods differ across application domains to optimize their design to improve end-users' understanding and positively influence their trust.

Additionally, Study 1 relied on an ML binary classification task utilizing a tabular dataset to train the predictive model that provided the decision recommendations to approve or reject the bank loan applications. As a result, this study focused on refining and evaluating explanation representations from the four selected local model-agnostic XAI methods specifically for the selected ML task and dataset type. Nevertheless, generating explanations for other ML tasks and dataset types has significant implications for the explanation representations and the information they provide. For example, the explanation representations of many XAI methods on multiclass image classification tasks rely on highlighting regions of an image to indicate which pixels were influential in the predicted class. Likewise, explanation representations in text classification need to be able to present the influence that

[26] https://scikit-learn.org/
[27] https://keras.io/
[28] https://www.tensorflow.org/

each word in the text had on the prediction. Therefore, future research must evaluate how end-users evaluate the explanation representations of local model-agnostic methods across different dataset types (e.g., visual, or textual) and ML tasks (e.g., regression, clustering, or multiclass classification).

Study 2 investigated the contingency factors of the effects of LIME explanations on end-users' trust in the context of biased AI systems utilizing the same scenario of bank loan applications from Study 1. The specific type of bias that was selected to contextualize this investigation was gender bias instantiated through an AI system that systematically favored male loan applicants over female applicants. The findings of this investigation demonstrated that explanations of a gender-biased AI system can increase end-users' perceived PCV and, in turn, decrease their trust in the system. However, these effects of explanations on PCV and trust are constrained to the selected gender bias context. Prior work has shown that explanations' effects on PCV and trust heavily depend on the specific context. For instance, W. Wang et al. (2018, 2019) showed that by disclosing a bias, explanations can help reduce the experienced PCV in the context of sponsor content. Therefore, an interesting research avenue involves investigating how explanations can affect PCV and trust across different types of bias.

Furthermore, Study 2 applied a social identity theory perspective to investigate how end-users' social identity moderated their PCV experience. The findings in this study showed that end-users' awareness of gender discrimination (i.e., stigma consciousness) affects their evaluation of explanations of the gender-biased AI system and their perceived PCV. Nevertheless, more research is needed to get a deeper understanding of the exact mechanisms that affect end-users' perception of a PCV. Specifically, future work should consider in more detail the threshold needed to experience a PCV and consider additional factors related to how end-users differ in this evaluation process, such as the degree to which they are personally affected or benefit from the biased decisions.

Moreover, in line with the theory on explanatory coherence (Thagard, 1989), Study 2 indicates that end-users' evaluation of explanations can be influenced by the overall consistency of attributes displayed in these explanations. The results of the two experiments in this study suggest that end-users' evaluation of explanations is highly subjective and can be easily influenced by the attributes displayed and the importance end-users place on specific attributes. Consequently, end-users evaluate not only the consistency of attributes displayed in explanations but also the consistency between them and their underlying assumptions about the attributes' relevance. Therefore, a motivated reasoning perspective (Kunda, 1990) should be followed in future research to theorize how end-users' pursuit of consistency between their own beliefs and the attributes displayed in explanations may result in an accurate or biased evaluation of explanations.

Additionally, Study 2 utilized LIME as a representative of local model-agnostic XAI methods to investigate the effects of explanations on end-users' trust. LIME was selected due to its relevance in the literature and popularity among practitioners. However, it is not clear if explanations from other XAI

methods would have the same effects on end-users' PCV and trust. Thus, future work should extend the research performed by this study to evaluate explanations of other XAI methods in the context of biased AI systems to understand how different explainability approaches might influence end-users' trust.

Study 3 relied on a DSR project to propose four design principles for interactive XAI systems and instantiate them into a system prototype that provides explanations based on SHAP. The design principles proposed in this study address the requirements interactive XAI systems must fulfill to help end-users understand their decisions. However, similarly to the other studies in this dissertation, Study 3 utilized the bank loan application domain with a binary classification ML task and a tabular dataset in its research efforts to instantiate the design principles into an interactive XAI system prototype. As a result, the system prototype was explicitly designed to instantiate these design principles while addressing some of the explainability requirements specific to this context and boundary conditions. Therefore, future research should explore how to instantiate these design principles in interactive XAI systems for other contexts. Afterward, these design principles should be evaluated with end-users to investigate their generalizability.

Furthermore, Study 3 proposed two interactive visualizations of local explanations based on the model-agnostic method SHAP. These visualizations provide an overview of the most influential attributes of the system's decision and allow end-users to explore details by themselves. Even though visual explanations of DICE and LIME methods were designed and implemented during the development of the system prototype, Study 3 did not evaluate them with end-users as part of the design research project. For DICE, it was observed that the XAI method would take considerable computing time to generate explanations and that it would not be feasible to implement it as part of the interactive explainable interface. Concerning LIME, several challenges were found regarding the generation of contra-intuitive explanations showing more attributes influencing the opposite class than the one predicted, which are caused by a class imbalance of the dataset. Future work could address these limitations by improving the implementation of DICE to deliver explanations in real-time. Moreover, future research could explore how to redesign LIME explanations to address the challenges of class imbalance datasets to avoid providing end-users with contra-intuitive explanations.

7 Conclusion

Explainability is increasingly being considered a critical capability of trustworthy AI systems that can ethically support decision-making processes. With the growing deployment of AI systems across many high-stake domains, increasing their transparency is imperative to allow end-users to understand the logic behind their decisions. In this light, there has been a surge of interest in XAI to produce systems that can explain their decisions to end-users in non-technical terms. In particular, there has been a focus on developing local model-agnostic explainable methods that can generate explanations of individual predictions for any predictive model due to their higher applicability and scalability. Motivated by the significance of explainability in the future development of AI systems, this dissertation investigates how to design visual representations of local model-agnostic XAI methods to increase end-users' understanding and trust in three studies.

In the first step, Study 1 utilized a user-centered perspective to examine the visual representation design of local explanations from XAI model-agnostic methods to understand how end-users evaluate them. To achieve this, Study 1 refined the representations of local explanations from the XAI model-agnostic methods Anchors, DICE, LIME, and SHAP following an interactive design process involving end-users. Afterward, Study 1 conducted a laboratory experiment leveraging eye-tracking technology to evaluate how end-users visually attend to these refined explanation representations' designs. Hence, Study 1 expands the understanding of how end-users utilize and visually attend to representations of local-model agnostic explanations. Moreover, Study 1 contributes with design knowledge for the visual representation of local model-agnostic explanations by providing empirical evidence on how they can help end-users understand AI systems' decisions and by providing an open-source reference implementation of the refined explanation representations.

Afterward, Study 2 investigated the contingency factors of the effects of explanations on end-users' trust in the context of biased AI systems by applying a PCV and social identity theory perspective. To accomplish this, Study 2 conducted two experiments investigating how end-users' awareness of gender discrimination (i.e., stigma consciousness) moderates the conflicting effects of LIME explanations on end-users' trust through perceived transparency and PCV when interacting with a gender-biased AI system. Therefore, Study 2 contextualizes PCV theory on gender-biased XAI systems to extend the current understanding of the impact of explanations on end-users' trust by providing insights into how they cognitively evaluate explanations. Furthermore, Study 2 provides insights into the challenges of evaluating explanations of biased AI systems and demonstrates how end-users' personal attitudes can shape their interaction with these systems.

Finally, Study 3 examined how to design interactive XAI systems to increase end-users' understanding and trust. To achieve this, Study 3 relied on a DSR project to propose design principles that address the

requirements XAI systems must fulfill to help end-users understand their decisions. Additionally, Study 3 instantiated these design principles in an interactive XAI system prototype, which was then evaluated with end-users. Hence, Study 3 contributes with design knowledge for interactive XAI systems that allow end-users to explore explanations, giving them more control over the information they receive. Moreover, Study 3 expands the understanding of how interactive XAI systems can help end-users understand their decisions.

Overall, this dissertation addresses researchers' calls (e.g., Adadi & Berrada, 2018; Miller, 2019; Mittelstadt et al., 2019; Ribera & Lapedriza, 2019) to extend the current understanding of how end-users evaluate local explanations from model-agnostic methods, the effects that these explanations can have on their trust in AI systems, as well as providing design knowledge for improving the design of visual representations. Specifically, this dissertation takes a holistic approach to understanding how to improve the visual design of the explanation representation of local model-agnostic explanations and the design of interactive XAI interfaces to increase end-users' understanding and trust.

8 References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3173574.3174156

Abdul, A., Von Der Weth, C., Kankanhalli, M., & Lim, B. Y. (2020). COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376615

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Aggarwal, A., Lohia, P., Nagar, S., Dey, K., & Saha, D. (2019). Black Box Fairness Testing of Machine Learning Models. *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 625–635. https://doi.org/10.1145/3338906.3338937

Agnar, A., & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, *7*(1), 39–59. https://doi.org/10.3233/AIC-1994-7104

AI HLEG. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. https://doi.org/10.2759/002360

Albert, E. T. (2019). AI in talent acquisition: a review of AI-applications used in recruitment and selection. *Strategic HR Review*, *18*(5), 215–221. https://doi.org/10.1108/SHR-04-2019-0024

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P., & Suh, J. (2015). ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 337–346. https://doi.org/10.1145/2702123.2702509

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks. *Knowledge-Based Systems*, *8*(6), 373–389. https://doi.org/10.1016/0950-7051(96)81920-4

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it biased against blacks. *ProPublica*, *23*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Apley, D. W., & Zhu, J. (2016). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1059--1086. https://doi.org/10.1111/rssb.12377

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI We Trust? Perceptions About Automated Decision-making by Artificial Intelligence. *AI and Society*, *35*(3), 611–623. https://doi.org/10.1007/S00146-019-00931-W/TABLES/4

Ardissono, L., Goy, A., Petrone, G., Segnan, M., & Torasso, P. (2003). Intrigue: Personalized Recommendation of Tourist Attractions for Desktop and Hand Held Devices. *Applied Artificial Intelligence*, *17*(8–9), 687–714. https://doi.org/10.1080/713827254

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., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2019). One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *ArXiv Preprint ArXiv:1909.03012*.

Bansal, G., Fok, R., Ribeiro, M. T., Wu, T., Zhou, J., Kamar, E., Weld, D. S., & Nushi, B. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764

Barlas, P., Kyriakou, K., Kleanthous, S., & Otterbacher, J. (2019). What Makes an Image Tagger Fair? *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, 95–103.

Barria-Pineda, J., Akhuseyinoglu, K., Želem-Ćelap, S., Brusilovsky, P., Milicevic, A. K., & Ivanovic, M. (2021). Explainable Recommendations in a Personalized Programming Practice System. *Proceedings of the International Conference on Artificial Intelligence in Education*, *12748 LNAI*, 64–76. https://doi.org/10.1007/978-3-030-78292-4_6

Berendt, B., & Preibusch, S. (2017). Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop—and Under the Looking Glass. *Big Data*, *5*(2), 135–152.

Bigras, É., Léger, P. M., & Sénécal, S. (2019). Recommendation Agent Adoption: How Recommendation Presentation Influences Employees' Perceptions, Behaviors, and Decision Quality. *Applied Sciences*, *9*(20), 4244. https://doi.org/10.3390/APP9204244

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI-17 Workshop on Explainable AI (XAI)*, 8–13. http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf

Bove, C., Aigrain, J., Lesot, M. J., Tijus, C., & Detyniecki, M. (2022). Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. *27th International Conference on Intelligent User Interfaces*, *22*, 807–819. https://doi.org/10.1145/3490099.3511139

Brauner, P., Philipsen, R., Calero Valdez, A., & Ziefle, M. (2019). What happens when decision support systems fail? — the importance of usability on performance in erroneous systems. *Behaviour & Information Technology*, *38*(12), 1225–1242.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bridle, J. S. (1989). Training Stochastic Model Recognition Algorithms Training Stochastic Model Recognition Algorithms as Networks can lead to Maximum Mutual Information Estimation of Parameters. *Advances in Neural Information Processing Systems*, *2*.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. *Proceedings of the International Conference on Intelligent User Interfaces, Proceedings IUI*, 454–464. https://doi.org/10.1145/3377325.3377498

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *Proceedings of the 2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Cabrera, A. A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., & Chau, D. H. (2019). FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning. *FAIRVIS:*

Visual Analytics for Discovering Intersectional Bias in Machine Learning, 46–56. https://doi.org/10.1109/VAST47406.2019.8986948

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Chakraborty, J., Majumder, S., Yu, Z., & Menzies, T. (2020). Fairway: A Way to Build Fair ML Software. *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 654–665. https://doi.org/10.1145/3368089.3409697

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789

Chollet, F. (2015). Keras. In *Keras*. GitHub. https://github.com/fchollet/keras

Chromik, M. (2021). *Human-centric Explanation Facilities: Explainable AI for the Pragmatic Understanding of Non-expert End Users*. Ludwig-Maximilians-Universität München.

