Human-centric Explanation Facilities: Explainable AI for the Pragmatic Understanding of Non-expert End Users

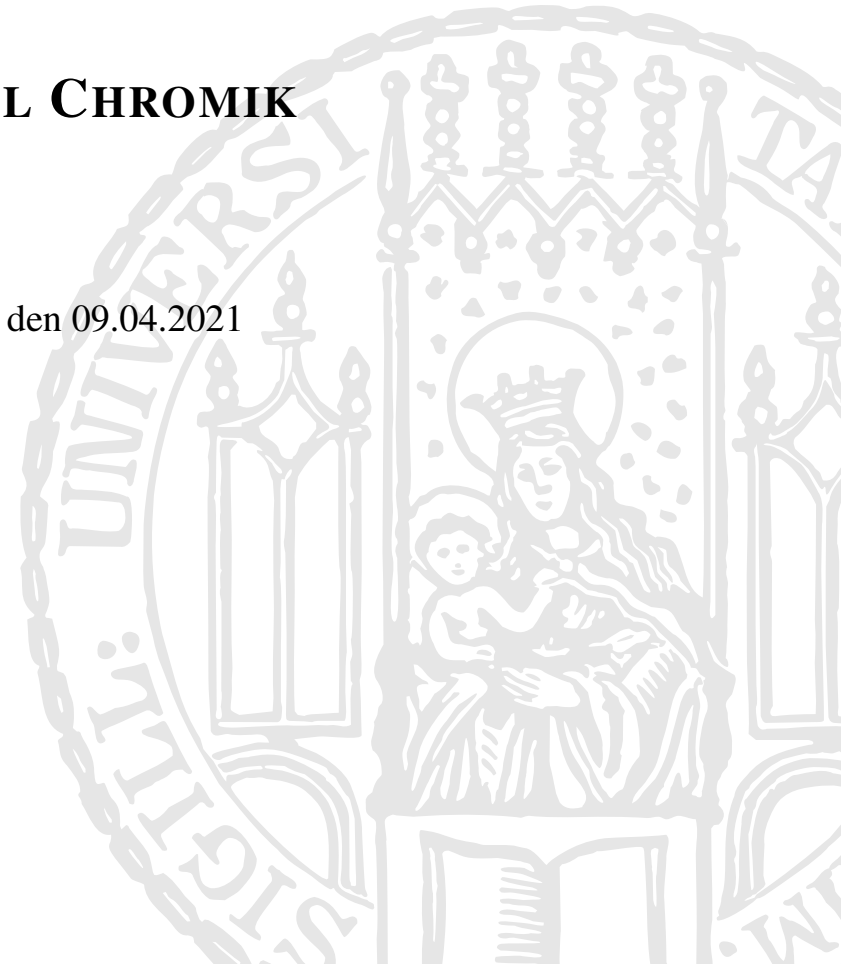
Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

eingereicht von

MICHAEL CHROMIK

München, den 09.04.2021



Erstgutachter:	Prof. Dr. Andreas Butz
Zweitgutachter:	Prof. Dr.-Ing. Klaus Diepold
Drittgutachter:	Prof. Dr. Dieter Kranzlmüller

Tag der mündlichen Prüfung: 28.06.2021

ABSTRACT

AI-infused systems support or automate decision-making in many sensitive contexts of society, such as medicine, education, criminal justice, loan approval, or recruiting. To accomplish this, these systems often leverage *machine learning (ML)* methods. The risk of unintended consequences caused by the black-box problem of these systems drove social, ethical, and legal calls for more interpretable and explainable systems for stakeholders beyond the systems' developers. This field is often referred to as *Explainable Artificial Intelligence (XAI)*. XAI researchers face the accusation of putting their emphasis on generating explanatory models for like-minded ML experts instead of tailoring them to the actual users, who often lack the technical background about ML.

This human-centric view on non-expert users of XAI is the focus of this thesis. By non-experts, we refer to users of XAI systems who have not been trained or educated in ML concepts, but who use the system's predictions to support or perform their tasks. The work presented in this thesis (i) examines human factors that may be encountered by end users when interpreting explanations from ML-based XAI systems, and (ii) explores the interaction space of XAI explanation facilities to foster a pragmatic understanding of the underlying ML model.

To this end, this thesis makes three contributions: (i) it empirically explores unknown human factors and cognitive biases that influence the end user understanding gained through XAI explanations, (ii) it conceptually structures inconsistencies within the involved research communities on how explanations are evaluated with human subjects in empirical XAI studies and how user interaction with explanation interfaces in XAI can be described, and (iii) it presents case studies of constructive explanation facility artifacts that fulfill the requirements of *naturalness* (use natural language explanations), *responsiveness* (allow follow-up interactions), and *sensitivity* (elicit end user beliefs to calibrate their expectations). In summary, this thesis raises XAI designers' awareness of the human aspects of interpretability in XAI.

ZUSAMMENFASSUNG

Mit *Künstlicher Intelligenz (KI)* angereicherte Systeme unterstützen oder automatisieren Entscheidungen in vielen sensiblen Bereichen der Gesellschaft, beispielsweise in der Medizin, im Bildungswesen, im Strafrecht, der Kreditvergabe oder der Personalrekrutierung. Dafür nutzen diese Systeme oftmals Methoden des *Maschinellem Lernens (ML)*. Das Risiko unbeabsichtigter Konsequenzen, das vom Blackbox-Problem des Maschinellen Lernens ausgehen kann, führte zu sozialen, ethischen und rechtlichen Forderungen nach besser interpretierbaren und erklärbaren Systemen für Nutzergruppen, die über die Entwickler dieser Systeme hinausgehen. Dieses Feld wird oft als *Erklärbare Künstliche Intelligenz (engl. Explainable Artificial Intelligence, XAI)* bezeichnet. XAI-Forscher sehen sich dem Vorwurf ausgesetzt, dass sie viel Wert darauf legen, Erklärungsmodelle für gleichgesinnte ML-Experten zu generieren, anstatt sie auf die eigentlichen Endnutzer, die oftmals keine technischen Experten sind, auszurichten.

Diese menschenzentrierte Sichtweise auf nicht-technische Endnutzer von XAI ist der Fokus dieser Arbeit. Mit nicht-technischen Endnutzern bezeichnen wir Nutzer von XAI-Systemen, die wenig mit den Konzepten des maschinellen Lernens vertraut sind, aber dessen Vorhersagen zur Unterstützung oder Erfüllung ihrer Aufgaben nutzen. Die in dieser Thesis vorgestellten Arbeiten (i) untersuchen menschliche Faktoren, die bei Endnutzern auftreten können, wenn sie Erklärungen von ML-basierten XAI-Systemen interpretieren und (ii) erforschen den Interaktionsraum von XAI-Nutzerschnittstellen *(engl. XAI Explanation Facilities)*, um das pragmatische Verständnis des zugrunde liegenden ML-Modells zu fördern.

Hierbei leistet diese Arbeit drei Beiträge: (i) sie erforscht empirisch unbekannte menschliche Faktoren und kognitive Verzerrungen, die das durch XAI-Erklärungen gewonnene Verständnis von nicht-technischen Endnutzern beeinflussen, (ii) sie strukturiert konzeptionell die Inkonsistenzen innerhalb der Forschungsgemeinden darüber, wie Erklärungen mit menschlichen Probanden in empirischen XAI-Studien evaluiert werden und wie die Interaktion mit XAI-Nutzerschnittstellen beschrieben werden kann, sowie (iii) sie präsentiert konstruktive Fallstudien von XAI-Nutzerschnittstellen, die den Anforderungen der *Natürlichkeit* (verwendet natürlichsprachliche Erklärungen), *Ansprechbarkeit* (ermöglicht Folgeinteraktionen) und *Einfühlsamkeit* (ermittelt die Überzeugungen der Nutzer, um deren Erwartungen zu kalibrieren). Zusammenfassend schärft diese Arbeit das Bewusstsein von XAI Designern für die menschlichen Aspekte der Interpretierbarkeit in XAI.

ACKNOWLEDGEMENT

This dissertation would not have been possible without the encouragement and support of so many people I was privileged to learn from. First and foremost, I would like to thank my advisor Prof. Dr. Andreas Butz for his constant guidance, patience, and personal support along the winding path of my research. For allowing me to explore diverse, often naïve, flashes of inspiration, and for guiding me in tailoring them into scientific outcomes. Thanks for reminding me that DNF is not an option in any context. Many thanks also go to the reviewers Prof. Dr.-Ing. Klaus Diepold and Prof. Dr. Dieter Kranzlmüller, who both supported me during my initial exploration of research directions and whose support I am glad to again have at the end of m doctoral research.