Coba, L., Zanker, M., Rook, L., & Symeonidis, P. (2019). Decision-making Strategies Differ in the Presence of Collaborative Explanations: Two Conjoint Studies. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 291–302. https://doi.org/10.1145/3301275

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward Personalized XAI: A Case Study in Intelligent Tutoring Systems. *Artificial Intelligence*, *298*, 103503. https://doi.org/10.1016/J.ARTINT.2021.103503

Coyle-Shapiro, J. A.-M., Pereira Costa, S., Doden, W., & Chang, C. (2019). Psychological Contracts: Past, Present, and Future. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*(1), 145–169. https://doi.org/10.1146/annurev-orgpsych-012218-015212

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3

DARPA. (2016). *Explainable Artificial Intelligence (XAI)*. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. *Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/https://doi.org/10.1109/SP.2016.42

Deng, J., & Brown, E. T. (2021). RISSAD: Rule-based Interactive Semi-Supervised Anomaly Detection. *Proceedings of the EuroVis 2021*. https://doi.org/10.2312/evs.20211050

Dhaliwal, J. S., & Benbasat, I. (1996). The Use and Effects of Knowledge-based System Explanations: Theoretical Foundations and a Framework for Empirical Evaluation. In *Information Systems Research* (Vol. 7, Issue 3, pp. 342–362). INFORMS. https://doi.org/10.1287/isre.7.3.342

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., & Zevenbergen, B. (2017).

Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. FAT/ML. https://www.fatml.org/resources/principles-for-accountable-algorithms

Dodge, J., Vera Liao, Q., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv Preprint ArXiv: 1702.08608*.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable Artificial Intelligence: A Survey. *Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, 210–215.

Doyle, D., Alexey Tsymbal, & Pádraig Cunningham. (2003). *A Review of Explanation and Explanation in Case-Based Reasoning*.

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository — German Credit Data*. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Duchowski, A. T. (2002). A Breadth-first Survey of Eye-Tracking applications. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 455–470. https://doi.org/10.3758/BF03195475

Duchowski, A. T. (2017). Eye Tracking Methodology: Theory and Practice. In *Eye Tracking Methodology: Theory and Practice: Third Edition* (Third Edition). Springer. https://doi.org/10.1007/978-3-319-57883-5/COVER

Dudley, J. J., & Kristensson, P. O. (2018). A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, *8*(2), 1–37. https://doi.org/10.1145/3185517

El Bekri, N., Kling, J., & Huber, M. F. (2019). A Study on Trust in Black Box Models and Post-hoc Explanations. *Proceedings of the International Workshop on Soft Computing Models in Industrial and Environmental Applications*, *950*, 35–46. https://doi.org/10.1007/978-3-030-20055-8_4

Erlei, A., Nekdem, F. A., Meub, L., Anand, A., & Gadiraju, U. (2020). Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*(1), 43–52.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T. R., Zerbe, N., & Holzinger, A. (2022). The Explainability Paradox: Challenges for XAI in Digital Pathology. *Future Generation Computer Systems*, *133*, 281–296. https://doi.org/10.1016/j.future.2022.03.009

FAT/ML. (2022). *Fairness, Accountability, and Transparency in Machine Learning*. FAT/ML. https://www.fatml.org/

Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach. *MIS Quarterly*, *46*(3), 1635–1660.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friendly, M. (2008). A Brief History of Data Visualization. *Handbook of Data Visualization*, 15–56. https://doi.org/10.1007/978-3-540-33037-0_2

Frye, C., Rowat, C., & Feige, I. (2020). Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-agnostic Explainability. *Proceedings of the Neural Information Processing Systems*, *33*, 1229–1239.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, *33*(2), 678–696. https://doi.org/10.1287/ISRE.2021.1079

Gartner. (2022). *Top Strategic Technology Trends for 2022*. Gartner. https://www.gartner.com/en/information-technology/insights/top-technology-trends

Gefen, D. (2000). E-commerce: The role of familiarity and trust. *Omega*, *28*(6), 725–737. https://doi.org/10.1016/S0305-0483(00)00021-9

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Gomez, O., Holter, S., Yuan, J., & Bertini, E. (2021). AdViCE: Aggregated Visual Counterfactual Explanations for Machine Learning Model Validation. *Proceedings of the 2021 IEEE Visualization Conference (VIS)*, 31–35. https://doi.org/10.1109/VIS49827.2021.9623271

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gower, J. C. (1971). A General Coefficient of Similarity and Some of its Properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, *23*(4), 497–530. https://doi.org/10.2307/249487

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. https://doi.org/10.1109/MIS.2019.2957223

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. *ArXiv Preprint ArXiv:1805.10820*.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5). https://doi.org/10.1145/3236009

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/JAMA.2016.17216

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/AIMAG.V40I2.2850

Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022). Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. *27th International Conference on Intelligent User Interfaces*, *22*, 537–548. https://doi.org/10.1145/3490099.3511111

Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., & Aggarwal, C. (2019). Efficient data representation by selecting prototypes with importance weights. *2019 IEEE International Conference on Data Mining (ICDM)*, 260–269. https://doi.org/10.1109/ICDM.2019.00036

H2O.ai. (2022). *H2O Driverless AI*. https://h2o.ai/platform/ai-cloud/make/h2o-driverless-ai/

Hartmann, K., & Wenzelburger, G. (2021). Uncertainty, Risk and the Use of Algorithms in Policy Decisions: A Case Study on Criminal Justice in the USA. *Policy Sciences*, *54*(2), 269–287. https://doi.org/10.1007/s11077-020-09414-y

Hase, P., & Bansal, M. (2020). Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, 5540–5552.

Haverinen, T. (2020). *Towards Explainable Artificial Intelligence (XAI)*. University of Jyväskylä.

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis : A Regression-Based Approach* (Second Edi). The Guilford Press.

Hayhoe, M., & Ballard, D. (2005). Eye Movements in Natural Behavior. *Trends in Cognitive Sciences*, *9*(4), 188–194. https://doi.org/10.1016/J.TICS.2005.02.009

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). *Explaining Collaborative Filtering Recommendations*. http://www.grouplens.org/

Heskes, T., Sijben, E., Bucur, I. G., & Claassen, T. (2020). Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. *Proceedings of the Neural Information Processing Systems*, *33*, 4778–4789.

Hitchcock, C. (2018). Causal Models. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. *ArXiv Preprint ArXiv:1812.04608*.

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300809

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/pdf/4615733.pdf

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press.

Hoque, M. N., & Mueller, K. (2021). Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making. *ArXiv Preprint ArXiv:2101.00633*.

IDC. (2022). *Worldwide Semiannual Artificial Intelligence Tracker*. https://www.idc.com/getdoc.jsp?containerId=IDC_P37251

Iivari, J. (2015). Distinguishing and contrasting two strategies for design science research. *European Journal of Information Systems*, *24*(1), 107–115. https://doi.org/10.1057/ejis.2013.35

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine Learning and Deep Learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/S12525-021-00475-2/TABLES/2

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How Can I Choose an Explainer? An Application-grounded Evaluation of Post-hoc Explanations. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 805–815. https://doi.org/10.1145/3442188.3445941

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://www.tandfonline.com/doi/abs/10.1207/S15327566IJCE0401_04

Johnson, H., & Johnson, P. (1993). Explanation Facilities and Interactive Systems. *Proceedings of the International Conference on Intelligent User Interfaces, Proceedings IUI, Part F1275*, 159–166. https://doi.org/10.1145/169891.169951

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on Deep Learning with Class Imbalance. *Journal of Big Data*, *6*(1), 1–54. https://doi.org/10.1186/S40537-019-0192-5/TABLES/18

Jussupow, E., Meza Martínez, M. A., Maedche, A., & Heinzl, A. (2021). Is This System Biased? – How Users React to Gender Bias in an Explainable AI System. *ICIS 2021 Proceedings*, *11*. https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/11

Jussupow, E., Meza Martínez, M. A., Maedche, A., & Heinzl, A. (2023). Why Individuals Trust and Not Trust Biased Explainable AI Systems: A Psychological Contract Violation and Social Identity Perspective. *Working Paper, to Be Submitted*.

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, *32*(3), 713–735. https://doi.org/10.1287/isre.2020.0980

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/FNINS.2022.883385

Kasinidou, M., Kleanthous, S., Barlas, P., & Otterbacher, J. (2021). "I agree with the decision, but they didn't deserve this": Future Developers' Perception of Fairness in Algorithmic Decisions. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 690–700. https://doi.org/10.1145/3442188.3445931

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis* (Vol. 344). John Wiley & Sons.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2019). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not Enough, Learn to Criticize! Criticism for Interpretability. *Advances in Neural Information Processing Systems*, 2280–2288.

Kim, M. P., Ghorbani, A., & Zou, J. (2019). Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 247–254. https://doi.org/10.1145/3306618.3314287

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Koh, P. W., & Liang, P. (2017). Understanding Black-box Predictions via Influence Functions. *Proceedings of the International Conference on Machine Learning*, 1885–1894.

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, 1–22. https://doi.org/10.1080/0960085X.2021.1927212

Krause, J., Perer, A., & Ng, K. (2016). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5686–5697. https://doi.org/10.1145/2858036.2858529

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, 126–137.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W. K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kurzhals, K., Fisher, B., Burch, M., & Weiskopf, D. (2015). Eye Tracking Evaluation of Visual Analytics. *Information Visualization*, *15*(4), 340–358. https://doi.org/10.1177/1473871615609787

Lakkaraju, H., & Bastani, O. (2020). "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85. https://doi.org/10.1145/3375627.3375833

Landecker, W., Thomure, M. D., Bettencourt, L. M. A., Mitchell, M., Kenyon, G. T., & Brumby, S. P. (2013). Interpreting Individual Classifications of Hierarchical Networks. *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*, 32–38. https://doi.org/10.1109/CIDM.2013.6597214

Law, P.-M., Malik, S., Du, F., & Sinha, M. (2020). The Impact of Presentation Style on Human-In-The-Loop Detection of Algorithmic Bias. *ArXiv Preprint ArXiv:2004.12388*.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature 2015 521:7553*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. *Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09*, 2119. https://doi.org/10.1145/1518701.1519023

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Liu, H., Lai, V., & Tan, C. (2021). Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 45. https://doi.org/10.1145/3479552

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic Eye movements and Cognition. *Trends in Cognitive Sciences*, *4*(1), 6–14. https://doi.org/10.1016/S1364-6613(99)01418-7

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/J.TICS.2006.08.004

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., & Pan, B. (2008). Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, *59*(7), 1041–1052. https://doi.org/10.1002/asi.20794

Lundberg, S., & Lee, S. (2016). *Shap*. https://github.com/slundberg/shap

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. In Citeseer (Ed.), *Proceedings of the 11th Australasian Conference on Information Systems* (pp. 6–8).

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 122–130.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*, *160*(4), 835–850.

Mayo Clinic. (2021). *Diabetic Retinopathy - Symptoms and causes - .* Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/diabetic-retinopathy/symptoms-causes/syc-20371611

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing And Validating Trust Measure for E-Commerce: An Integrative Typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the*

Association for Information Systems, *16*(9), 2. https://doi.org/10.17705/1jais.00408

Meza Martínez, M. A., & Maedche, A. (2023). Designing Interactive Explainable AI Systems for Lay Users. *Manuscript Accepted in the International Conference on Information Systems (ICIS 2023)*.

Meza Martínez, M. A., Nadj, M., Langner, M., Toreini, P., & Maedche, A. (2023). Does This Explanation Help? Designing Local Model-Agnostic Explanation Representations and an Experimental Evaluation Using Eye-Tracking Technology. *ACM Transactions on Interactive Intelligent Systems (TiiS), Just Accep*. https://doi.org/10.1145/3607145

Meza Martínez, M. A., Nadj, M., & Maedche, A. (2019). Towards an Integrative Theoretical Framework of Interactive Machine Learning Systems. *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Ming, Y., Qu, H., & Bertini, E. (2019). RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level Control Through Deep Reinforcement Learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mojsilovic, A. (2019). *Introducing AI Explainability 360*. IBM Research Blog. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/

Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 417--431.

Moore, J. D., & Swartout, W. R. (1988). *Explanation in Expert Systems: A Survey*. https://apps.dtic.mil/sti/citations/ADA206283

Morrison, E. W., & Robinson, S. L. (1997). When employees feel betrayed: A model of how psychological contract violation develops. *Academy of Management Review*, *22*(1), 226–256. https://doi.org/10.5465/amr.1997.9707180265

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617. https://doi.org/10.1145/3351095.3372850

Muddamsetty, S. M., Jahromi, M. N. S., Ciontos, A. E., Fenoy, L. M., & Moeslund, T. B. (2022). Visual Explanation of Black-box Model: Similarity Difference and Uniqueness (SIDU) Method. *Pattern Recognition*, *127*, 108604. https://doi.org/10.1016/J.PATCOG.2022.108604

Myers, M. D. (2002). *Qualitative Research in Information Systems: A Reader*. SAGE.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2022). How Different Explanations Impact Trust Calibration: The Case of Clinical Decision Support Systems. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4098528

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *ArXiv Preprint ArXiv:1802.00682*.

Narkar, S., Zhang, Y., Liao, Q. V., Wang, D., & Weisz, J. D. (2021). Model LineUpper: Supporting Interactive Model Comparison at Multiple Levels for AutoML. *26th International Conference on Intelligent User Interfaces*, 170–174. https://doi.org/10.1145/3397481.3450658

Narla, A., Kuprel, B., Sarin, K., Novoa, R., & Ko, J. (2018). Automated Classification of Skin Lesions: From Pixels to Practice. *Journal of Investigative Dermatology*, *138*(10), 2108–2110. https://doi.org/10.1016/J.JID.2018.06.175

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *ArXiv Preprint ArXiv:1909.09223*.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., & Gogate, V. (2021). Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. *26th International Conference on Intelligent User Interfaces*, 340–350. https://doi.org/10.1145/3397481.3450639

Oyserman, D., Uskul, A. K., Yoder, N., Nesse, R. M., & Williams, D. R. (2007). Unfair treatment and self-regulatory focus. *Journal of Experimental Social Psychology*, *43*(3), 505–512. https://doi.org/10.1016/j.jesp.2006.05.014

Pavlou, P. A., & Gefen, D. (2005). Psychological Contract Violation in Online Marketplaces: Antecedents, Consequences, and Moderating Role. *Information Systems Research*, *16*(4), 372–399. https://doi.org/10.1287/isre.1050.0065

Payne, J. W. (1976). Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis. *Organizational Behavior and Human Performance*, *16*(2), 366–387. https://doi.org/10.1016/0030-5073(76)90022-2

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pethig, F., & Kroenung, J. (2020). Specialized Information Systems for the Digitally Disadvantaged. *Journal of the Association for Information Systems*, *20*(10), 1412–1446. https://doi.org/10.17705/1jais.00573

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. *ArXiv Preprint ArXiv:1806.07421*. https://arxiv.org/abs/1806.07421

Pfeuffer, N., Baum, L., Stammer, W., Abdel-Karim, B. M., Schramowski, P., Bucher, A. M., Hügel, C., Rohde, G., Kersting, K., & Hinz, O. (2023). Explanatory Interactive Machine Learning. *Business & Information Systems Engineering*, 1–25. https://doi.org/10.1007/S12599-023-00806-X

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four Principles of Explainable Artificial Intelligence*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312

Pinel, E. C. (1999). Stigma consciousness: The psychological legacy of social stereotypes. *Journal of Personality and Social Psychology*, *76*(1), 114–128. https://doi.org/10.1037/0022-3514.76.1.114

Pinel, E. C. (2004). You're Just Saying That Because I'm a Woman: Stigma Consciousness and Attributions to Discrimination. *Self and Identity*, *3*(1), 39–51. https://doi.org/10.1080/13576500342000031

Polley, S., Koparde, R. R., Gowri, A. B., Perera, M., & Nuernberger, A. (2021). Towards Trustworthiness in the Context of Explainable Search. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2580–2584. https://doi.org/10.1145/3404835.3462799

Poole, A., & Ball, L. J. (2006). Eye Tracking in HCI and Usability Research. In C. Ghaoui (Ed.), *Encyclopedia of Human Computer Interaction* (1st ed., pp. 211–219). IGI Global. https://doi.org/10.4018/978-1-59140-562-7.ch034

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and Measuring Model Interpretability. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–52. https://doi.org/10.1145/3411764.3445315

Rasmussen, P. M., Schmah, T., Madsen, K. H., Lund, T. E., Yourganov, G., Strother, S. C., & Hansen, L. K. (2012). Visualization of nonlinear classification models in neuroimaging: Signed sensitivity maps. *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing - BIOSIGNALS 2012*, 254–263.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Ribeiro, M. T. (2016). Lime: Explaining the predictions of any machine learning classifier. In *Github Repository*. https://github.com/marcotcr/lime

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-Agnostic Interpretability of Machine Learning. *ArXiv Preprint ArXiv:1606.05386*. https://doi.org/10.48550/arxiv.1606.05386

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Ribera, M., & Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. *Proceedings of the IUI Workshops*, *2327*, 38.

Robinson, S. L. (1996). Trust and breach of the psychological contract. *Administrative Science Quarterly*, *41*(4), 574–599. https://doi.org/10.2307/2393868

Robnik-Šikonja, M., & Bohanec, M. (2018). Perturbation-Based Explanations of Prediction Models. In J. Zhou & F. Chen (Eds.), *Human and Machine Learning* (1st ed., pp. 159–175). Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0_9

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. https://doi.org/10.1038/s42256-019-0048-x

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/S11263-015-0816-Y/FIGURES/16

Schneider, J., & Handali, J. (2019). Personalized Explanation in Machine Learning: A Conceptualization. *ArXiv Preprint ArXiv:1901.00770*. https://doi.org/10.48550/arxiv.1901.00770

Schoeffer, J., Machowski, Y., & Kuehl, N. (2021). A Study on Fairness and Trust Perceptions in Automated Decision Making. *ArXiv:2103.04757*. https://doi.org/10.48550/arXiv.2103.04757

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. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Sevastjanova, R., Jentner, W., Sperrle, F., Kehlbeck, R., Bernard, J., El-Assady, M., Jentner, W., Sperrle, F., Kehlbeck, R., & El-Assady, M. (2021). QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3–4), 1–38. https://doi.org/10.1145/3429448

Shandilya, A., Dash, A., Chakraborty, A., Ghosh, K., & Ghosh, S. (2020). Fairness for Whom? Understanding the Reader's Perception of Fairness in Text Summarization. *IEEE International Conference on Big Data (Big Data)*, 3692–3701.

Shapley, L. S. (2016). 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*. Princeton University Press. https://doi.org/10.1515/9781400881970-018/HTML

Sharma, S., Zhang, Y., Aliaga, J. M. R. o., Bouneffouf, D., Muthusamy, V., & Varshney, K. R. (2020). Data Augmentation for Discrimination Prevention and Bias Disambiguation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 358–364. https://doi.org/10.1145/3375627.3375865

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature 2016 529:7587*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Smith, B. (2018). *IBM Watson OpenScale: Operate and Automate AI with Trust*. Watson Blog; IBM. https://www.ibm.com/blogs/watson/2018/10/ibm-ai-openscale-operate-and-automate-ai-with-trust/

Sokol, K., & Flach, P. (2020). Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. https://doi.org/10.1145/3351095.3372870

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype Threat. *Annual Review of Psychology*, *67*, 415–437. https://doi.org/10.1146/annurev-psych-073115-103235

Spinner, T., Schlegel, U., Schäfer, H., & El-Assady, M. (2020). ExplAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629

Springer, A., Whittaker, S., Springer, A., & Whittaker, S. (2020). Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Transactions on Interactive Intelligent Systems*, *10*(4), 1–32. https://doi.org/10.1145/3374218

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *ArXiv:2103.12016*. http://arxiv.org/abs/2103.12016

Steele, C. M., & Aronson, J. (1995). Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with Group Image: The Psychology of Stereotype and Social Identity Threat. *Advances in Experimental Social Psychology*, *34*, 379–440. https://doi.org/10.1016/S0065-2601(02)80009-0

Stowell, E., Lyson, M. C., Saksono, H., Wurth, R. C., Jimison, H., Pavel, M., & Parker, A. G. (2018). Designing and Evaluating mHealth Interventions for Vulnerable Populations: A Systematic Review. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3173574.3173589

Strobel, B., Saß, S., Lindner, M. A., & Köller, O. (2016). Do Graph Readers Prefer the Graph Type Most Suited to a Given Task? Insights from Eye Tracking. *Journal of Eye Movement Research*, *9*(4), 1–15. https://doi.org/10.16910/jemr.9.4.4

Sun, Y., & Sundar, S. S. (2022). Exploring the Effects of Interactive Dialogue in Improving User Control for Explainable Online Symptom Checkers. *Conference on Human Factors in Computing Systems - Proceedings*, 1–7. https://doi.org/10.1145/3491101.3519668

Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. *M. J. Metzger and A. J. Flanagin*, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

Sundar, S. S. (2020). Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/JCMC/ZMZ026

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328.

Swartout, W. R. (1983). XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence*, *21*(3), 285–325. https://doi.org/10.1016/S0004-3702(83)80014-9

Swartout, W. R. (1985). Explaining and Justifying Expert Consulting Programs. In *Computer-assisted Medical Decision Making* (Vol. 2, pp. 254–271). Springer New York. https://doi.org/10.1007/978-1-4612-5108-8_15

Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 109–119. https://doi.org/10.1145/3397481.3450662

Tajfel, H. (1982). Social Psychology of Intergroup Relations. *Annual Review of Psychology*, *33*(1), 1–39. https://doi.org/10.1146/annurev.ps.33.020182.000245

Teodorescu, M. H. M., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human–ML Augmentation. *MIS Quarterly*, *45*(3), 1483–1499. https://doi.org/10.25300/MISQ/2021/16535

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. https://doi.org/10.1017/S0140525X00057046

Thomas, L. V., Deng, J., & Brown, E. T. (2021). FacetRules: Discovering and Describing Related Groups. *Proceedings of the 2021 IEEE Workshop on Machine Learning from User Interactions (MLUI)*, 21–26. https://doi.org/10.1109/MLUI54255.2021.00008

Tintarev, N., & Masthoff, J. (2007). A Survey of Explanations in Recommender Systems. *2007 IEEE 23rd International Conference on Data Engineering Workshop*, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

Tjoa, E., & Guan, C. (2019). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793–4813.

Tomprou, M., & Lee, M. K. (2022). Employment Relationships in Algorithmic Management: A Psychological Contract Perspective. *Computers in Human Behavior*, *126*, 106997. https://doi.org/10.1016/J.CHB.2021.106997

Turek, M. (2018). *Explainable artificial intelligence (XAI)*. Defense Advanced Research Projects Agency (DARPA). https://www.darpa.mil/program/explainable-artificial-intelligence

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An Explainable Artificial Intelligence System for Small-unit Tactical Behavior. *Proceedings of the National Conference on Artificial Intelligence*, 900–907. www.aaai.org

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)*, *31*. https://doi.org/10.2139/ssrn.3063289

Wallace, L. E., Wegener, D. T., & Petty, R. E. (2020). When Sources Honestly Provide Their Biased Opinion: Bias as a Distinct Source Perception With Independent Effects on Credibility and Persuasion. *Personality and Social Psychology Bulletin*, *46*(3), 439–453. https://doi.org/10.1177/0146167219858654

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-driven User-Centric Explainable AI. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300831

Wang, R., Harper, F. M., & Zhu, H. (2020). Factors Influencing Perceived Fairness in Algorithmic

Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376813

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, *23*(4), 217–246. https://doi.org/10.2753/MIS0742-1222230410

Wang, W., & Benbasat, I. (2016). Empirical Assessment of Alternative Designs for Enhancing Different Types of Trusting Beliefs in Online Recommendation Agents. *Journal of Management Information Systems*, *33*(3), 744–775. https://doi.org/10.1080/07421222.2016.1243949

Wang, W., & Wang, M. (2019). Effects of Sponsorship Disclosure on Perceived Integrity of Biased Recommendation Agents: Psychological Contract Violation and Knowledge-Based Trust Perspectives. *Information Systems Research*, *30*(2), 507–522. https://doi.org/10.1287/isre.2018.0811

Wang, W., Xu, J., & Wang, M. (2018). Effects of Recommendation Neutrality and Sponsorship Disclosure on Trust vs. Distrust in Online Recommendation Agents: Moderating Role of Explanations for Organic Recommendations. *Management Science*, *64*(11), 5198–5219. https://doi.org/10.1287/mnsc.2017.2906

Weerts, H. J. P., van Ipenburg, W., & Pechenizkiy, M. (2019). A Human-Grounded Evaluation of SHAP for Alert Processing. *ArXiv Preprint ArXiv:1907.03324*.

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM*, *62*(6), 70–79. https://doi.org/10.1145/3282486

WHI. (2020). *2020 Workshop on Human Interpretability in Machine Learning (WHI)*. WHI. https://sites.google.com/view/whi2020/home

Xu, K., Lei Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning*, 2048–2057.

Yan, J. N., Gu, Z., Lin, H., & Rzeszotarski, J. M. (2020). Silva: Interactively Assessing Machine Learning Fairness Using Causality. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces*, *20*, 189–201. https://doi.org/10.1145/3377325.3377480

Yap, B. W., Rani, K. A., Abd Rahman, H. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets. *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, *285 LNEE*, 13–22. https://doi.org/10.1007/978-981-4585-18-7_2

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Yuan, J., Nov, O., & Bertini, E. (2021). An Exploration and Validation of Visual Factors in Understanding Classification Rule Sets. *Proceedings of the 2021 IEEE Visualization Conference (VIS)*, 6–10. https://doi.org/10.1109/VIS49827.2021.9623303

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *ArXiv Preprint ArXiv:1904.12991*.

Zhou, B., Sun, Y., Bau, D., & Torralba, A. (2018). Interpretable Basis Decomposition for Visual Explanation. *Proceedings of the European Conference on Computer Vision (ECCV)*, 119–134.

Zhou, J., Arshad, S. Z., Luo, S., & Chen, F. (2017). *Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making* (pp. 23–39). Springer, Cham. https://doi.org/10.1007/978-3-

319-68059-0_2

Appendix

Appendix A: Study I

Appendix A1: German Credit Dataset Attributes

	Description
Amount (EUR)	Loan amount of the application in EUR
Duration (months)	Duration of loan application repayment in months
Guarantee	Information about additional people that act as a guarantee for the loan
Purpose	Purpose of the loan
Account Balance	Status of the checking account of the applicant
Assets	Most valuable available assets of the applicant
Available Income	Percentage of available income after fix costs
Housing	Information whether the housing of the applicant is owned or rented
Loan History	Information about the payment of previous loans
Number of Previous Loans	Number of previous loans of the applicant at this bank
Other Loans	Information about additional loans
Savings Account	Status of the savings account of the applicant
Age (years)	Age of the applicant in years
Employment	Duration of present employment
Job	Type of job of the applicant
Number of dependents	Number of people that depend financially on the applicant
Residence Duration	Duration living in current residence
Telephone	Information whether a telephone line is registered under the applicants name

Loan Details Financial Status Personal Information

Figure 37: Description of German credit dataset attributes used to train AI system.

Appendix A2: Measures

All measured on a 7-Likert scale Strongly disagree – Strongly agree, unless stated otherwise.

Table 14: Measures used in Study 1.