I also thank all colleagues and collaborators who enriched my journey. Especially, to my colleagues at the *LMU Media Informatics and HCI Groups* for welcoming me so openly at my first IDC in Venice. Thank you Malin Eiband for sparking my interest in XAI, our joint projects, and joint status meetings from the wardrobe. Thank you Sarah Völkel, Malin Eiband, Kai Holländer, and Daniel Buscheck for introducing me to scientific work in the XAI context (*#explainabiliTea*) and navigating my baby steps at my first scientific conference. Thanks to Martin Schüssler for the pre-Christmas and time-zone-bridging night shifts. Many thanks also to my companions at the *Center for Digital Technology and Management (CDTM)* for all the creative sprints, constructive discussions, pragmatic doings, and the *mostly awesome* journey we have been through. Thank you to Gesa Biermann, Patrick Bilic, Laura Bechthold, Aaron Defort, Theresa Doppstadt, Elizaveta Felsche, Michael Fröhlich, Philipp Hofsommer, Philipp Hulm, Till Kröger, Florian Korte, Florian Lachner, Julia Mecheels, Kilian Moser, Philipp Nägelein, Stefan Nothelfer, Tom Schelo, Andrea Socher, Stefanie Weniger, and especially Uta Weber. Thank you to the countless amazing Centerlings with whom I had the honour of sharing this journey. Because of you, this time made me grow professionally, practically, and as a human being. Thank you for constantly challenging my status quo while pushing the boundaries of what I deemed possible.

I am also very grateful and in deep respect to my parents who, with their courage, enabled my sisters and me a life with freedom of choice. Without that, none of this would have been possible for me. Thank you to my amazing friends Johannes, Lukas, Steffen, Johannes, and Johannes who inspired and motivated me to start the journey of academic education. Last but not least, a very special thank you to Judith, who has always stood by my side along this journey, on good days and not so good ones. I'm thrilled about the many chapters to come.

I thank you all. Ich Danke Euch. Dziękuję wam.

TABLE OF CONTENTS

List of Figures

List of Tables

1 Introduction

1.1 Motivation

"Our human, social, and civic dilemmas are becoming technical.
And our technical dilemmas are becoming human, social, and civic."

B. Christian, The Alignment Problem, 2020

Unintended Consequences of AI in Sensitive Areas of Society Modern intelligent systems often leverage methods of artificial intelligence (AI) and directly expose them to end users. These so-called *AI-infused systems* [3] navigate us to our next destination and protect us from unwanted emails. They recommend us movies, music, or products, and serve tailored advertisements. Often, they build on non-linear *machine learning* (ML) methods that enable accurate predictions. AI-infused systems based on ML outperformed human performance in complex tasks such as speech recognition, language translation, and games [80, 108]. They find patterns in large volumes of data in reasonably little time and, thus, hold the promise to augment human decision-making.

This promise led to a controversially discussed proliferation of AI-infused systems into sensitive areas of our societies, such as credit scoring [21], algorithmic trading [44], education [61], recruiting [113], predictive policing [45], and criminal justice [18]. There were multiple cases of unintended consequences [98] that revealed a *"mismatch between human-intended and model-learned solutions"* of AI-infused systems [35]. For example, a decision support system for dermatologists has *"inadvertently learned that rulers are malignant"* instead of melanomas [87]. Further, search engines were shown to display *"fewer instances of an ad related to high paying jobs"* to women than to men [22], recruiting systems downgraded resumes that included words such as *"women's"* [28], and systems for the risk assessment or reoffenders were claimed to show disproportionately higher classification errors for people of color [33]. These unintended behaviors often result from the black-box problem of AI [102] and their tendency to learn shortcuts.

The Black-Box Problem of AI In traditional software development, engineers crafted a deductive model based on explicit rules and turned them into code. In such a context, it is possible to inspect which parts of the code are executed. Such systems are explainable by definition. A different approach is taken with machine learning (ML) systems. From the origins of the perceptron [99] to today's deep neural networks, ML-based AI systems are programmed *"to learn from experience"* [105]. Those systems do not establish rules in advance. Instead, ML-based systems are probabilistic, non-deterministic and often non-linear. For any input fed into a ML model, the output depends on the underlying training data and training parameters. From this training data, the algorithms autonomously infer implicit rules. This implicit inference comes at the expense of interpretability regarding the prediction process and effectively turns the model into a black-box (i.e., a situation in which it is possible to observe the inputs and outputs, but where the internal operations are not disclosed nor interpretable to humans) [102].

To assess the utility of a non-linear ML model, engineers mostly rely on performance metrics that compare the outcomes of the prediction process, such as the number or ratio of true and false predictions (e.g., accuracy, AUC, or F1) [86]. However, these performance metrics do not say anything about the ML model being *"right for the right reasons"* [100]. A model may learn shortcuts that do not generalize outside the training data or that are considered as unfair or discriminating [35]. For instance, Ribeiro et al. [97] trained an ML model with the goal of distinguishing dogs from wolves in images. While the metrics indicated promising performance, their explainability method showed that the model achieved its performance by distinguishing images with areas that did or did not contain snow in the background. ML models rely on correlation instead of causation. Bias and unfairness may creep into the prediction behavior if a discriminatory confounding factor is not included in the ML model. Thus, achieving high accuracy scores on a held-out test set may not always result in understandable and trustworthy systems that serve the underlying domain problem.

Democratizing AI Supervision through Explainability Explainability is considered as one way to prevent or monitor the emergence of undesired consequences of AI-infused systems. As these systems are introduced into more sensitive contexts of society, there is a growing acceptance that they must be capable of explaining their behavior in human-understandable terms to stakeholders beyond the developers. Regulatory, organizational, and societal stakeholders partake in the discussion and emphasize the importance of explainability for trustworthy systems. For example, the EU *General Data Protection Regulation (GDPR)* grants individuals affected by automated decisions the right for *"meaningful information about the logic involved"* [39]. The EU AI HLEG[1] considers explainability as a key element of trustworthy AI systems [32]. More concretely, the NIST[2] defined four fundamental principles of *explanation supply*, *meaningful explanation*, *explanation accuracy*, and *knowledge limits* for explainable AI systems [88]. Following these calls, companies aim at adopting explainability to manage their algorithmic risks. According to a survey by IBM, 68% of business executives believe that their customers will demand more AI explainability in the upcoming three years [54].

This surge for *explainable AI (XAI)* shows that AI-infused systems are no longer the sole matter of developers. Understanding AI-infused systems at least on a simplified level is a key to participating in these discussions. However, most works in XAI focus on the computational aspects of generating explanations for black-box ML models. Limited research concerns the human-centered design of XAI, for example, providing non-technical users with some intuition why certain predictions are made, the system's underlying assumptions or constraints, or means to calibrate their trust. If questions of *reflexive skepticism* from non-technical users are left unanswered by an AI-infused system, this will have a direct impact on trust, decision-making and eventually their adoption [123]. As the human use of computing is the subject of inquiry in the problem solving field of HCI [92], the HCI community *"should take a leading role [...] by providing explainable and comprehensible AI, and useful and usable AI"* [123].

[1] High-Level Expert Group on Artificial Intelligence

[2] National Institute of Standards and Technology of the U.S. Department of Commerce

1.2 Thesis Statement

AI-infused systems based on machine learning (ML) support or automate decision-making in many sensitive contexts of society. The risk of unintended consequences caused by the black-box problem of these systems drove social, ethical, and legal calls for more interpretable and explainable systems for stakeholders beyond the systems' developers. Research in this field is often referred to as *Explainable Artificial Intelligence (XAI)*. This dissertation centers around the human-centric challenges of making XAI systems comprehensible and usable for non-expert users from an HCI perspective. Making decisions of intelligent systems *comprehensible* to non-experts has been an active research field since the era of rule-based expert systems in the 1970s and 1980s. In my work, I build around the concept of *explanation facilities* that dates back to this time. It centers around the idea that an explanation interface that targets non-experts should include multiple modes of explanation and interaction. A *usable* explanation facility should fulfill the requirements of fidelity, naturalness, responsiveness, flexibility, sensitivity, extensibility, portability, and adaptivity [81]. Although the term has been used less often in the context of ML-based intelligent systems, the underlying human-centric principles remain valid and have the potential to address the call for *"usable, practical, and effective transparency that works for and benefits people"* [1]. Recently, XAI researchers have been accused of putting too much emphasis on generating explanatory models for like-minded ML experts (*"inmates running the asylum"* [77]) instead of tailoring them to the actual users who are often non-experts.

The work presented in this thesis (i) examines human factors that XAI designers must account for when exposing non-expert users to XAI explanations, (ii) conceptually analyzes the design space of involving them into the XAI explanation process through interaction or evaluation, and (iii) constructively explores different designs of interactive XAI explanation facilities for non-expert users. With non-expert users, I refer to end users of XAI systems who have not been trained or skilled with machine learning concepts but use its predictions to perform their tasks. Due to the diverse use cases of XAI explanations and contexts of its users, it is unlikely that one explanation will result in equally effective understanding for all. Thus, designing for understanding requires taking into account cognitive and pragmatic aspects of explanation [76, 93, 31].