Measure	**Items**
Constructs	
Trust (adapted from Hoffman et al., 2018)	I am confident in the AI system. I feel that it works well.
	The recommendations of the AI system are very predictable.
	The AI system is very reliable. I can count on it to be correct all the time.
	I feel safe that when bank employees rely on the AI system, they will get the right answers.
	I am skeptical of the AI system.
	The AI system can perform the task of deciding loan applications better than a novice human user.
Understandability (adapted from Madsen & Gregor, 2000)	The system uses appropriate methods to provide explanations for decision recommendations.
	The system has good knowledge about this type of problem built into it.
	The system produces explanations for decision recommendations that are as good as those which a highly competent person could produce.
	The system makes use of all the knowledge and information available to it to produce explanations for decision recommendations.
Explanation Satisfaction (adapted from Hoffman et al., 2018)	From the explanations, I understand how the AI system makes recommendations.
	The explanations of how the AI system makes recommendations are satisfying.
	The explanations of how the AI system makes recommendations have sufficient detail.
	The explanations of how the AI system makes recommendations seem complete.
	The explanations of how the AI system makes recommendations are useful to my goals.

Measure	Items
Controls	
Domain knowledge (adapted from (heng et al., 2019)	How much experience have you had in the past with tasks similar to the evaluation of loan applications? No experience; A little experience; Some experience; A lot of experience.
Machine learning knowledge (adapted from Cheng et al., 2019)	How much knowledge of machine learning do you have? No knowledge; A little knowledge – I know basic concepts in machine learning; Some knowledge – I have used machine learning before; A lot of knowledge – I apply machine learning frequently to my work or I create machine learning applications.
Programming knowledge (adapted from Cheng et al., 2019)	How much knowledge in programming knowledge do you have? No knowledge; A little knowledge - I know basic programming concepts; Some knowledge - I have coded a few programs before; A lot of knowledge - I code programs frequently.
Technical literacy (adapted from Cheng et al., 2019)	I am confident using computers.
	I can make use of computer programming to solve a problem.
	I understand how Amazon recommends products for me to purchase.
	I use computers whenever I can.
	I understand how my credit score is calculated.
	I understand how my email provider's spam filter works.

Appendix A3: Statistical Analyses of Iterative Design Process

Table 15: Independent Kruskal-Wallis test for trust in the evaluation of the second design iteration.

Independent-Samples Kruskal-Wallis Test for Trust Summary	
Total N	235
Test Statistic	1.751[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.626
a. The test statistic is adjusted for ties.	

Table 16: Independent Kruskal-Wallis test for understandability in the evaluation of the second design iteration.

Independent-Samples Kruskal-Wallis Test for Understandability Summary	
Total N	235
Test Statistic	0.805[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.848
a. The test statistic is adjusted for ties.	

Table 17: Independent Kruskal-Wallis test for forward-prediction score in the evaluation of the second design iteration.

Independent-Samples Kruskal-Wallis Test for Forward-prediction Score Summary	
Total N	235
Test Statistic	9.963[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	**0.019**
a. The test statistic is adjusted for ties.	

Table 18: Post-hoc pairwise comparison with Bonferroni correction for the forward-prediction score in the evaluation of the second design iteration.

Pairwise Comparisons of Explanation Representation for Forward-prediction Score					
Sample 1-Sample 2	**Test Statistic**	**Std. Error**	**Std. Test Statistic**	**Sig.**	**Adj. Sig.[a]**
SHAP-LIME	18.154	12.270	1.480	0.139	0.834
SHAP-DICE	20.913	12.270	1.704	0.088	0.530
SHAP-Anchors	**38.950**	**12.372**	**3.148**	**0.002**	**0.010**
LIME-DICE	2.758	12.111	0.228	0.820	1.000
LIME-Anchors	20.795	12.215	1.702	0.089	0.532
DICE-Anchors	18.037	12.215	1.477	0.140	0.839
Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .050. a. Significance values have been adjusted by the Bonferroni correction for multiple tests.					

Table 19: Independent Kruskal-Wallis test for satisfaction in the evaluation of the third design iteration.

Independent-Samples Kruskal-Wallis Test for Satisfaction Summary	
Total N	252
Test Statistic	1.415[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.702
a. The test statistic is adjusted for ties.	

Table 20: Independent Kruskal-Wallis test for trust in the evaluation of the third design iteration.

Independent-Samples Kruskal-Wallis Test for Trust Summary	
Total N	252
Test Statistic	1.897[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.594
a. The test statistic is adjusted for ties.	

Table 21: Independent Kruskal-Wallis test for understandability in the evaluation of the third design iteration.

Independent-Samples Kruskal-Wallis Test for Understandability Summary	
Total N	252
Test Statistic	1.381[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.710
a. The test statistic is adjusted for ties.	

Table 22: Independent Kruskal-Wallis test for forward-prediction score in the evaluation of the third design iteration.

Independent-Samples Kruskal-Wallis Test for Forward-prediction Score Summary	
Total N	252
Test Statistic	6.056[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.109
a. The test statistic is adjusted for ties.	

Appendix A4: Descriptive Statistics for Control Variables in Eye-Tracking Experiment

Table 23: Mean and standard deviation for control variables in the eye-tracking experiment.

Control Variable	Mean	SD
Domain Knowledge	2.26	0.991
Machine Learning Knowledge	2.11	0.809
Programming Knowledge	2.79	1.032
Technical Literacy	5.6579	0.754
Age	23.32	2.810

Table 24: Distribution of categories for control variables in the eye-tracking experiment.

Control Variable	Category	Frequency	Percentage
Domain Knowledge	No knowledge	3	15.8
	Slightly familiar	11	57.9
	Somewhat familiar	3	15.8
	Moderately familiar	1	5.3
	Extremely familiar	1	5.3
	Total	**19**	**100**
Machine Learning Knowledge	No knowledge	4	21.1
	A little knowledge	10	52.6
	Some knowledge	4	21.1
	A lot of knowledge	1	5.3
	Total	**19**	**100**
Programming Knowledge	No knowledge	2	10.5
	A little knowledge	6	31.6
	Some knowledge	5	26.3
	A lot of knowledge	6	31.6
	Total	**19**	**100**
Gender	Female	6	31.6
	Male	13	68.4
	Total	**19**	**100**
Area of Study	Biochemistry	1	5.3
	Chemical Engineering	3	15.8
	Civil Engineering	1	5.3
	Computer Science	3	15.8
	Industrial Engineering	6	31.6
	Information Systems	1	5.3
	International Business	1	5.3
	Mathematics	1	5.3
	Mechanical Engineering	2	10.5
	Total	**19**	**100**

Appendix A5: Statistical Analyses of Self-reported Measures Eye-Tracking Experiment

Table 25: Repeated measures ANCOVA analysis for satisfaction in the eye-tracking experiment.

Tests of Within-Subjects Effects for Satisfaction							
Measure: Satisfaction							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	Sphericity Assumed	1.393	3	0.464	0.382	0.766	0.031
Explanation Representation * Domain Knowledge	Sphericity Assumed	5.353	3	1.784	1.468	0.239	0.109
Explanation Representation * Technical Literacy	Sphericity Assumed	4.610	3	1.537	1.265	0.301	0.095
Explanation Representation * ML Knowledge	**Sphericity Assumed**	**10.938**	**3**	**3.646**	**3.001**	**0.043**	**0.200**
Explanation Representation * Programming Knowledge	Sphericity Assumed	4.510	3	1.503	1.237	0.310	0.093
Explanation Representation * Gender	Sphericity Assumed	1.761	3	0.587	0.483	0.696	0.039
Explanation Representation * Age	Sphericity Assumed	1.761	3	0.587	0.483	0.696	0.039
Error (Explanation Representation)	Sphericity Assumed	43.743	36	1.215			

Table 26: Friedman test for usefulness in the eye-tracking experiment.

Friedman Test Usefulness	
Ranks	
	Mean Rank
Usefulness Anchors	1.970
Usefulness DICE	2.370
Usefulness Grade LIME	2.610
Usefulness SHAP	3.050
Test Statistics[a]	
N	19
Chi-Square	7.190
df	3
Asymp. Sig.	**0.066**
a. Friedman Test	

Table 27: Post-hoc pairwise Wilcoxon signed-ranks test for usefulness in the eye-tracking experiment.

Wilcoxon Signed Ranks Test Usefulness				
Ranks				
		N	Mean Rank	Sum of Ranks
Usefulness DICE - Usefulness Anchors	Negative Ranks	7[a]	10.29	72.00
	Positive Ranks	11[b]	9.00	99.00
	Ties	1[c]		
	Total	19		
Usefulness Grade LIME - Usefulness Anchors	Negative Ranks	7[d]	7.50	52.50
	Positive Ranks	12[e]	11.46	137.50
	Ties	0[f]		
	Total	19		
Usefulness SHAP - Usefulness Anchors	Negative Ranks	3[g]	11.17	33.50
	Positive Ranks	14[h]	8.54	119.50
	Ties	2[i]		
	Total	19		
Usefulness Grade LIME - Usefulness DICE	Negative Ranks	8[j]	8.00	64.00
	Positive Ranks	11[k]	11.45	126.00
	Ties	0[l]		
	Total	19		
Usefulness SHAP - Usefulness DICE	Negative Ranks	6[m]	7.67	46.00
	Positive Ranks	12[n]	10.42	125.00
	Ties	1[o]		
	Total	19		
Usefulness SHAP - Usefulness Grade LIME	Negative Ranks	7[p]	8.86	62.00
	Positive Ranks	11[q]	9.91	109.00
	Ties	1[r]		
	Total	19		

a. Usefulness DICE < Usefulness Anchors
b. Usefulness DICE > Usefulness Anchors
c. Usefulness DICE = Usefulness Anchors
d. Usefulness Grade LIME < Usefulness Anchors
e. Usefulness Grade LIME > Usefulness Anchors
f. Usefulness Grade LIME = Usefulness Anchors
g. Usefulness SHAP < Usefulness Anchors
h. Usefulness SHAP > Usefulness Anchors
i. Usefulness SHAP = Usefulness Anchors
j. Usefulness Grade LIME < Usefulness DICE
k. Usefulness Grade LIME > Usefulness DICE
l. Usefulness Grade LIME = Usefulness DICE
m. Usefulness SHAP < Usefulness DICE
n. Usefulness SHAP > Usefulness DICE
o. Usefulness SHAP = Usefulness DICE
p. Usefulness SHAP < Usefulness Grade LIME
q. Usefulness SHAP > Usefulness Grade LIME
r. Usefulness SHAP = Usefulness Grade LIME

Test Statistics[a]		
	Z	Asymp. Sig.
Usefulness DICE - Usefulness Anchors	-0.592[b]	0.554
Usefulness Grade LIME - Usefulness Anchors	-1.715[b]	0.086
Usefulness SHAP - Usefulness Anchors	-2.042[b]	0.041
Usefulness Grade LIME - Usefulness DICE	-1.254[b]	0.210
Usefulness SHAP - Usefulness DICE	-1.723[b]	0.085
Usefulness SHAP - Usefulness Grade LIME	-1.031[b]	0.302

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.

Table 28: Holm-Bonferroni correction for post-hoc pairwise Wilcoxon signed-ranks test for usefulness in the eye-tracking experiment.

Holm-Bonferroni Correction for Wilcoxon Signed Ranks Test Usefulness					
	Asymp. Sig. (Ascending Order)	**i**	**Adjusted α Formula**	**Adjusted α (α = 0.05)**	**Significant**
Usefulness SHAP - Usefulness Anchors	0.041	1	$\frac{\alpha}{m+1-i}$	0.008	No
Usefulness SHAP - Usefulness DICE	0.085	2		0.010	No
Usefulness LIME - Usefulness Anchors	0.086	3		0.013	No
Usefulness LIME - Usefulness DICE	0.210	4		0.017	No
Usefulness SHAP - Usefulness LIME	0.302	5		0.025	No
Usefulness DICE - Usefulness Anchors	0.554	6		0.050	No

Table 29: Repeated Measures ANCOVA analysis for usefulness in the eye-tracking experiment.

Tests of Within-Subjects Effects for Usefulness							
Measure: Usefulness							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	Sphericity Assumed	14.479	3	4.826	1.640	0.197	0.120
Explanation Representation * Domain Knowledge	Sphericity Assumed	16.635	3	5.545	1.885	0.150	0.136
Explanation Representation * Technical Literacy	Sphericity Assumed	9.496	3	3.165	1.076	0.372	0.082
Explanation Representation * ML Knowledge	Sphericity Assumed	25.536	3	8.512	2.893	0.049	0.194
Explanation Representation * Programming Knowledge	Sphericity Assumed	20.369	3	6.790	2.308	0.093	0.161
Explanation Representation * Gender	**Sphericity Assumed**	**90.564**	**3**	**30.188**	**10.260**	**<0.001**	**0.461**
Explanation Representation * Age	Sphericity Assumed	2.178	3	0.726	0.247	0.863	0.020
Error (Explanation Representation)	Sphericity Assumed	105.924	36	2.942			

Table 30: Friedman test for rank of explanation representation in the eye-tracking experiment.

Friedman Test Ranks	
Ranks	
	Mean Rank
Rank Anchors	1.840
Rank DICE	2.470
Rank LIME	2.740
Rank SHAP	2.950
Test Statistics[a]	
N	19
Chi-Square	7.863
df	3
Asymp. Sig.	**0.049**
a. Friedman Test	

Table 31: Post-hoc pairwise Wilcoxon signed-ranks test for rank of explanation representation in the eye-tracking experiment.

Wilcoxon Signed Ranks Test Rank of Explanation Representation				
Ranks				
		N	Mean Rank	Sum of Ranks
Rank DICE - Rank Anchors	Negative Ranks	6[a]	8.92	53.50
	Positive Ranks	13[b]	10.50	136.50
	Ties	0[c]		
	Total	19		
Rank LIME - Rank Anchors	Negative Ranks	6[d]	6.00	36.00
	Positive Ranks	13[e]	11.85	154.00
	Ties	0[f]		
	Total	19		
Rank SHAP - Rank Anchors	Negative Ranks	4[g]	9.50	38.00
	Positive Ranks	15[h]	10.13	152.00
	Ties	0[i]		
	Total	19		
Rank LIME - Rank DICE	Negative Ranks	9[j]	8.78	79.00
	Positive Ranks	10[k]	11.10	111.00
	Ties	0[l]		
	Total	19		
Rank SHAP - Rank DICE	Negative Ranks	6[m]	11.92	71.50
	Positive Ranks	13[n]	9.12	118.50
	Ties	0[o]		
	Total	19		
Rank SHAP - Rank LIME	Negative Ranks	10[p]	8.35	83.50
	Positive Ranks	9[q]	11.83	106.50
	Ties	0[r]		
	Total	19		
a. Rank DICE < Rank Anchors b. Rank DICE > Rank Anchors c. Rank DICE = Rank Anchors d. Rank LIME < Rank Anchors e. Rank LIME > Rank Anchors f. Rank LIME = Rank Anchors g. Rank SHAP < Rank Anchors h. Rank SHAP > Rank Anchors i. Rank SHAP = Rank Anchors j. Rank LIME < Rank DICE k. Rank LIME > Rank DICE l. Rank LIME = Rank DICE m. Rank SHAP < Rank DICE n. Rank SHAP > Rank DICE o. Rank SHAP = Rank DICE p. Rank SHAP < Rank LIME q. Rank SHAP > Rank LIME r. Rank SHAP = Rank LIME				
Test Statistics[a]				
	Z		Asymp. Sig.	
Rank DICE - Rank Anchors	-1.754[b]		0.079	
Rank LIME - Rank Anchors	-2.438[b]		0.015	
Rank SHAP - Rank Anchors	-2.343[b]		0.019	
Rank LIME - Rank DICE	-0.657[b]		0.511	
Rank SHAP - Rank DICE	-0.959[b]		0.337	
Rank SHAP - Rank LIME	-0.486[b]		0.627	
a. Wilcoxon Signed Ranks Test b. Based on negative ranks.				

Table 32: Holm-Bonferroni correction for post-hoc pairwise Wilcoxon signed-ranks test for rank of explanation representation in the eye-tracking experiment.

Holm-Bonferroni Correction for Wilcoxon Signed Ranks Test for Rank of Explanation Representation					
	Asymp. Sig. (Ascending Order)	**i**	**Adjusted α Formula**	**Adjusted α (α = 0.05)**	**Significant**
Rank LIME - Rank Anchors	0.015	1	$\frac{\alpha}{m+1-i}$	0.008	No
Rank SHAP - Rank Anchors	0.019	2		0.010	No
Rank DICE - Rank Anchors	0.079	3		0.013	No
Rank SHAP - Rank DICE	0.337	4		0.017	No
Rank LIME - Rank DICE	0.511	5		0.025	No
Rank SHAP - Rank LIME	0.627	6		0.050	No

Appendix A6: Definition of AOIs for Eye-tracking Evaluation

Table 33: AOI for each explanation representation for eye-tracking analysis.

	Anchors	**DICE**	**LIME**	**SHAP**
AOI1	Attributes Information Table			
AOI2	Explanation Representation			
AOI2.1	First rule	First Counterfactual	Top influencing attributes	
AOI2.2	Second rule	Second Counterfactual	Bottom influencing attributes	
AOI2.3	Third rule	Third Counterfactual	Negative influencing attributes	
AOI2.4	Fourth rule		Positive influencing attributes	
AOI2.5	Fifth rule			Probability area

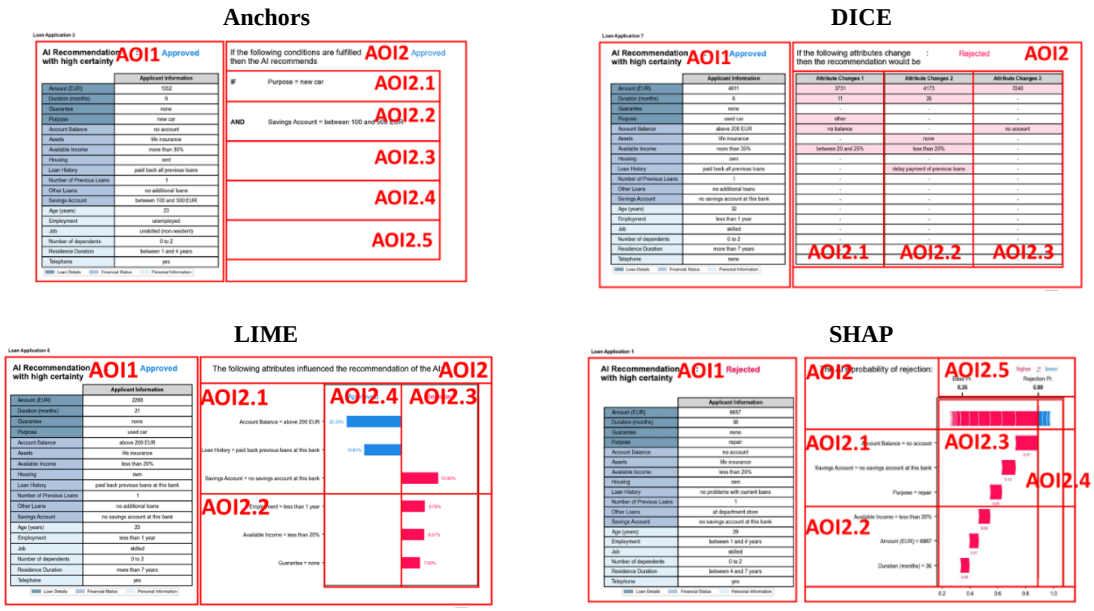


Figure 38: Visualization of AOI defined for the explanation representations for Anchors, DICE, LIME, and SHAP.

Appendix A7: Statistical Analyses of Eye-Tracking Data

Table 34: Friedman test for fixation duration in complete visualization in the eye-tracking experiment.

Friedman Test Fixation Duration in Complete Visualization	
Ranks	
	Mean Rank
Anchors Rejected - Total Fixation Duration	3.950
Anchors Approved - Total Fixation Duration	2.530
DICE Rejected - Total Fixation Duration	4.790
DICE Approved - Total Fixation Duration	5.790
LIME Rejected - Total Fixation Duration	5.050
LIME Approved - Total Fixation Duration	5.580
SHAP Rejected - Total Fixation Duration	4.260
SHAP Approved - Total Fixation Duration	4.050
Test Statistics[a]	
N	19
Chi-Square	24.298
df	7
Asymp. Sig.	**0.001**
a. Friedman Test	

Table 35: Post-hoc pairwise Wilcoxon signed-ranks test for fixation duration in complete visualization in the eye-tracking experiment.

Wilcoxon Signed Fixation Duration in Complete Visualization				
Ranks				
		N	Mean Rank	Sum of Ranks
Anchors Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	13[a]	12.08	157.00
	Positive Ranks	6[b]	5.50	33.00
	Ties	0[c]		
	Total	19		
DICE Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	6[d]	9.67	58.00
	Positive Ranks	13[e]	10.15	132.00
	Ties	0[f]		
	Total	19		
DICE Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	5[g]	5.40	27.00
	Positive Ranks	14[h]	11.64	163.00
	Ties	0[i]		
	Total	19		
LIME Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	6[j]	10.33	62.00
	Positive Ranks	13[k]	9.85	128.00
	Ties	0[l]		
	Total	19		
LIME Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	8[m]	5.88	47.00
	Positive Ranks	11[n]	13.00	143.00
	Ties	0[o]		
	Total	19		
SHAP Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	8[p]	8.50	68.00
	Positive Ranks	11[q]	11.09	122.00
	Ties	0[r]		
	Total	19		
SHAP Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	Negative Ranks	10[s]	8.00	80.00
	Positive Ranks	9[t]	12.22	110.00
	Ties	0[u]		
	Total	19		
DICE Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**Negative Ranks**	**4[v]**	**3.75**	**15.00**
	Positive Ranks	**15[w]**	**11.67**	**175.00**
	Ties	**0[x]**		
	Total	**19**		

DICE Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**Negative Ranks**	**2[y]**	**2.5**	**5**
	Positive Ranks	**17[z]**	**10.88**	**185**
	Ties	**0[aa]**		
	Total	**19**		
LIME Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	Negative Ranks	4[ab]	5.25	21
	Positive Ranks	15[ac]	11.27	169
	Ties	0[ad]		
	Total	19		
LIME Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**Negative Ranks**	**3[ae]**	**4.67**	**14**
	Positive Ranks	**16[af]**	**11**	**176**
	Ties	**0[ag]**		
	Total	**19**		
SHAP Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**Negative Ranks**	**3[ah]**	**3.67**	**11**
	Positive Ranks	**16[ai]**	**11.19**	**179**
	Ties	**0[aj]**		
	Total	**19**		
SHAP Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	Negative Ranks	7[ak]	6.29	44
	Positive Ranks	12[al]	12.17	146
	Ties	0[am]		
	Total	19		
DICE Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	Negative Ranks	8[an]	6.63	53
	Positive Ranks	11[ao]	12.45	137
	Ties	0[ap]		
	Total	19		
LIME Rejected - Total Fixation Duration - DICE Rejected - Total Fixation Duration	Negative Ranks	9[aq]	10.22	92
	Positive Ranks	10[ar]	9.8	98
	Ties	0[as]		
	Total	19		
LIME Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	Negative Ranks	7[at]	10.43	73
	Positive Ranks	12[au]	9.75	117
	Ties	0[av]		
	Total	19		
SHAP Rejected - Total Fixation Duration - DICE Rejected - Total Fixation Duration	Negative Ranks	11[aw]	10.27	113
	Positive Ranks	8[ax]	9.63	77
	Ties	0[ay]		
	Total	19		
SHAP Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	Negative Ranks	9[az]	11.33	102
	Positive Ranks	10[ba]	8.8	88
	Ties	0[bb]		
	Total	19		
LIME Rejected - Total Fixation Duration - DICE Approved - Total Fixation Duration	Negative Ranks	11[bc]	10.82	119
	Positive Ranks	8[bc]	8.88	71
	Ties	0[be]		
	Total	19		
LIME Approved - Total Fixation Duration - DICE Approved - Total Fixation Duration	Negative Ranks	9[bf]	11.56	104
	Positive Ranks	10[bg]	8.6	86
	Ties	0[bh]		
	Total	19		
SHAP Rejected - Total Fixation Duration - DICE Approved - Total Fixation Duration	Negative Ranks	15[bi]	10.27	154
	Positive Ranks	4[bj]	9	36
	Ties	0[bk]		
	Total	19		
SHAP Approved - Total Fixation Duration - DICE Approved - Total Fixation Duration	Negative Ranks	14[bl]	9.86	138
	Positive Ranks	5[bm]	10.4	52
	Ties	0[bn]		
	Total	19		
LIME Approved - Total Fixation Duration - LIME Rejected - Total Fixation Duration	Negative Ranks	9[bo]	8.78	79
	Positive Ranks	10[bp]	11.1	111
	Ties	0[bq]		
	Total	19		
SHAP Rejected - Total Fixation Duration - LIME Rejected - Total Fixation Duration	Negative Ranks	10[br]	11.1	111
	Positive Ranks	9[bs]	8.78	79
	Ties	0[bt]		
	Total	19		