To this end, this thesis makes three contributions: (i) It empirically explores unknown human factors that influence the process of understanding XAI explanations. I show that a cognitive bias of illusory understanding can emerge when non-expert users are consuming explanations [P4]. Further, I discuss the risks of deception through placebic explanations by providers of XAI explanations [P5]. (ii) It conceptually categorizes inconsistencies within the involved research communities on how explanations have been evaluated with human subjects in empirical XAI studies [P8]. Further, it conceptually outlines different types of user interaction with explanation interfaces in XAI [P3] and proposes design principles for the human-centric design of explanation interfaces. (iii) Building on these insights, it presents evaluated prototypes of an XAI explanation facilities that fulfill the requirements of *sensitivity* (elicit user beliefs to calibrate expectations) [P6], *naturalness* (use natural language explanations) [P1], and *responsiveness* (allowing follow-up interactions) [P1, P2].

This chapter motivated and grounded the context of this dissertation. Chapter 2 introduces central concepts relevant for the research goal. Chapter 3 presents the addressed research questions and the contribution(s) towards them. Chapter 4 reflects on the gained insights and outlines an agenda for future research.

2
Background and Definitions

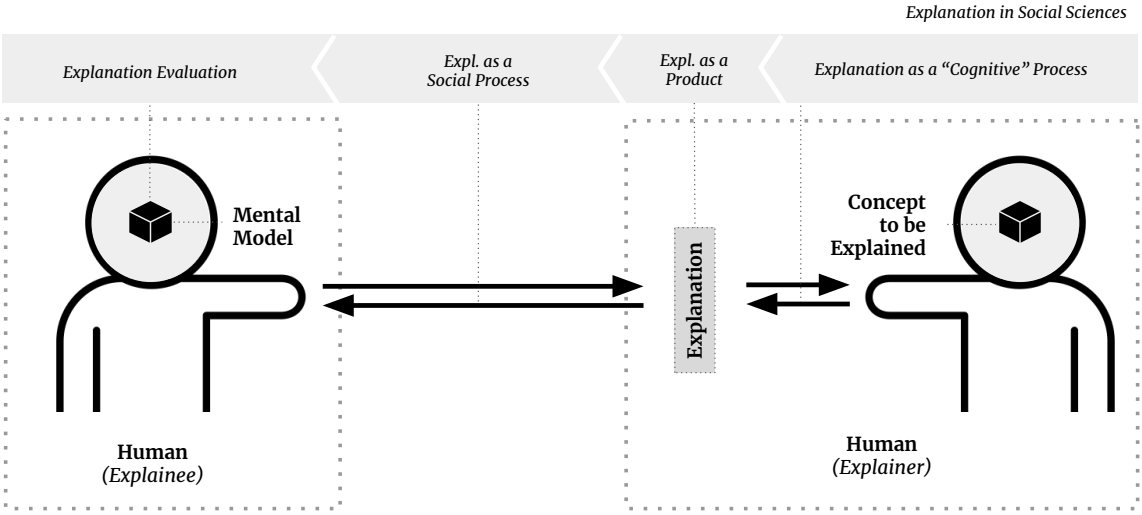


Figure 2.1: The explanation process through the lens of the social sciences.

2.1 Explanation in Social Sciences

I base my work on research on explanation in the social and cognitive sciences. First, I introduce the explanation process and different notions of the word *"explanation"*. Second, I outline how humans consume explanations to build their understanding. Finally, I describe how humans evaluate the quality of explanations during an explanation process.

Explanation Process The Oxford English dictionary defines explanation as *"a statement or piece of writing that tells you how something works or makes something easier to understand."*[1]. Miller defines explanation as either a process or a product. On the one hand, an explanation describes the *cognitive process* of identifying the cause(s) of a particular event. At the same time, it is a *social process* between an *explainer* (sender of an explanation) and the *explainee* (receiver of an explanation) with the goal of transferring knowledge about the cognitive process. Lastly, an explanation can describe the *product* that results from the cognitive process and which aims to answer a why-question [76]. The social sciences primarily consider the transactional nature of explanations between individuals. Explanation is seen as an attempt to communicate understanding between social and interacting agents. The motivation of an explainee to seek an explanation can be diverse. Keil distinguishes five explanation needs [58]: (i) *prediction* to anticipate similar events more effectively in the future; (ii) *diagnosis* to understand why a system failed and restore it; (iii) *justification* as an act of persuasion;

[1] https://www.oxfordlearnersdictionaries.com/us/definition/english/explanation?q=explanation

(iv) *accusation* to determine a guilty party; (v) *aesthetic pleasure* to increase the appreciation of an explanandum in others, e.g., explaining mysteries or poems.

Explanations and Mental Models The causal complexity of the real world makes the process of explanation incomplete or flawed. Thus, people develop heuristics to deal with the missing details and recognize flawed explanations [58]. According to Norman, people form theories for any system they interact with to reason about what they observe [90]. A *mental model* refers to a person's mental understanding of how a system works and how their behavior affects it. They can resemble logical patterns or image-like representations of a system's inner working [58]. They are formed for all kinds of systems including objects, people, and services. A respective mental model is adjusted with every interaction and helps the person to reflect their belief about the value they can expect from the system. Thus, people may use explanations provided by an explainer to adjust their mental model of the explained concept. In contrast, explainers use their mental model to formulate explanations.

Explanation Evaluation Psychologists and social scientists investigated how humans evaluate explanations for decades. Within their disciplines, *explanation evaluation* refers to the process applied by an explainee for determining if an explanation is satisfactory given the current explanation need [76]. Explanations are evaluated based on their source, process, and content:

Evaluating the Explanation Source: Explanation is evaluated based on the explainer conveying the explanation. The explainee assesses the motivational states and the competence of the explainer, i.e., explainees assess if the explainer speaks from an area of expertise or is bluffing or posturing in any way [58]. An explanation may be discounted due to motivational states when a conflict of interest becomes salient to the explainee [75]. The explainee may also discount the explanation if the explainer is perceived as incompetent because of intoxication, a lack of education, or excessive use of emotions [58].

Evaluating the Explanation Process: Explanations form an interactive conversation [48]. During this conversation people typically expect the explainer to follow general rules of conversation. The conversational statements are supposed to be linked together and form a cooperative effort to achieve the goal of information exchange. A widely accepted set of rules of conversation are *Grice's maxims* [41]. They consist of four aspects that are expected from an explainer: (i) *quality*: say only to true statements that you believe in; (ii) *quantity*: say only as much as necessary but not more; (iii) *relation*: say only to statements that are relevant for the respective context; (iv) *manner*: say it in a comprehensible way, i.e., avoid ambiguities. Even if explanations are presented in a visual way, instead of text or verbal, they should be assessed according to these properties [77].

Evaluating the Explanation Content: Most of explanations in everyday contexts follow some notion of cause and effect. The primary criterion of evaluating the content of an explanation is whether the explanation helps them to understand the underlying cause [76]. Even when an explanation contains causal and non-causal elements (e.g., correlations), the former ones dominate the explainee's judgement [85]. This is also reflected in *counterfactual thinking*. People often sense the meaning of casual relations through *"would have"* relationships. This means, with other conditions remaining the same, a particular event B would not have happened if an event A would not have happened first [68]. Scholars conducted experiments where they presented participants with different types of explanations as treatments. In practice, choosing one explanation over another is often an arbitrary choice heavily influenced by cognitive biases and heuristics [58]. For instance, humans are more likely to

accept explanations that are consistent with their prior beliefs due to *illusory correlations*. Chapman and Chapman demonstrated that humans discount strong correlations that do not match their mental model while overestimating weak correlations that do match with it [16]. According to the *inference to the best explanation* process, explainees may prefer explanations with less predictive power but a simpler internal structure (i.e., with fewer causes) over explanations with higher predictive power but seemingly unrelated elements [46]. Further, the *illusion of explanatory depth* (IOED) has been demonstrated in many contexts [101, 66, 78]. According to IOED, humans have a robust bias of overconfidence regarding their understanding of how complex systems work. After being asked to explain their understanding, people significantly reduce their estimation of their own knowledge.

Summary: Explanation is an Iterative and Heuristic Process

Explaining must be distinguished from understanding. Explaining depends on what and how something is explained by whom, while understanding also depends to whom it is explained. It is highly unlikely that a given cause can be explained in a way that satisfies every explainee. A suitable explanation for one purpose may be irrelevant for another. For an explanation process to be effective, it is essential to know the intended context of use and account for potential cognitive biases throughout the conversation.