SHAP Approved - Total Fixation Duration - LIME Rejected - Total Fixation Duration	Negative Ranks	12[bu]	9.92	119
	Positive Ranks	7[bv]	10.14	71
	Ties	0[bw]		
	Total	19		
SHAP Rejected - Total Fixation Duration - LIME Approved - Total Fixation Duration	Negative Ranks	14[bx]	9.93	139
	Positive Ranks	5[by]	10.2	51
	Ties	0[bz]		
	Total	19		
SHAP Approved - Total Fixation Duration - LIME Approved - Total Fixation Duration	Negative Ranks	14[ca]	10	140
	Positive Ranks	5[cb]	10	50
	Ties	0[cc]		
	Total	19		
SHAP Approved - Total Fixation Duration - SHAP Rejected - Total Fixation Duration	Negative Ranks	9[cd]	10.67	96
	Positive Ranks	10[ce]	9.4	94
	Ties	0[cf]		
	Total	19		

a. Anchors Approved - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
b. Anchors Approved - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
c. Anchors Approved - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
d. DICE Rejected - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
e. DICE Rejected - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
f. DICE Rejected - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
g. DICE Approved - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
h. DICE Approved - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
i. DICE Approved - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
j. LIME Rejected - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
k. LIME Rejected - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
l. LIME Rejected - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
m. LIME Approved - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
n. LIME Approved - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
o. LIME Approved - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
p. SHAP Rejected - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
q. SHAP Rejected - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
r. SHAP Rejected - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
s. SHAP Approved - Total Fixation Duration < Anchors Rejected - Total Fixation Duration
t. SHAP Approved - Total Fixation Duration > Anchors Rejected - Total Fixation Duration
u. SHAP Approved - Total Fixation Duration = Anchors Rejected - Total Fixation Duration
v. DICE Rejected - Total Fixation Duration < Anchors Approved - Total Fixation Duration
w. DICE Rejected - Total Fixation Duration > Anchors Approved - Total Fixation Duration
x. DICE Rejected - Total Fixation Duration = Anchors Approved - Total Fixation Duration
y. DICE Approved - Total Fixation Duration < Anchors Approved - Total Fixation Duration
z. DICE Approved - Total Fixation Duration > Anchors Approved - Total Fixation Duration
aa. DICE Approved - Total Fixation Duration = Anchors Approved - Total Fixation Duration
ab. LIME Rejected - Total Fixation Duration < Anchors Approved - Total Fixation Duration
ac. LIME Rejected - Total Fixation Duration > Anchors Approved - Total Fixation Duration
ad. LIME Rejected - Total Fixation Duration = Anchors Approved - Total Fixation Duration
ae. LIME Approved - Total Fixation Duration < Anchors Approved - Total Fixation Duration
af. LIME Approved - Total Fixation Duration > Anchors Approved - Total Fixation Duration
ag. LIME Approved - Total Fixation Duration = Anchors Approved - Total Fixation Duration
ah. SHAP Rejected - Total Fixation Duration < Anchors Approved - Total Fixation Duration
ai. SHAP Rejected - Total Fixation Duration > Anchors Approved - Total Fixation Duration
aj. SHAP Rejected - Total Fixation Duration = Anchors Approved - Total Fixation Duration
ak. SHAP Approved - Total Fixation Duration < Anchors Approved - Total Fixation Duration
al. SHAP Approved - Total Fixation Duration > Anchors Approved - Total Fixation Duration
am. SHAP Approved - Total Fixation Duration = Anchors Approved - Total Fixation Duration
an. DICE Approved - Total Fixation Duration < DICE Rejected - Total Fixation Duration
ao. DICE Approved - Total Fixation Duration > DICE Rejected - Total Fixation Duration
ap. DICE Approved - Total Fixation Duration = DICE Rejected - Total Fixation Duration
aq. LIME Rejected - Total Fixation Duration < DICE Rejected - Total Fixation Duration
ar. LIME Rejected - Total Fixation Duration > DICE Rejected - Total Fixation Duration
as. LIME Rejected - Total Fixation Duration = DICE Rejected - Total Fixation Duration
at. LIME Approved - Total Fixation Duration < DICE Rejected - Total Fixation Duration
au. LIME Approved - Total Fixation Duration > DICE Rejected - Total Fixation Duration
av. LIME Approved - Total Fixation Duration = DICE Rejected - Total Fixation Duration
aw. SHAP Rejected - Total Fixation Duration < DICE Rejected - Total Fixation Duration
ax. SHAP Rejected - Total Fixation Duration > DICE Rejected - Total Fixation Duration
ay. SHAP Rejected - Total Fixation Duration = DICE Rejected - Total Fixation Duration
az. SHAP Approved - Total Fixation Duration < DICE Rejected - Total Fixation Duration
ba. SHAP Approved - Total Fixation Duration > DICE Rejected - Total Fixation Duration
bb. SHAP Approved - Total Fixation Duration = DICE Rejected - Total Fixation Duration
bc. LIME Rejected - Total Fixation Duration < DICE Approved - Total Fixation Duration
bd. LIME Rejected - Total Fixation Duration > DICE Approved - Total Fixation Duration
be. LIME Rejected - Total Fixation Duration = DICE Approved - Total Fixation Duration

bf. LIME Approved - Total Fixation Duration < DICE Approved - Total Fixation Duration
bg. LIME Approved - Total Fixation Duration > DICE Approved - Total Fixation Duration
bh. LIME Approved - Total Fixation Duration = DICE Approved - Total Fixation Duration
bi. SHAP Rejected - Total Fixation Duration < DICE Approved - Total Fixation Duration
bj. SHAP Rejected - Total Fixation Duration > DICE Approved - Total Fixation Duration
bk. SHAP Rejected - Total Fixation Duration = DICE Approved - Total Fixation Duration
bl. SHAP Approved - Total Fixation Duration < DICE Approved - Total Fixation Duration
bm. SHAP Approved - Total Fixation Duration > DICE Approved - Total Fixation Duration
bn. SHAP Approved - Total Fixation Duration = DICE Approved - Total Fixation Duration
bo. LIME Approved - Total Fixation Duration < LIME Rejected - Total Fixation Duration
bp. LIME Approved - Total Fixation Duration > LIME Rejected - Total Fixation Duration
bq. LIME Approved - Total Fixation Duration = LIME Rejected - Total Fixation Duration
br. SHAP Rejected - Total Fixation Duration < LIME Rejected - Total Fixation Duration
bs. SHAP Rejected - Total Fixation Duration > LIME Rejected - Total Fixation Duration
bt. SHAP Rejected - Total Fixation Duration = LIME Rejected - Total Fixation Duration
bu. SHAP Approved - Total Fixation Duration < LIME Rejected - Total Fixation Duration
bv. SHAP Approved - Total Fixation Duration > LIME Rejected - Total Fixation Duration
bw. SHAP Approved - Total Fixation Duration = LIME Rejected - Total Fixation Duration
bx. SHAP Rejected - Total Fixation Duration < LIME Approved - Total Fixation Duration
by. SHAP Rejected - Total Fixation Duration > LIME Approved - Total Fixation Duration
bz. SHAP Rejected - Total Fixation Duration = LIME Approved - Total Fixation Duration
ca. SHAP Approved - Total Fixation Duration < LIME Approved - Total Fixation Duration
cb. SHAP Approved - Total Fixation Duration > LIME Approved - Total Fixation Duration
cc. SHAP Approved - Total Fixation Duration = LIME Approved - Total Fixation Duration
cd. SHAP Approved - Total Fixation Duration < SHAP Rejected - Total Fixation Duration
ce. SHAP Approved - Total Fixation Duration > SHAP Rejected - Total Fixation Duration
cf. SHAP Approved - Total Fixation Duration = SHAP Rejected - Total Fixation Duration

Test Statistics[a]

	Z	Asymp. Sig.
Anchors Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-2.495[b]	0.013
DICE Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-1.489[c]	0.136
DICE Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-2.736[c]	0.006
LIME Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-1.328[c]	0.184
LIME Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-1.932[c]	0.053
SHAP Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-1.087[c]	0.277
SHAP Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	-.604[c]	0.546
DICE Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**-3.220[c]**	**0.001**
DICE Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**-3.622[c]**	**<0.001**
LIME Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	-2.978[c]	0.003
LIME Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**-3.260[c]**	**0.001**
SHAP Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**-3.380[c]**	**<0.001**
SHAP Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	-2.052[c]	0.04
DICE Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	-1.690[c]	0.091
LIME Rejected - Total Fixation Duration - DICE Rejected - Total Fixation Duration	-.121[c]	0.904
LIME Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	-.885[c]	0.376
SHAP Rejected - Total Fixation Duration - DICE Rejected - Total Fixation Duration	-.724[b]	0.469
SHAP Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	-.282[b]	0.778

LIME Rejected - Total Fixation Duration - DICE Approved - Total Fixation Duration	-.966[b]	0.334
LIME Approved - Total Fixation Duration - DICE Approved - Total Fixation Duration	-.362[b]	0.717
SHAP Rejected - Total Fixation Duration - DICE Approved - Total Fixation Duration	-2.374[b]	0.018
SHAP Approved - Total Fixation Duration - DICE Approved - Total Fixation Duration	-1.730[b]	0.084
LIME Approved - Total Fixation Duration - LIME Rejected - Total Fixation Duration	-.644[c]	0.52
SHAP Rejected - Total Fixation Duration - LIME Rejected - Total Fixation Duration	-.644[b]	0.52
SHAP Approved - Total Fixation Duration - LIME Rejected - Total Fixation Duration	-.966[b]	0.334
SHAP Rejected - Total Fixation Duration - LIME Approved - Total Fixation Duration	-1.771[b]	0.077
SHAP Approved - Total Fixation Duration - LIME Approved - Total Fixation Duration	-1.811[b]	0.07
SHAP Approved - Total Fixation Duration - SHAP Rejected - Total Fixation Duration	-.040[b]	0.968
a. Wilcoxon Signed Ranks Test b. Based on positive ranks. c Based on negative ranks.		

Table 36: Holm-Bonferroni correction for post-hoc pairwise Wilcoxon signed-ranks test for fixation duration in complete visualization in the eye-tracking experiment.

Holm-Bonferroni Correction for Wilcoxon Signed Ranks Test Fixation Duration in Complete Visualization					
	Asymp. Sig. (Ascending Order)	**i**	**Adjusted α Formula**	**Adjusted α (α = 0.05)**	**Significant**
DICE Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**0.001**	**1**		**0.002**	**Yes**
SHAP Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**0.001**	**2**		**0.002**	**Yes**
DICE Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**0.001**	**3**		**0.002**	**Yes**
LIME Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	**0.001**	**4**		**0.002**	**Yes**
LIME Rejected - Total Fixation Duration - Anchors Approved - Total Fixation Duration	0.003	5		0.002	No
DICE Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.006	6		0.002	No
Anchors Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.013	7		0.002	No
SHAP Rejected - Total Fixation Duration - DICE Approved - Total Fixation Duration	0.018	8		0.002	No
SHAP Approved - Total Fixation Duration - Anchors Approved - Total Fixation Duration	0.040	9		0.003	No
LIME Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.053	10		0.003	No
SHAP Approved - Total Fixation Duration - LIME Approved - Total Fixation Duration	0.070	11		0.003	No
SHAP Rejected - Total Fixation Duration - LIME Approved - Total Fixation Duration	0.077	12		0.003	No
SHAP Approved - Total Fixation Duration - DICE Approved - Total Fixation Duration	0.084	13		0.003	No
DICE Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	0.091	14		0.003	No
DICE Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.136	15		0.004	No

LIME Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.184	16		0.004	No
SHAP Rejected - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.277	17		0.004	No
LIME Rejected - Total Fixation Duration - DICE Approved - Total Fixation Duration	0.334	18		0.005	No
SHAP Approved - Total Fixation Duration - LIME Rejected - Total Fixation Duration	0.334	19		0.005	No
LIME Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	0.376	20		0.006	No
SHAP Rejected - Total Fixation Duration - DICE Rejected - Total Fixation Duration	0.469	21		0.006	No
LIME Approved - Total Fixation Duration - LIME Rejected - Total Fixation Duration	0.520	22		0.007	No
SHAP Rejected - Total Fixation Duration - LIME Rejected - Total Fixation Duration	0.520	23		0.008	No
SHAP Approved - Total Fixation Duration - Anchors Rejected - Total Fixation Duration	0.546	24		0.010	No
LIME Approved - Total Fixation Duration - DICE Approved - Total Fixation Duration	0.717	25		0.013	No
SHAP Approved - Total Fixation Duration - DICE Rejected - Total Fixation Duration	0.778	26		0.017	No
LIME Rejected - Total Fixation Duration - DICE Rejected - Total Fixation Duration	0.904	27		0.025	No
SHAP Approved - Total Fixation Duration - SHAP Rejected - Total Fixation Duration	0.968	28		0.050	No

Table 37: Two-way repeated measures ANOVA analysis for fixation duration on complete visualization in the eye-tracking experiment.

Tests of Within-Subjects Effects for Fixation Duration in Complete Visualization							
Measure: Fixation Duration in Complete Visualization							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	**Sphericity Assumed**	**6752001027.316**	**3**	**2250667009.105**	**5.277**	**0.003**	**0.227**
Error (Explanation Representation)	Sphericity Assumed	23029406979.934	54	426470499.628			
Loan Decision	Sphericity Assumed	11834064.105	1	11834064.105	0.051	0.823	0.003
Error (Loan Decision)	Sphericity Assumed	4143997381.145	18	230222076.730			
Explanation Representation * Loan Decision	**Sphericity Assumed**	**1600538948.895**	**3**	**533512982.965**	**3.618**	**0.019**	**0.168**
Error (Explanation Representation * Loan Decision)	Sphericity Assumed	7963910197.855	54	147479818.479			

Table 38: Two-way measures ANOVA analysis for fixation duration on explanation representation as a percentage of complete visualization in the eye-tracking experiment.

Tests of Within-Subjects Effects for Fixation Duration on Explanation Representation as Percentage of Complete Visualization							
Measure: Fixation Duration on Explanation Representation as Percentage of Complete Visualization							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	**Sphericity Assumed**	**3169.250**	**3**	**1056.417**	**2.989**	**0.039**	**0.142**
Error (Explanation Representation)	Sphericity Assumed	19083.370	54	353.396			
Loan Decision	Sphericity Assumed	112.583	1	112.583	0.401	0.535	0.022
Error (Loan Decision)	Sphericity Assumed	5059.105	18	281.061			
Explanation Representation * Loan Decision	Sphericity Assumed	178.266	3	59.422	0.324	0.808	0.018
Error (Explanation Representation *Loan Decision)	Sphericity Assumed	9913.711	54	183.587			

Table 39: Post-hoc pairwise comparison with Bonferroni correction for fixation duration on explanation representation as a percentage of complete visualization in the eye-tracking experiment.