2.2 Explainable AI

This work is based in the interdisciplinary research field of explainable AI. First, I define my notion of AI-infused systems as well as explainability and interpretability. Second, I categorize the types of methods for explainability as well as the different stakeholders of AI-infused systems. Finally, I isolate my focus from related research fields.

AI-infused Systems I build on the notion of *AI-infused systems* by Amershi et al. [3]. They define them as *"systems that have features harnessing AI capabilities that are directly exposed to the end user"*. AI capabilities in this context refer to *"activities that we associate with human thinking, activities such as decision-making, problem solving, learning"* [7]. AI-infused systems resemble the notion of an *intelligent system* which *"embodies one or more capabilities that have traditionally been associated more strongly with humans than with computers, such as the abilities to perceive, interpret, learn, use language, reason, plan, and decide"* [55]. As part of this thesis, I focus on systems where the intelligent behavior results from a black-box *machine learning* (ML) component. ML is a subset of methods to achieve AI. It refers to *"a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty* [86]. More formally, Mitchell defines ML as *"a computer program [that] is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"* [79]. From a technical perspective, ML is typically split into *supervised* learning methods, which focus on predictions based on labeled training data, *unsupervised* learning methods, which find relationships in unlabeled data, and *reinforcement learning*, which optimizes some notion of reward by interacting with an environment. Furthermore, multiple nuances of learning methods exist along the spectrum between supervised and unsupervised. As part of this thesis, I focus on supervised learning.

Interpretability of AI-infused Systems The field of *explainable artificial intelligence (XAI)* deals with methods and techniques that make the predictions and processes of ML-based systems understandable to human users. The term XAI was first used by van Lent et al. [67] in 2004 as part of their work on explanations of military simulations. It is closely related to the notion of *interpretable machine learning (IML)*. IML often refers to research on models and algorithms that are considered as inherently interpretable while XAI often refers to the generation of (post-hoc) explanations or means of introspection for black-box models [103, 124]. However, the lines between IML and XAI are often seamless and the terms are often used interchangeably. To date, there is no agreement on standard definitions for XAI and IML [2, 76, 42]. For instance, DARPA's XAI program subsumes both terms under the objective of *"[enabling] human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"* [43].

Similarly, the terms *explainability* and *interpretability* are often used interchangeably. Kulesza et al. [63] define explainability as the capability of an ML system to accurately explain the reasons for its predictions to an end user. Similarly, Doshi-Velez and Kim [24] describe interpretability as a model's *"ability to explain or to present in understandable terms to a human."* As such, these definitions are rather *system-centric* and focus on a system's functionality to provide explanations. In contrast, Miller [76] takes a more outcome-oriented *human-centered* perspective calling it *"the degree to which an observer can understand the cause of a decision"*. Similarly, Biran and Cotton consider systems as interpretable *"if their operations can be understood by a human"* [10]. Going even further, Kim et al. [59] consider a method as interpretable if a human can not only understand, but consistently predict a model's predictions. These examples illustrate the different poles of the discussion. Lipton points out that interpretability is not a monolithic concept but encompasses distinctive ideas like model transparency, trust, and fairness [71]. In this thesis, I build on the definitions by Tomsett et al. [117] as following: *Transparency* is the degree to which the system provides information about its inner workings or structure. *Explanation* refers to *"the information provided by a system to outline the cause and reason for a decision or output"*. Similarly, I define an *XAI system* as an AI-infused system that offers some form of explanation. *Explainability* is the degree to which as system can provide explanations for the underlying causes. In contrast, *interpretation* is the understanding of a user about the underlying cause and thus closely related to their mental model of the XAI system. Based on this, *interpretability* is the degree of understanding that a user gains by using explanation and transparency.

Categorization of Explainability Methods Lipton [71] distinguishes two categories of methods for interpretability: *transparency* and *post-hoc interpretability*. *Transparency* (sometimes referred to as *ante-hoc* interpretability [111]) refers to exposing the mechanisms by which the model works in an understandable way. It aims to incorporate explainability directly into a ML model. For this, often the complexity of the model is restricted (e.g., limiting the number of non-zero features [97] or enforcing monotonic constraints [37, 89]). As a result, the model is assumed to be inherently interpretable and thus the opposite of a black-box. In contrast, *post-hoc interpretability* is applied when the model is not inherently transparent. It adds explainability after the training of the ML model by analyzing its input and output relationship. Typically, post-hoc interpretability does not claim to precisely explain the mechanisms and algorithms at work. Instead, it is about conveying *"useful information of any kind"* [71] to help users building an accurate understanding of the model behavior.

In this thesis, I focus on primarily on post-hoc explanation methods. A large variety of post-hoc methods exists [42, 4]. On a technical level, they can be distinguished by their ML model requirements and

their scope of interpretability. *Model-specific* methods leverage characteristics of a ML model type, i.e., to accelerate the computation, but are limited to this specific type. In contrast, *model-agnostic* approaches do not pose any model requirements and treat every ML model as a black-box, even if it is not. For instance, KernelSHAP is a computation-intense model-agnostic method to calculate the contribution of individual feature values. In contrast, TreeSHAP is an optimized version limited to tree-based models, such as random forests [72]. Further, the scope of post-hoc methods can be local or global. *Local* methods aim to justify the ML prediction of an individual instance. In contrast, *global* methods aim to describe the prediction behavior on a more holistic level for a set of instances. Further, different categorizations of post-hoc explanation styles have been proposed [9, 23, 42, 4]. For instance, Arrieta et al. [4] distinguish between *textual explanations* (generate symbols that represent the model behavior), *visual explanations* (visualize model behavior graphically), *explanations by example* (explain through representative instances), *explanations by simplification* (approximate behavior through a simplified inherently interpretable model), and *explanation by feature relevance* (quantify the contribution of individual features towards a prediction).

Stakeholders in XAI Focusing only on the generation of explanations ignores that explainability and interpretability are two different goals. The former depends on what is explained and how it is explained, while the latter also depends on to whom it is explained. In its basic form interpretability in XAI involves two roles: the XAI system as the explainer and a human user as the explainee. However, this form is often not sufficient to describe the diverse stakeholders of an AI-infused system deployed in the real world. To comply with demands and regulations, organizations provide explanation facilities to wider non-technical audiences [8]. Different role-based models for interpretability have been proposed [49, 117, 121, 74, 6, 8]. An important role in a real-world setting is the *deployer* [121] (also referred to as *business owner* [6]) who owns the system, releases it, and is accountable for potentially undesired consequences caused by the system. From a supervision perspective, Belle and Papantonis [6] add the roles of (internal) *model risk assessors* (also referred as *AI managers* [74]), who challenge and approve the model on behalf of the business owner, and (external) *regulators*, who inspect the impact of the model on its users and individuals affected by it. Further, Hind et al. [49] distinguish two types of end users that consume the explanations provided by an AI-infused system: *end user decision makers* (also referred to as *operators* or *executors* [117]), who are often subject-matter experts that leverage the explanations to inform their decisions (e.g., physicians, loan officers, or judges), and *affected end users* (also referred to as *decision subjects* [117]), who are affected by an individual decision (e.g., patients, loan applicants, defendants). While developers are mainly interested in technical details on how an underlying ML model works, the other roles often have limited ML knowledge and focus more on understanding what input parameters drive the model's predictions and when the predictions can be trusted [6]. As a result, an effective XAI system needs to model the user's context and background and provide personalized explainability.

This thesis focuses on the two types of end users. End user decision makers may be accountable for their prediction-informed decisions. Thus, they utilize explanations to assure the underlying model is *trustworthy* (i.e., *"they can reasonably trust a model's outputs"* [8]). As such, they require interpretability on a local (i.e., to argue for individual predictions) as well as global level (i.e., to understand capabilities and limitations of the AI-infused decision support). Following [117], I refer to this as *operator-interpretability*. In contrast, affected end users may seek local explanations to challenge their individual decision or understand how they need to change their parameters to influence the decision. Following [117], I refer to this as *contestability*.

Related Concepts There are other concepts that are closely related to interpretability in AI. Based on my definition of interpretability, I describe common ground and distinctions.

Scrutability: This concept is about *"allowing users to tell the system if it is wrong"* [115]. It is widespread in the context of social recommender systems [5, 62]. These provide users with individualized recommendations based on an estimated model of a user's preferences. Some notions of scrutability are limited to the aspect of explainability [13]. For instance, Cheverst [17] refers to *"the ability of a user to interrogate her user model in order to understand the system's behavior"*. However, typically scrutability it is not limited to receiving explanations but also allows users to debug or correct system assumptions [120]. As such, scrutability extends interpretability through some feed forward or control interaction that provides end users *"with a direct and meaningful way to revise their [user] model"* [5, 110].

Interactive Machine Learning (IML): This concept describes *"an interaction paradigm in which a user [...] iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review"*. The paradigm is centered around the back-and-forth dialogue between a user and a system. Unlike scrutability, IML is not about the individual user model but about any problem domain where the objective may be unclear and data labels unavailable a priori. It aims to combine the complementary strengths of both: Users can train a ML model without explicit programming knowledge by demonstrating or labelling samples while the model benefits from users' domain expertise. In each iteration both sides are directly influencing each other's behavior (co-adaptive) [27]. In contrast to interpretability, IML largely treats ML as a black-box. Users may change inputs or parameters of the model and observe its changes in the output, however, explaining why certain changes occur (explainability) is often not the primary concern [63].

Study of Machine Behavior: This concept describes the empirical analysis of the behavior of intelligent machines and their effects on humans in the wild. For this, it draws analogies from the empirical study of animal behavior and involves not only the AI design and engineering disciplines but those that study biological agents. The concept distinguishes a *proximate view* that investigates how the machine functions and an *evolutionary view* that investigates why a certain type of behavior evolved and how it adapts to human stakeholders over time [95]. Aspects of the proximate view center on the causal mechanisms of a machine and how they shape human behavior (e.g. as decision aids) and thus overlap with the focus of this thesis. Unlike interpretability, which often focuses on the current behavior of a single static system, this concept also inquires the mutual influences between multiple systems and multiple users over time. As such, it takes a macro perspective on human-AI interaction.

Summary: Interpretability for End Users of XAI Systems

I center this thesis in the context of AI-infused intelligent systems that have a black-box component and that are exposed to end users with limited AI knowledge. I focus on the process and degree of understanding that end users, such as decision-makers or decision-subjects, obtain when exposed to post-hoc explanations (i.e., the interpretability). Unlike research on scrutability and interactive ML, I focus on situations where the end users have no means to influence the ML behavior. Unlike the study on machine behavior, I focus mainly on the micro perspective of the interaction between a single XAI system and a single end user.

3

Human-centric Explanation Facilities

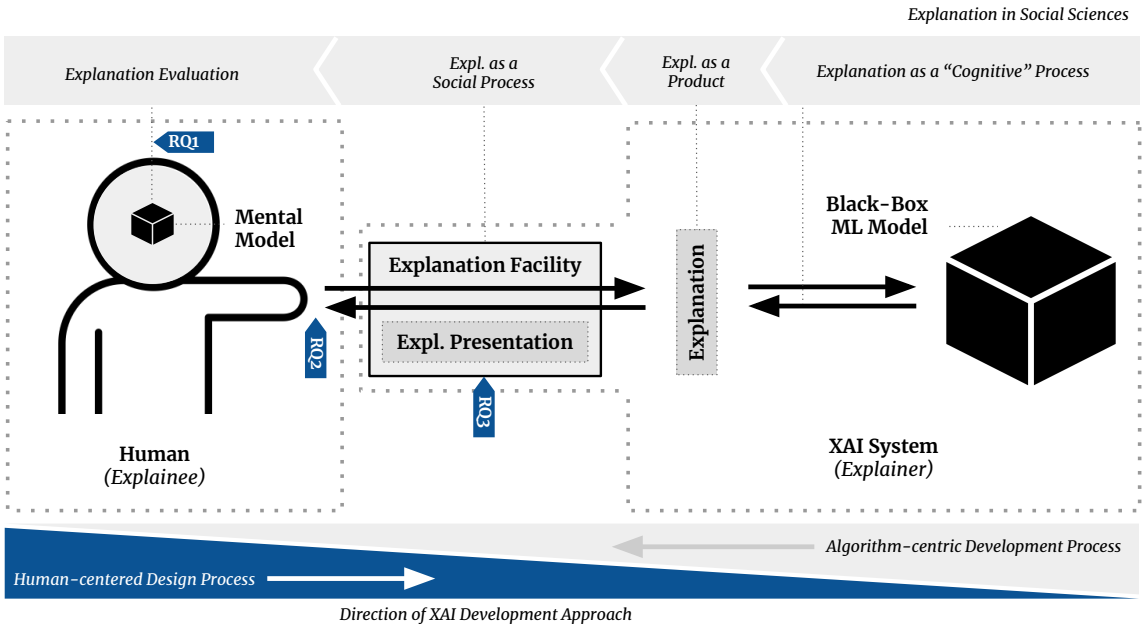


Figure 3.1: In this thesis, I analyze the XAI development approach through the lens of the social sciences [76]. I contribute to research on cognitive *human factors* in XAI (RQ1), concepts of *human involvement* in XAI (RQ2), and constructive designs for *human-centric* XAI (RQ3).

3.1 Research Problems and Questions

When humans are questioned about their actions, they provide arguments to the questioning party and thus convey an underlying reasoning. Similarly, AI-infused systems should be capable of justifying their behavior by providing some notion of explanation. XAI research approaches this challenge from multiple perspectives. The *algorithm-centric* perspective focuses on technical methods and solutions that can explain the behavior of the underlying ML model. Many formal and mathematical methods have been developed that explain the inner workings of ML models. However, despite their formal rigor, they often lack usability and practical efficacy for real users [1].

There is a growing acceptance that building XAI systems requires a multidisciplinary effort involving technical, HCI, cognitive, and domain-specific perspectives. This is also reflected in the applied research methods by some parts of the XAI research community. The focus shifted from evaluating the algorithmic performance of an AI system to evaluating the human performance and satisfaction with an AI-infused system [107]. System developers put human users at the center and apply user-centered participatory design methods that include all stakeholders into the development and evaluation process - not just system engineers [31]. Evaluations with human subjects verify the system fits the users' needs, goals, and beliefs. In this chapter, I present HCI research problems which emerge from

a *human-centric* perspective on XAI systems. I structure them based on the categorization of HCI research problems by Oulasvirta and Hornbæk [92] into (i) *empirical* (describing real-world phenomena related to the human use of computing), (ii) *conceptual* (explaining unconnected phenomena in the human use of computing), and (iii) *constructive* (understanding the construction of artefacts related to the human use of computing). These serve as guiding questions for this thesis.

Empirical: Human Factors in XAI The algorithm-centric perspective focuses primarily on what about the AI can accurately be explained. This approach, while it may yield factually correct explanations, may not be sufficient to generate trust with the user [93]. It neglects how an explanation is evaluated by the human recipient in practice. Following a human-centric perspective, Paez [93] argues to instead make the *pragmatic* understanding of users, that results from any form of explanation, the unit of analysis. Similarly, Eiband et al. propose a *pragmatic* perspective on understanding that balances the cognitive load of an explanation, their seamless integration into a user workflow, and a sufficient understanding instead of a comprehensive one [31]. The path from explanation to understanding is not straight forward but influenced by cognitive biases and reasoning fallacies [120]. Further, non-technical end users may not be aware of the *"forms of uncertainty that are baked into ML predictions"* [127]. Still, the amount of research that empirically explores the role of such human factors in the context of XAI is limited (e.g., [127, 106, 119, 91, 60, 14]). For example, Schaffer et al. [106] show that people with low competence at a given task tend to overestimate their task understanding and thus ignore XAI explanations. Nourani et al. [91] show that a positive first impression by end users of an XAI system may lead to automation bias. More research has been conducted in the older field of *decision support systems* investigating automation [15, 109, 73], anchoring [36, 65], and confirmation biases [53, 112]. Building on the insights from the field of decision support systems and cognitive sciences may yield strategies on how to identify and mitigate cognitive biases. Thus, this thesis is guided by the following research question:

RQ1: *How is end users' understanding of XAI explanations impacted by human factors?*

Conceptual: Human Involvement in XAI Beyond their interest as a stakeholder of an XAI system, humans may serve different roles in a human-centric XAI development process. Most naturally, they are the *consumer* of the deployed XAI artifact that results from the development process. As such, prior research frames XAI as a human-agent interaction problem [76] between a human user and an AI agent towards an explanatory goal that is mediated through an explanation user interface (XUI). However, there are *conceptual inconsistencies* about the role of user interaction on end users' understanding of XAI systems. XAI research often implicitly assumes that there is a single message to be conveyed through an explanation [1, 76]. However, in decision-making situations that demand interpretability, it is unlikely that a single static explanation can address all concerns and questions of a user. This resonates with the social science perspective that considers explanation as a social process. Further, humans may serve the role of the *evaluator* who informs, guides, or assesses the XAI development process. Prior surveys identified a need for more rigid empirical evaluations of XAI explanations [2, 77, 25]. Yet, since there is no consensus on evaluation methods, the comparison and validation of diverse explanation techniques is an open challenge [2, 24]. Thus, there are *conceptual inconsistencies* within the involved research communities on how human understanding should be evaluated in empirical XAI studies. Reflecting on the involvement of humans as consumers and evaluators, this thesis addresses the following research questions:

RQ2a: *How is user understanding of XAI evaluated in empirical studies?*

RQ2b: *How is interactivity used in XAI to promote user understanding?*

Constructive: Human-centric Design of XAI for Non-Expert Users The call for explainability in AI systems is not new. In the 1970s and 1980s explanation facilities were incorporated into so called *expert systems*. XAI researchers pointed out the importance of building on these *"insights from more than four decades of human-centered research on explanation in AI systems"* [83]. In this thesis, I explore the notion of so called *explanation facilities* that dates back to this time. It centers around the idea that an explanation interface targeting non-experts should include multiple modes of explanation that fulfill the requirements of *fidelity* (generate accurate explanations), *naturalness* (explanations in natural language following a dialogue), *responsiveness* (allow follow-up questions and alternative explanations), *flexibility* (make use of multiple explanation methods to allow different explanations for different contexts), *sensitivity* (provided explanations should be informed by the user's knowledge, goal, context, and prior interaction), *extensibility* (allow to include novel explanation methods), *portability* (allow to be tailored to a specific domain), and *adaptivity* (automatically learn from interaction over time) [81]. The DARPA XAI program illustrates the XAI process as a two-staged approach. It distinguishes between the explainable model, which taps into the ML model to generate explanations, and the explanation interface, which the user directly interacts with [43]. Such a two-staged approach disentangles the XAI process into analyzing the ML model behavior and communicating it to the user. Thus, in the last part of this thesis, I explore how the requirements by Moore and Paris [81] may be leveraged for the design of usable XAI explanation interfaces.

RQ3: *How can interactive explanation facilities be designed to promote end users' understanding, taking human factors into account?*

Table 3.1 summarizes the identified research problems and guiding research questions.

Research Problem	Description	Research Question
Empirical	There are **unknown phenomena** and **unknown factors** that influence the interpretability of XAI explanations due to heuristics in users' explanation evaluation.	**RQ1:** How is end users' understanding of XAI explanations impacted by human factors?
Conceptual	There are **conceptual inconsistencies** within the involved research communities on how human understanding should be evaluated in XAI.	**RQ2a:** How is user understanding of XAI evaluated in empirical studies?
	Furthermore, there are **conceptual inconsistencies** about the role of user interaction on end users' understanding of XAI systems.	**RQ2b:** How is interactivity used in XAI to promote user understanding?
Constructive	There are **partial** and **ineffective solutions** that focus on end users' understanding and that account for human factors in XAI.	**RQ3:** How can interactive explanation facilities be designed to promote end users' understanding, taking human factors into account?

Table 3.1: The HCI research problems [92] addressed in this thesis take a human-centric perspective on XAI systems. In particular, we focus on the interaction between end users and XAI systems from an HCI perspective.

3.2 Contributing Publications

The dissertation is cumulative, i.e., it consists of multiple peer-reviewed publications. All contributing publications are listed in the reference list below. The *Appendix: Original Publications* contains the original publications as published or submitted. Throughout this dissertation, a contributing publication is referenced with a prepended "P" (e.g. [P4]). As most contributing publications resulted from joint work, I use the scientific "we" in the following sections.

[P1] Michael Chromik. "Making SHAP Rap: Bridging Local and Global Insights through Interaction and Narratives". To appear in *18th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'21)*. (Cited on pp. 3, 21, 22, A 1).

[P2] Michael Chromik. "reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP". In: *Proceedings of 18th European Conference on Computer-Supported Cooperative Work (ECSCW 2020)*. EUSSET, 2020. DOI: 10.18420/ecscw2020_p06 (cited on pp. 3, 20–22, A 1).

[P3] Michael Chromik and Andreas Butz. "Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces". To appear in *18th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT'21)*. 2021 (cited on pp. 3, 18, 19, 22, 23, A 1).

[P4] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. "I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI". In: *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI 2021)*. ACM, 2021 (cited on pp. 3, 14, 15, 17, 22, 23, A 1).

[P5] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. "Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems". In: *Explainable Smart Systems Workshop at the 24th International Conference on Intelligent User Interfaces (IUI 2019)*. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-7.pdf. 2019 (cited on pp. 3, 16, 17, 22, 23, A 1).

[P6] Michael Chromik, Florian Fincke, and Andreas Butz. "Mind the (Persuasion) Gap: Contrasting Predictions of Intelligent DSS with User Beliefs to Improve Interpretability". In: *Companion Proceedings of the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. EICS '20 Companion. ACM, 2020. DOI: 10.1145/3393672.3398491 (cited on pp. 3, 20–22, A 1).

[P7] Michael Chromik, Florian Lachner, and Andreas Butz. "ML for UX? - An Inventory and Predictions on the Use of Machine Learning Techniques for UX Research". In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM, 2020. DOI: 10.1145/3419249.3420163 (cited on pp. 16, 17, 22, A 1).

[P8] Michael Chromik and Martin Schuessler. "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI". In: *Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies Workshop on Intelligent User Interfaces (IUI 2020)*. http://ceur-ws.org/Vol-2582/paper9.pdf. 2020 (cited on pp. 3, 18, 19, 22, 23, A 1).

3.3 Empirical: Human Factors in XAI

RQ1: *How is end users' understanding of XAI explanations impacted by human factors?*

3.3.1 [P4] The Illusion of Explanatory Depth in XAI

Summary: Research in cognitive sciences indicates that humans often form an inaccurate understanding of complex systems and overrate the depth of the understanding they gain from explanations [82]. Rozenblit and Keil named this type of overconfidence bias the *illusion of explanatory depth (IOED)* [101]. No empirical has been published on the potential of an IOED in XAI although some researchers speculated that it may be at play when users deal with explanations from XAI systems [19, 111, 82, 57].

In this conference publication, we investigated the robustness of the perceived understanding that end users gain from local XAI explanations. For this, we exposed participants to a widely used local explanation method, namely Shapley based feature attributions, and examined their understanding. We measured participants' subjective and objective understanding at multiple stages of the study procedure through self-ratings and different tests of understanding (e.g., self-explanation and mental simulation of the XAI system behavior). We applied a mixed-method approach consisting of a moderated think-aloud study (40 participants) and an unmoderated crowd sourcing study (107 participants) to account for analytical and heuristic modes of reasoning. Our results show that, on average, participants in both studies decreased their perceived understanding over time, indicating an IOED effect. Participants who were guided by heuristic thinking spent significantly less time and had a significantly lower objective understanding. Still, they reported higher perceived understanding and were more confident in their predictions of the XAI system behavior than their analytical counterparts. With our work, we highlight the need of XAI systems to capture wrong or incomplete mental models of end users to support interpretability, e.g., by adjusting the form or phrasing of an explanation. Further, we describe the observed reasoning and interaction strategies that participants applied during their exploration of local XAI explanations. Our approach and insights inform future work on the design of interactive explanation facilities that elicit user's mental model of the underlying ML model and account for human factors in the interpretation of XAI explanations.

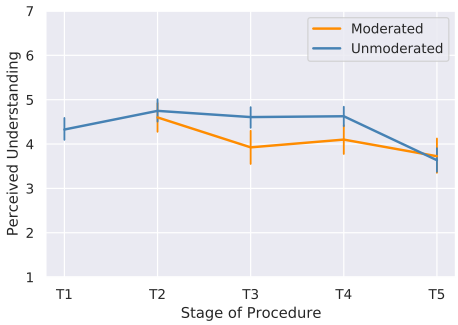


Figure 3.2: The means of perceived understanding in the moderated and unmoderated studies after different tests of understanding.

Author Contributions: I came up with the research idea, concept, study design, and technical implementation of the apparatus. Further, I was the leading author of this publication. Malin Eiband provided feedback throughout this process. Felicitas Buchner conducted and analyzed the moderated user study. Adrian Krüger conducted and analyzed the unmoderated user study. Andreas Butz provided feedback and revised the publication for clarity and conciseness.

3.3.2 [P5] Dark Patterns of Explainability in XAI

Summary: Humans take the producer of the explanation, their intentions, and integrity into account. The evaluation of non-experts may differ if the explanation has been provided by external parties compared to trusted internal parties [75, 58]. Prior work in XAI raised concerns that the call for explainability may result in explanations that pursue other goals than promoting user understanding due to misaligned interests [121].