Pairwise Comparisons						
Measure: Fixation Duration on Explanation Representation as Percentage of Complete Visualization						
(I) Explanation Representation	**(J) Explanation Representation**	**Mean Difference (I-J)**	**Std. Error**	**Sig.[a]**	**95% Confidence Interval for Difference[a]**	
					Lower Bound	**Upper Bound**
Anchors	DICE	-0.983	4.017	1.000	-12.885	10.918
	LIME	-6.335	5.077	1.000	-21.378	8.708
	SHAP	-11.395	4.699	0.156	-25.316	2.525
DICE	Anchors	0.983	4.017	1.000	-10.918	12.885
	LIME	-5.352	3.504	0.864	-15.733	5.029
	SHAP	**-10.412**	**3.905**	**0.094**	**-21.983**	**1.159**
LIME	Anchors	6.335	5.077	1.000	-8.708	21.378
	DICE	5.352	3.504	0.864	-5.029	15.733
	SHAP	-5.060	4.481	1.000	-18.335	8.215
SHAP	Anchors	11.395	4.699	0.156	-2.525	25.316
	DICE	**10.412**	**3.905**	**0.094**	**-1.159**	**21.983**
	LIME	5.060	4.481	1.000	-8.215	18.335
Based on estimated marginal means a. Adjustment for multiple comparisons: Bonferroni.						

Table 40: Two-way repeated measures ANOVA analysis number of fixations on explanation representation as a percentage of complete visualization in the eye-tracking experiment.

Tests of Within-Subjects Effects for Number of Fixations on Explanation Representation as Percentage of Complete Visualization							
Measure: Fixation Duration on Explanation Representation as Percentage of Complete Visualization							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	**Sphericity Assumed**	**4496.950**	**3**	**1498.983**	**6.097**	**0.001**	**0.253**
Error(Explanation Representation)	Sphericity Assumed	13275.561	54	245.844			
Loan Decision	Sphericity Assumed	210.435	1	210.435	1.360	0.259	0.070
Error(Loan Decision)	Sphericity Assumed	2785.370	18	154.743			
Explanation Representation * Loan Decision	Sphericity Assumed	123.176	3	41.059	0.316	0.814	0.017
Error(Explanation Representation * Loan Decision)	Sphericity Assumed	7009.436	54	129.804			

Table 41: Post-hoc pairwise comparison with Bonferroni correction for number of fixations on explanation representation as a percentage of complete visualization in the eye-tracking experiment.

Pairwise Comparisons						
Measure: Number of Fixations on Explanation Representation as a Percentage of Complete Visualization						
(I) Explanation Representation	**(J) Explanation Representation**	**Mean Difference (I-J)**	**Std. Error**	**Sig.[b]**	**95% Confidence Interval for Difference[b]**	
					Lower Bound	**Upper Bound**
Anchors	DICE	4.641	3.302	1.000	-5.141	14.423
	LIME	-6.246	4.129	0.886	-18.480	5.988
	SHAP	-9.378	3.811	0.145	-20.668	1.913
DICE	Anchors	-4.641	3.302	1.000	-14.423	5.141
	LIME	**-10.888***	**3.302**	**0.024**	**-20.670**	**-1.105**
	SHAP	**-14.019***	**3.299**	**0.003**	**-23.794**	**-4.244**
LIME	Anchors	6.246	4.129	0.886	-5.988	18.480
	DICE	**10.888***	**3.302**	**0.024**	**1.105**	**20.670**
	SHAP	-3.131	3.657	1.000	-13.966	7.703
SHAP	Anchors	9.378	3.811	0.145	-1.913	20.668
	DICE	**14.019***	**3.299**	**0.003**	**4.244**	**23.794**
	LIME	3.131	3.657	1.000	-7.703	13.966

Based on estimated marginal means
* The mean difference is significant at the .05 level.
b. Adjustment for multiple comparisons: Bonferroni.

Table 42: Two-way repeated measures ANOVA analysis fixation duration on counterfactuals (%) in the eye-tracking experiment.

Tests of Within-Subjects Effects for Fixation Duration on Counterfactuals (%)							
Measure: Fixation Duration on Counterfactuals (%)							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Counterfactual	**Sphericity Assumed**	**2488.570**	**2**	**1244.285**	**7.488**	**0.002**	**0.294**
Error (Counterfactual)	Sphericity Assumed	5981.901	36	166.164			
Loan Decision	Sphericity Assumed	0.000	1	0.000	.	.	1.000
Error (Loan Decision)	Sphericity Assumed	0.000	18	0.000			
Loan Decision * Counterfactual	Sphericity Assumed	261.783	2	130.892	0.490	0.616	0.027
Error (Loan Decision * Counterfactual)	Sphericity Assumed	9608.190	36	266.894			

Table 43: Post-hoc pairwise comparison with Bonferroni correction for fixation duration on counterfactuals (%) in the eye-tracking experiment.

Pairwise Comparisons						
Measure: Fixation Duration on Counterfactuals (%)						
(I) Counterfactual	**(J) Counterfactual**	**Mean Difference (I-J)**	**Std. Error**	**Sig.[b]**	**95% Confidence Interval for Difference[b]**	
					Lower Bound	**Upper Bound**
First	Second	-0.962	2.915	1.000	-8.656	6.732
	Third	**9.395***	**2.713**	**0.008**	**2.234**	**16.556**
Second	First	0.962	2.915	1.000	-6.732	8.656
	Third	**10.357***	**3.221**	**0.014**	**1.856**	**18.858**
Third	**First**	**-9.395***	**2.713**	**0.008**	**-16.556**	**-2.234**
	Second	**-10.357***	**3.221**	**0.014**	**-18.858**	**-1.856**
Based on estimated marginal means * The mean difference is significant at the .05 level. b. Adjustment for multiple comparisons: Bonferroni.						

Table 44: Two-way repeated measures ANOVA analysis fixation duration on top attributes for LIME and SHAP (%) in the eye-tracking experiment.

Tests of Within-Subjects Effects for Fixation Duration on Top Attributes for LIME and SHAP (%)							
Measure: Fixation Duration on Top Attributes for LIME and SHAP (%)							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	**Sphericity Assumed**	**1190.885**	**1**	**1190.885**	**4.581**	**0.046**	**0.203**
Error (Explanation Representation)	Sphericity Assumed	4678.823	18	259.935			
Loan Decision	Sphericity Assumed	5.493	1	5.493	0.021	0.887	0.001
Error (Loan Decision)	Sphericity Assumed	4743.984	18	263.555			
Explanation Representation * Loan Decision	**Sphericity Assumed**	**716.768**	**1**	**716.768**	**2.666**	**0.120**	**0.129**
Error (Explanation Representation *Loan Decision)	Sphericity Assumed	4840.270	18	268.904			

Table 45: Post-hoc pairwise comparison with Bonferroni correction for fixation duration on top attributes for LIME and SHAP (%) in the eye-tracking experiment.

Pairwise Comparisons						
Measure: Fixation Duration on Top Attributes for LIME and SHAP (%)						
(I) Explanation Representation	**(J) Explanation Representation**	**Mean Difference (I-J)**	**Std. Error**	**Sig.[b]**	**95% Confidence Interval for Difference[b]**	
					Lower Bound	**Upper Bound**
LIME	**SHAP**	**7.917***	**3.699**	**0.046**	**0.146**	**15.688**
SHAP	**LIME**	**-7.917***	**3.699**	**0.046**	**-15.688**	**-0.146**

Based on estimated marginal means
* The mean difference is significant at the .05 level.
b. Adjustment for multiple comparisons: Bonferroni.

Table 46: Two-way repeated measures ANOVA analysis fixation duration on positive attributes for LIME and SHAP (%) in the eye-tracking experiment.

Tests of Within-Subjects Effects for Fixation Duration on Positive Attributes for LIME and SHAP (%)							
Measure: Fixation Duration on Positive Attributes for LIME and SHAP (%)							
Source		**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Explanation Representation	**Sphericity Assumed**	**9929.903**	**1**	**9929.903**	**34.506**	**<0.001**	**0.657**
Error (Explanation Representation)	Sphericity Assumed	5179.882	18	287.771			
Loan Decision	Sphericity Assumed	572.462	1	572.462	3.183	0.091	0.150
Error (Loan Decision)	Sphericity Assumed	3237.641	18	179.869			
Explanation Representation * Loan Decision	**Sphericity Assumed**	**5735.807**	**1**	**5735.807**	**46.216**	**<0.001**	**0.720**
Error (Explanation Representation * Loan Decision)	Sphericity Assumed	2233.958	18	124.109			

Table 47: Post-hoc pairwise comparison with Bonferroni correction for fixation duration on positive attributes for LIME and SHAP (%) in the eye-tracking experiment.

Pairwise Comparisons						
Measure: Fixation Duration on Positive Attributes for LIME and SHAP (%)						
(I) Explanation Representation	**(J) Explanation Representation**	**Mean Difference (I-J)**	**Std. Error**	**Sig.[b]**	**95% Confidence Interval for Difference[b]**	
					Lower Bound	**Upper Bound**
LIME	**SHAP**	**-22.861***	**3.892**	**<0.001**	**-31.037**	**-14.685**
SHAP	**LIME**	**22.861***	**3.892**	**<0.001**	**14.685**	**31.037**

Based on estimated marginal means
* The mean difference is significant at the .05 level.
b. Adjustment for multiple comparisons: Bonferroni.

Table 48: Post-hoc pairwise comparison with Bonferroni correction for fixation duration on positive attributes for rejected and approved loans (%) in the eye-tracking experiment.

Pairwise Comparisons						
Measure: Fixation Duration on Positive Attributes for Approved and Rejected Loans (%)						
(I) Loan Decision	**(J) Loan Decision**	**Mean Difference (I-J)**	**Std. Error**	**Sig.[a]**	**95% Confidence Interval for Difference[a]**	
					Lower Bound	**Upper Bound**
Rejected	**Approved**	**5.489**	**3.077**	**0.091**	**-0.975**	**11.953**
Approved	**Rejected**	**-5.489**	**3.077**	**0.091**	**-11.953**	**0.975**
Based on estimated marginal means a Adjustment for multiple comparisons: Bonferroni.						

Appendix A8: Guide for Semi-structured Interviews

General Experience During Experiment

- How did you feel participating in this experiment?
- Were the experiment instructions clear and easy to follow?

General Perception of the System

- How did you feel in general about your interaction with the AI system, independently of the type of explanation that the system provided?
- Do you think the system was reliable?
- How would you feel if bank employees used this system when deciding loan applications?
- Do you think the system provided fair recommendations?

General Perception on Explanations

- Do you think that providing explanations helps to understand why the system made a certain decision recommendation for a loan application?
- How would you feel if the system didn't provide any explanations for the decisions it makes?

Preference Between Explanation Types

- From the four different types of explanations that the system provided to explain why it made a certain decision recommendation, how did you rank the four types of explanations according to your preference?

Preference Between Explanation Types

- Why was this explanation type __ your 1st/2nd/3rd/4th preferred type of explanation?
- What did you like about this type of explanation?
- What didn't you like about this type of explanation?
- Do you think that this type of explanation was useful to understand the decisions made by the system?
- How would you grade the usefulness of this explanation on a scale from 0 to 10?

Alternative Explanations

- Could you imagine another way in which the system could provide you with an explanation of why a given decision was made?
- What information do you think could be useful and wasn't provided by any of the explanations?

Appendix B: Study II

Appendix B1: Deviations from the Preregistration

Table 49: Deviations from the preregistration.

Deviation from the Preregistration	**Justification**
Change in the research model for Experiment 1	The self-developed scale of accepting bias as contingent was preregistered as a moderator between bias detection and trusting intentions. This variable was replaced with PCV and changed the role to a mediator. • Reasons in theory: The proposed relationship between bias detection and trusting intentions was already based on the PCV theory in the preregistration. It was assumed that accepting bias as contingent would reduce the PCV and therefore reduce the negative effect of bias detection on trusting intentions. The measure of PCV was formerly intended to be an additional check for the proposed effects. For this experiment, the direct justification was followed and included PCV as the primary variable instead of accepting bias as contingent. • Measurement reasons: Accepting bias as contingent caused multiple measurement concerns. First, in the exploratory factor analysis, the items had a strong cross-loading with items of trusting intentions. This reduced the model's quality and made it impossible to use it as a moderator. Furthermore, the scale was strongly correlated with PCV and did not offer additional explanatory insights. Because of this conceptual similarity and for theoretical reasons, the established scale for PCV was used instead
Additional mediation pathways	• The direct effect pathway from the experimental manipulation on trusting intentions was included in Study1 +2. No hypothesis was tested this way. • For Experiment 2, an additional direct pathway from the priming manipulation to stigma consciousness was included, as this was not clearly elaborated in the preregistration.
Plausibility of Explanations (Experiment 1)	• An assessment of the plausibility of explanations as a second dependent variable was preregistered. However, a factor analysis revealed that combining all plausibility items into one factor was not possible. Therefore, this analysis was removed, and all ratings were considered separately in a repeated measure ANOVA.
Wording	• In line with prior literature and to increase clarity, the wording of stigma consciousness (preregistered as stereotype consciousness) and perceived bias (preregistered as bias detection) was adjusted to correspond with the label of the scale in the referenced literature. • The wording of the hypotheses was adjusted to increase clarity.

Appendix B2: Measures

Table 50: Measures.

Measure	**Items**
Variables for the overall evaluation and model testing	
Trusting intentions (adapted from McKnight et al., 2002)	We plan to introduce the developed algorithm into a German bank. Would you recommend bank employees to use this version of the algorithm?
	I would feel comfortable that bank employees act on the information given by this algorithm.
	I would not hesitate to recommend bank employees to use this algorithm.
	I would feel confident if bank employees acted on the decision recommendation given by this algorithm.
Perceived bias (adapted from Wallace et al. 2020)	How much would you see the algorithm as having a biased perspective?
	How much would you see the algorithms' decision recommendation as a product of bias in the data?
	To what extent do you feel that the algorithms' decision recommendation is a product of bias?
	The algorithm failed to meet its obligations to me.