In this workshop publication, we present a set of dark design patterns of explainability as a thought-provoking discussion paper. We build on the concept of dark design patterns in UX by Brignull [11] and Gray et al. [40]. These describe *"a user interface that has been carefully crafted to trick users [...] and do[es] not have the user's interests in mind"* [12]. We transfer this concept to the design of explanation facilities. We provide examples of dark patterns in the phrasing of explanations and in the interaction with explanation interfaces. We discuss situations of opposing interests between the explainer and explainee in XAI that could be argued as questionable or unethical. For example, the dark pattern of *obstruction* makes it intentionally hard to get (useful) explanation about the AI decision-making and thus result in users shunning from the additional effort to question the system. With our work, we reflect on the practical challenges and human complexities of a mandatory call for explainability, such as the *right to explanation* as part of the *General Data Protection Regulation (GDPR)* [114], when the interests of the explainer and explainee are not aligned. Further, we discuss the role of HCI design practitioners in the ethical design of explanation facilities and propose dark patterns as a baseline for human-subject evaluations in XAI.

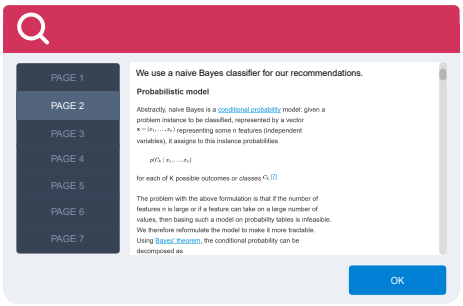


Figure 3.3: Example of the *information overload* dark pattern. The given explanation is lengthy and uses technical language not suitable for non-experts.

Author Contributions: Daniel Buschek and Malin Eiband came up with the initial idea and concept of this work. Daniel Buschek and Sarah Völkel contributed substantially to the description and visualization of the dark patterns. I was the leading author and contributed substantially to the motivation, background, dark pattern examples, discussion, and overall alignment of this publication.

3.3.3 [P7] The Potentials of (X)AI for User Experience Research

Summary: An active field of HCI research explores how UX designers and UX researchers can enhance the user experience of AI-infused systems through effective human-AI interaction [64, 38, 3, 125]. However, little research has been conducted on how UX practitioners could leverage ML methods to enhance the UX activities themselves [26, 126].

In this conference publication, we empirically investigated the potential synergies between empathy-focused user experience research (UXR) and data-driven ML techniques. UXR activities rely on generating insights about the targeted users' perspectives to inform the design process of products and services. Thus, we speculated that explainability would be an important aspect for the acceptance of ML for UXR. To understand the current practices in the field, we surveyed 49

practitioners from the fields of UX and ML. Further, we complemented our learnings through 13 semi-structured interviews with UX practitioners who were educated or experienced in ML. Our results indicate that the disciplines of ML and UX are increasingly overlapping and that UX practitioners envision opportunities to automate their mundane tasks, complement their decisions with data-driven insights from multiple sources, and enrich UXR with insights from users' emotional worlds. Challenges were perceived to result from an increasing obligation to utilize quantitative data over qualitative insights, ensuring the effectiveness of ML-based UXR after deployment, and a more restrictive access to user data. Explainability was a minor concern. With our work, we provide insights about the impact of ML on current UX practices, its technological potentials as well as its social and organizational challenges. Further, we identify the real-time UX evaluation of products and services through ML as a promising use case for future research.

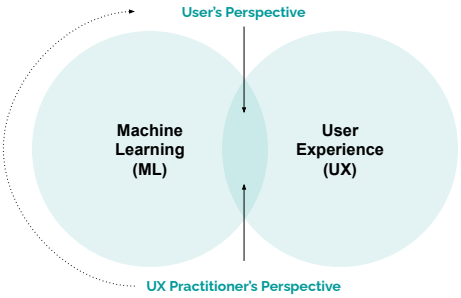


Figure 3.4: We investigated UX practitioners' perspective on ML for UX activities.

Author Contributions: Florian Lachner contributed substantially to this research. He came up with the initial idea and methodology of this work. He also conducted the qualitative expert interviews and designed the survey. I contributed substantially to the analysis of the qualitative and qualitative study data and I was the leading author of the publication. Andreas Butz provided feedback and revised the publication over multiple iterations.

Summary: RQ1: Human Factors May Have an Impact on Interpretability

My work shows that the path from explainability to interpretability of XAI systems is not straight forward and cannot be taken for granted. It requires attention and careful consideration by XAI developers and should take the individual mental models of end users into account. Regarding the impact of human factors on end user understanding of XAI explanations, this thesis contributes (i) two user studies that show that users may form an illusory perception of the interpretability they gain from local XAI explanations [P4], (ii) a set of dark patterns in XAI, which show that explainability and interpretability can deceptively be decoupled, and which may inform the ethical design of XAI [P5], and (iii) two user studies that elicit the technical potentials and social challenges of ML methods for UX activities [P7].

3.4 Conceptual: Human Involvement in XAI

RQ2a: *How is user understanding of XAI evaluated in empirical studies?*

3.4.1 [P8] A Categorization of Human-Subject Evaluation of XAI

Summary: Prior surveys identified a need for more rigid empirical evaluation of XAI explanations [2, 77, 25]. Yet, since there is no consensus on evaluation methods, the comparison and validation of diverse explanation techniques is an open challenge [2, 24]. There are two approaches to explainability evaluation. *Functional* evaluation through mathematically quantifiable metrics and *human-subject* evaluation through user studies.

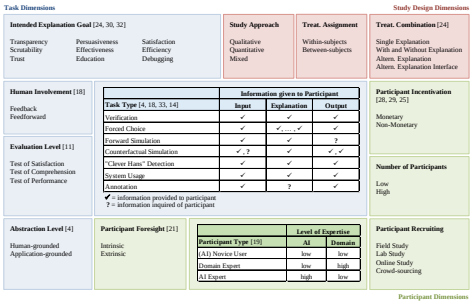


Figure 3.5: Categorization of human subject evaluation in XAI based on *task-related*, *participant-related*, and *study design-related* dimensions.

In this workshop publication, we analyzed the latter. To show an explanation method's utility for practical use cases [94], each promising functional evaluation should be succeeded by human-subject evaluations at some point in time. We conducted a literature review based on 653 search results and analyzed a sample of 34 publications that either report or discuss study design decisions in evaluations of XAI explanations with human subjects. We consolidate our insights into a categorization based on *task-related*, *participant-related*, and *study design-related* characteristics. For example, the dimensions *type of user task* presents seven strategies which have been proposed to elicit the quality of explanations. We categorize them by the information provided to the participant and the information inquired in return. With our work, we inform researchers and practitioners about the design and reporting of user studies that assess the utility of XAI explanations.

Author Contributions: I came up with the research idea, concept, methodology, and were the leading author of this publication. Martin Schuessler contributed significantly to the development of the categorization and revised the publication for clarity and conciseness.

RQ2b: *How is interactivity used in XAI to promote user understanding?*

3.4.2 [P3] A Categorization and Design Principles for Human-XAI Interaction

Summary: Prior research frames XAI as a human-agent interaction problem [76]. As such, it is about the interplay between a human user and an AI agent towards an explanatory goal that is mediated through an explanation user interface (XUI). Tintarev and Masthoff [116] distinguish seven explanatory goals, such as satisfaction, effectiveness, or efficiency. These goals are often in conflict with each other. Thus, designers of XUI *"need to make trade-offs while choosing or designing the form of interface"* [118].

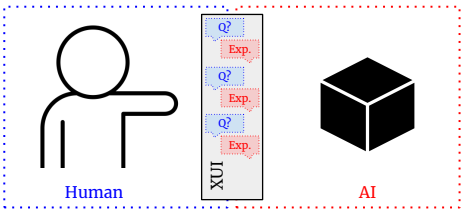


Figure 3.6: The concept of *human-XAI interaction as a dialogue* aims to facilitate a natural and iterative conversation about AI behavior with the goal of transparency or scrutability.

In this conference publication, we took an HCI perspective on how prior XAI research approached these trade-offs and how it designed the interplay between the user and the XAI system. We conducted a structured literature review based on 146 search results and analyzed 91 publications that either present constructive research involving an XUI or conceptual research addressing interaction in XAI. From there, we built on the conceptualization of human-computer interaction by Hornbæk and Oulasvirta [52] and narrowed it down to seven concepts of human-XAI interaction. Further, we describe four observed design principles for interactive XUI, discuss why each is relevant, and how it could be implemented. With our work, we organize the current literature that involves an XUI and contribute a categorization for the HCI community to describe existing and new works in XAI based on the form of user involvement. Further, we describe design principles that serve researchers and practitioners as a starting point for planning and designing human-centric XAI systems.

Author Contributions: I came up with the research concept and was the leading author of this publication. Andreas Butz provided feedback and revised the publication for clarity and conciseness.

Summary: RQ2: Human Involvement in XAI

My work highlights that the human explainee should not be considered a passive receiver of an XAI explanation. The HCI community perceives human involvement as a vital part for the success of the XAI development process. Regarding the type of human involvement in XAI, this thesis contributes (i) a categorization of human-subject evaluation approaches taken in prior literature [P8] and (ii) a categorization based on prior literature of the type of interplay between an explainee and an XAI system that is mediated through an explanation interface [P3].