Psychological Contract Violation (adapted from Pavlou and Gefen 2005; Wang and Wang 2019)	The algorithm did a good job of meeting its obligations to me. (*Reversed*)
	The algorithm fulfilled the most important obligations to me. (*Reversed*)
	The algorithm failed to meet its obligations to me.
Stigma Consciousness to female gender (adapted from Pethig and Kroenung 2020; Pinel 1999), short-scale	Please answer the following questions about your general attitudes about women in society
	Stereotypes about gender have not affected me personally. (Reversed)
	I never worry that my behavior will be viewed as stereotypically female. (Reversed)
	Most men do not judge women on the basis of their gender. (Reversed)
Transparency (adapted from Wang and Benbasat 2016), only used in Experiment 1	This algorithm made its reasoning process clear to me.
	I could easily understand this algorithm's reasoning process.
	It was easy for me to understand the inner workings of this algorithm.
	This algorithm's logic in providing the decision recommendation was clear to me.
	This algorithm hid important information that might reflect badly on the decision recommendation. (*Reversed, excluded from analysis)*
Controls	
Familiarity with the task (adapted from Doshi-Velez and Kim 2017; Gefen 2000; Kim et al. 2019) 7-point Likert scale	I am familiar with loan applications in general.
	I am familiar with the process to apply for a loan.
	I am familiar with evaluation criteria of loan decisions.
Algorithm knowledge (Cheng et al., 2019)	How much knowledge of computer algorithms do you have? No knowledge; A little knowledge – I know basic algorithm concepts; Some knowledge – I have used algorithms before; A lot of knowledge – I apply algorithms frequently to my work or I create algorithms
Machine learning knowledge (adapted from Cheng et al. 2019)	How much knowledge of machine learning do you have? No knowledge; A little knowledge – I know basic concepts in machine learning; Some knowledge – I have used machine learning before; A lot of knowledge – I apply machine learning frequently to my work or I create machine learning applications
Disposition to trust (Gefen, 2000)	I generally have faith in humanity.
	I feel that people are generally reliable.
	I generally trust other people unless they give me reasons not to.
Exploratory analysis (H3b)	
Plausibility of the decision recommendation	Please evaluate the algorithm's decision recommendation to approve or reject the loan:
	I think that the algorithm's decision recommendation is plausible.
	Experiment 2 (*additionally*): I agree with the algorithm's decision recommendation.
Open question	Please explain your selected answer for the plausibility (Experiment 2: and agreement) of the algorithm's decision recommendation
Manipulation check in Experiment 1	
Perceived neutrality (adapted from (McKinney et al., 2002; Wang & Wang, 2019)	On the basis of your interaction with the algorithm, do you expect the algorithm ... … to favor certain attributes that you may not agree with? (Reversed) … to be free of bias when providing loan decisions? … to provide misleading loan decisions? (Reversed)
Comprehension check	
Comprehension Check 1	Which of the following statements about the algorithm is true? The algorithm evaluates loan applications – The algorithm organizes loan applications – The algorithm generates loan applications – The algorithm selects loan applications
Comprehension Check 2	Which of the following statements about the algorithm is true? The explanation presents attributes according to their influence on the decision recommendation. – The explanation shows random attributes from each applicant. - All explanations are the same for each loan application.
Comprehension Check 3	Please check the above explanation – Which attribute had the strongest influence on the algorithm's decision recommendation to "approve" the loan application (Slightly different examples provided in Experiment 1 and Experiment 2) The account balance of the loan applicant is above 200 Euro. - The loan's applicant loan history. - The loan applicant has paid back all previous loans at the bank.

Appendix B3: Demographic Background and Control Variables

Table 51: Demographic background and control variables**.**

Variable /Item		**No. of participants (%) in Experiment 1**	**Mean of Experiment 1 (SD)**	**No. of participants (%) in Experiment 2**	**Mean of Experiment 2 (SD)**
Age		-	33.78 (12.58)		36.19 (13.52)
Gender	Male	186 (54.9)	-	103 (45.5)	-
	Female	144 (42.5)	-	114 (50.4)	-
	Non-binary (coded as female)	8 (2.4)	-	8 (3.5)	-
	Other (missing data)	1 (0.3)	-	1 (0.4)	
Education	None	-	-	-	-
	Elementary	-	-	-	-
	Middle school / Junior high school	2 (0.6)	-	-	-
	Secondary	68 (20.1)	-	39 (17.3)	-
	Post-secondary, non-tertiary	52 (15.3)	-	35 (15.5)	-
	Tertiary (higher professional education, university education)	184 (54.3)	-	112 (49.6)	-
	Post-tertiary	31 (9.1)		40 (17.7)	-
	Other	2 (0.6)	-	-	-
Control variables	Familiarity with loan applications		4.38 (1.72)	-	4.72 (1.63)
	Algorithm knowledge		2.31 (0.71)	-	2.33 (0.75)
	Machine learning knowledge		2.09 (0.68)	-	2.09 (0.71)
	Disposition to trust		4.70 (1.43)	-	4.40 (1.40)

Appendix B4: Measurement Model (CFA) for Experiment 1

Table 52: Measurement model (CFA) for Experiment 1.

Construct	Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Perceived Bias	Bias1	**0.878**	-0.021	0.055	0.019	-0.02
	Bias2	**0.899**	0.036	0.007	-0.008	0.045
	Bias3	**0.985**	-0.045	-0.021	-0.005	-0.02
Stigma Consciousness	SC1	-0.017	**0.871**	0.032	-0.019	0.023
	SC2	-0.071	**0.817**	0.057	0.051	0.057
	SC3	0.111	**0.566**	-0.135	-0.046	-0.12
Perceived PCV	PCV1	0.029	0.019	**0.66**	0.013	-0.161
	PCV2	-0.009	0.007	**0.954**	-0.047	0.046
	PCV3	0.037	0.038	**0.834**	-0.016	-0.039
Perceived Transparency	Transp1	0.016	0.007	0.086	**0.934**	0.077
	Transp2	-0.008	-0.021	-0.059	**0.956**	-0.085
	Transp3	-0.01	0.011	0.036	**0.882**	0.083
	Transp4	0.005	0.012	-0.109	**0.905**	-0.057
Trusting Intentions	Trust1	-0.027	0.028	-0.102	-0.014	**0.869**
	Trust2	-0.03	-0.012	0	0.021	**0.884**
	Trust3	0	0.015	-0.048	0.026	**0.936**

Appendix B5: Results of ANCOVA analysis in Experiment 1

Table 53: Results of ANCOVA analysis in Experiment 1.

Trusting intentions	**df**	**F**	**Sig.**	**Partial Eta-Square**
Adjusted Model	**9**	**6.746**	**<0.001**	**0.156**
Intercept	1	5.967	0.015	0.018
Controls				
Age	1	3.611	0.058	0.011
Gender	**1**	**11.597**	**<0.001**	**0.034**
Highest degree	1	2.624	0.106	0.008
Task familiarity	1	9.406	0.002	0.028
Disposition to trust	1	5.187	0.023	0.016
Algorithm knowledge	1	4.137	0.043	0.012
Machine learning knowledge	1	1.867	0.173	0.006
Experimental treatment				
Treatment Group (Control vs. Neutral vs. Biased)	**2**	**9.992**	**<0.001**	**0.057**
R-Square = .156 (corrected R-square = .133)				

Appendix B6: Measurement Model (CFA) for Experiment 2

Table 54: Measurement model (CFA) for Experiment 2.

Construct	Item	Factor 1	Factor 2	Factor 3
Perceived Bias	Bias1	**0.945**	-0.008	-0.007
	Bias2	**0.925**	0.011	0.014
	Bias3	**0.97**	-0.013	-0.013
Perceived PCV	PCV1	0.104	**0.613**	-0.243
	PCV2	-0.043	**0.984**	-0.044
	PCV3	0.011	**0.877**	0
Trusting Intentions	Trust1	0.008	-0.072	**0.912**
	Trust2	-0.031	-0.023	**0.858**
	Trust3	0.004	-0.02	**0.955**

Appendix B7: Results of ANCOVA analysis in Experiment 2

Table 55: Results of ANCOVA analysis in Experiment 2

Trusting intentions	**df**	**F**	**Sig.**	**Partial Eta-Square**
Adjusted Model	**10**	**3.905**	**<0.001**	**0.154**
Intercept	1	1.047	0.307	0.005
Controls				
Age	1	0.113	0.737	0.001
Gender	1	6.352	0.012	0.029
Highest degree	1	0.633	0.427	0.003
Task familiarity	1	1.63	0.203	0.008
Disposition to trust	1	9.196	0.003	0.041
Algorithm knowledge	1	0.612	0.435	0.003
Machine learning knowledge	1	0.272	0.602	0.001
Experimental treatment				
AI System (Neutral vs. Biased)	**1**	**13.094**	**<0.001**	**0.058**
Priming (No vs. Priming)	1	2.539	0.113	0.012
AI System * Priming	1	2.243	0.136	0.01
R-Square = .154 (corrected R-Square = .115)				

Appendix C: Study III

Appendix C1: Guide for Semi-structured Interviews

General Questions about the Study

- How did you find the study?
- How did you feel participating in this study?
- Before and during the study, what was your incentive to participate?
- What were your goals for the study? What did you imagine differently?

General Perception of the System

- How did you feel in general about your interaction with the AI system?
- Do you think the system was reliable?
- Do you think the system provided fair recommendations?
- How would you feel if bank employees used this system when deciding on loan applications you apply for in a bank?
- Could you please briefly explain how the system works?

Perception of System Functionality

- Do you think that the system's functionality helps you understand why the system made a certain decision recommendation for a loan application?
- How would you feel if the system didn't provide functionality that allows you to understand how it makes decision recommendations?

Perception of What-if analysis

- What do you think about the system's functionality that allows you to modify the attributes of the bank loan application to observe how the system's decision recommendation would change?
- Do you think that this functionality is useful? Does this function help you to understand how the system makes decision recommendations?
- Was this functionality easy to understand? Was it easy to use?
- Would you change anything regarding this functionality to make the system better?
- Could you imagine another way the system could provide you with an alternative functionality to help you more?

Perception of Explanations

- What do you think about the explanations that the system provided for how it made each decision recommendation?
- Were the explanations clear and easy to understand?
- Do you think that providing such explanations helps you understand why the system made a certain decision recommendation for a loan application?
- How would you feel if the system didn't provide explanations for its decision recommendations?
- Would you change anything regarding the explanations to make the system better?
- Could you imagine another way the system could provide you with an alternative explanation to help you more?

Extra Information on System

- Is there anything else you would like to share regarding the system or its functionality?

List of Publications

Journal Publications (Working Paper)

Jussupow, E., **Meza Martínez, M. A.**, Maedche, A., & Heinzl, A. (2023). *Why Individuals Trust and Not Trust Biased Explainable AI Systems: A Psychological Contract Violation and Social Identity Perspective*. Working paper, to be submitted.

Journal Publications (Published)

Meza Martínez, M. A., Nadj, M., Langner, M., Toreini, P., & Maedche, A. (2023). *Does This Explanation Help? Designing Local Model-Agnostic Explanation Representations and an Experimental Evaluation Using Eye-Tracking Technology*. ACM Transactions on Interactive Intelligent Systems (TiiS), Special Issue on Human-centered Explainable AI. Just Accepted. https://doi.org/10.1145/3607145

Conference Proceedings (Published)

Jussupow, E., **Meza Martínez, M. A.**, Maedche, A., & Heinzl, A. (2021). *Is This System Biased? – How Users React to Gender Bias in an Explainable AI System*. In Proceedings of the 42nd International Conference on Information Systems (ICIS 2021), Austin: AISel. https://aisel.aisnet.org/icis2021/hci_robot/hci_robot/11

Meza Martínez, M. A., Nadj, M., & Maedche, A. (2019). *Towards an Integrative Theoretical Framework of Interactive Machine Learning Systems*. Proceedings of the 27th European Conference on Information Systems (ECIS).

Exler, A., Kramer, S., **Meza Martínez, M. A**., Navolskyi, C., Vogt, M., & Beigl, M. (2017). *Suitability of Event-Based Prompts in Experience Sampling Studies Focusing on Location Changes*. In Proceedings of International Conference on Wireless Mobile Communication and Healthcare – (MobiHealth 2017), 163–168.

Conference Proceedings (Accepted)

Meza Martínez, M. A., & Maedche, A. (2023). *Designing Interactive Explainable AI Systems for Lay Users*. Manuscript Accepted in the International Conference on Information Systems (ICIS 2023).

Author Workshop

Jussupow, E., **Meza Martínez, M. A.**, Maedche, A., & Heinzl, A. (2021). *Is This System Biased? – How Users React to Gender Bias in an Explainable AI System*. In MISQ Author Workshop.

Theses

Meza Martínez, M. A. (2018). *Developing a Forecasting Tool for Industrial Energy Time Series*. Karlsruhe Institute of Technology.

Eidesstattliche Versicherung

gemäß § 13 Abs. 2 Ziff. 3 der Promotionsordnung des Karlsruher Instituts für Technologie für die Fakultät für Wirtschaftswissenschaften

1. Bei der eingereichten Dissertation zu dem Thema *Visual Representation of Explainable Artificial Intelligence Methods: Design and Empirical Studies* handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, den 21.11.2023

_____

Miguel Angel Meza Martínez