3.5 Constructive: Human-centric XAI for Non-Expert Users

RQ3: *How can interactive explanation facilities be designed to promote end users' understanding, taking human factors into account?*

3.5.1 [P2] A Proposal for a Responsive Explanation Facility Framework

Summary: In this poster publication, I build on the notion that usable explanation facilities should be *responsive* (allow follow-up questions and alternative explanations) [81]. I propose and outline a web-based UI framework for interactive explanations based on the explainability framework SHAP. It aims to enable end users to interactively explore the ML model behavior and verify their hypotheses about it. Furthermore, from the trail of the user interactions, the XAI system may derive information about the user's mental model and preferences to personalize the provided explanations.

Author Contributions: I came up with the proposal and was the leading author of this publication.

3.5.2 [P6] A Sensitive Explanation Facility Based on User Belief Elicitation

Summary: Prior research shows that a lack of interpretability can lead to users mistrusting, misusing, or rejecting a system [70, 84]. Often these result from a perceived mismatch between users' expectations and the actual behavior of a system [29]. The requirement of *sensitivity* calls for explanations that are informed by the user's knowledge, goal, context, and prior interaction [81].

In this conference publication, we explored how multi-criteria decision-making may be used as a basis for sensitive explanation facilities. In a real-world case study, we investigated the interpretability needs of decision makers of an AI-infused decision support system in the construction industry. We followed a human-centered design process to derive requirements and user needs. Based on these, we explored design opportunities for usable explanations using prototypes and user studies. We used the multi-criteria decision-making technique *Analytic Hierarchy Process (AHP)* [104, 34] to elicit a user's belief about the decision-making situation and contrasted their belief with the ML prediction. This approach allows identifying persuasion gaps, i.e., situations in which the XAI system and the user base their decision on different criteria. Further, we report insights from a formative evaluation with 7 domain experts.

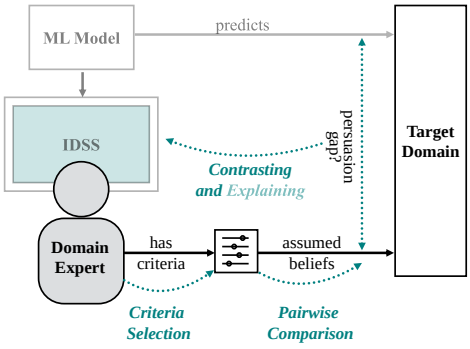


Figure 3.7: Contrasting user beliefs (elicited through AHP) and ML predictions to identify persuasion gaps.

Author Contributions: I came up with the research idea and was the leading author of this publication. Florian Fincke contributed significantly to the methodology, apparatus, and analysis of the user studies. Andreas Butz provided feedback and revised the publication for clarity and conciseness.

3.5.3 [P1] Bridging Local and Global Insights through Interaction

Summary: Shapley explanations are widely used in practice. However, they are often presented as visualization and thus leave their interpretation to the user. As such, even ML experts have difficulties interpreting them [57]. On the other hand, combining visual cues with textual rationales has been shown to facilitate understanding and communicative effectiveness [30, 20].

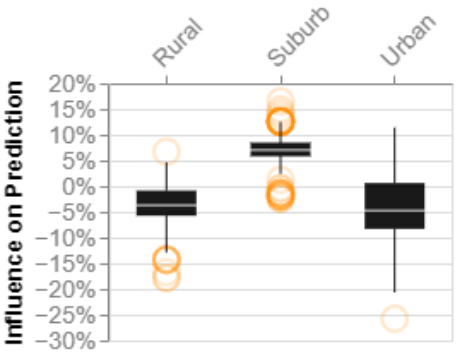


Figure 3.8: We use the distribution of local Shapley explanations for each feature value to provide insights about the global ML model behavior.

In this conference submission, we pursued to improve the accessibility of Shapley explanations for end users. We build around the requirements of *naturalness*, which calls for explanations in natural language, and *responsiveness*, which calls for multiple complementary explanations methods [81]. We present an interactive explanation facility artifact that provides local Shapley explanations and complement them with global explanations about prior predictions. Since Shapley values are additive and consistent, the explanations of individual instances can be aggregated over multiple instances to approximate the global prediction behavior of the ML model. Further, we provide reassuring rationales in natural language to support user understanding.

Author Contributions: I came up with the research proposal and was the leading author.

Summary: RQ3: Human-centric Explanation Facilities

The ultimate purpose of XAI is to foster human understanding. Therefore, it is not sufficient to limit the boundaries of the XAI system to generating an accurate explanation of the AI behavior (*explanation as a product*). In my work, I took the approach that an XAI system designed to foster human understanding needs an explanation facility that moderates the social process of explanation between the user and the XAI system. The goal of the explanation facility is to close the gap between the real AI behavior and the explainee's mental model of it.

In [P2] and [P1], we propose a reactive approach. The social process of explanation is driven by the explainee until the explainee reaches a satisficing level. The explanation facility is not aware of the gap and does not keep track of it. The explanation facility in [P2] offers different explanation requests and follows the requirement of *responsiveness* [81]. In [P1], we present an explanation facility that provides local as well as global explanations and complements them with textual explanations in natural language. As such, it follows the requirements of *responsiveness* and *naturalness* [81]. In [P6], we propose an active approach. Here, the explanation facility guides the explainee through an onboarding process at the beginning of the interaction to elicit the user's beliefs about a target domain. In this way, the explanation facility becomes aware of potential gaps and can keep track of them. Further, explainee and XAI system can drive the social process of explanation (e.g., by proactively highlighting predictions that violate the user's mental model). This approach follows the requirement of *sensitivity* by Moore and Paris [81].

3.6 Contribution

This thesis strives to advance the body of knowledge about the human use of XAI systems. In Table 3.2, I summarize the research contribution of the contributing publications. I categorize them based on the seven types of knowledge in HCI proposed by Wobbrock and Kientz [122].

Research Question	Knowledge Type	Contribution
RQ1: How is end users' understanding of XAI explanations impacted by human factors?	Empirical	We present results from an empirical examination with two user studies indicating that end users of XAI systems may form an illusory understanding of the AI prediction behavior if their explanation evaluation is unguided. [P4]. Further, we present results from an online survey and semi-structured interviews with UX and ML practitioners. We derive an inventory of the technological potentials and social challenges of applying ML to UX activities [P7].
	Opinion	We present a thought-provoking essay about potential dark patterns that may result from deceptive explanation facilities when the interests of explainee and explainer are not aligned. [P5].
RQ2a: How is user understanding of XAI evaluated in empirical studies?	Survey & Theoretical	We organize a sample of prior literature that evaluated XAI explanations through user studies. From this, we derive a categorization based on task-related, participant-related and study-design related characteristics that guide the evaluation of XAI artifacts [P8].
RQ2b: How is interactivity used in XAI to promote user understanding?	Survey & Theoretical	We organize prior literature on explanation user interfaces based on a systematic meta-analysis. From this, we derive a categorization of the type of interplay between a user and an XAI system. Further, we propose guiding design principles for the design of human-centric explanation facilities to promote user understanding [P3].
RQ3: How can interactive explanation facilities be designed to promote end users' understanding, taking human factors into account?	Artifact	We present a prototype of an explanation facility targeting non-expert users of XAI systems that embeds local Shapley explanations in an accessible spreadsheet-like user interface [P8]. In [P1], we extend this explanation facility with global explanations and complement them with textual explanations to improve accessibility. Further, we present a prototype of a sensitive explanation facility that is aware of their users' beliefs about the target domain [P6]. The prototype resulted from a human-centric design process with financial decision makers in the construction industry. Lastly, we outline and propose the development of a framework for a responsive explanation facility that allows various follow-up explanations based on the explainability framework SHAP [P2].

Table 3.2: Summary of types of knowledge in HCI [122] presented in this thesis.

4
Discussion

4.1 A Research Agenda for Human-centric XAI

This thesis presents aspects and approaches relevant to the design of human-centric XAI systems. It highlights the importance of guiding users to achieve interpretability. However, further work is required along the lines of user modeling, user adaptation, and performance-based XAI evaluation to achieve truly effective XAI systems that foster human understanding.

Constructing User Models Mental models are the blueprints of a person's understanding about a complex system. Our research indicates that for causally complex systems, such as XAI systems, users may form an inaccurate understanding when explanations are merely presented following a *human-XAI interaction as information transmission* concept [P4, P3]. Similarly, Hoffman and Muller [50] consider a single explanation artifact, such as a statement, image, or alike, as not sufficient to qualify as *"being an explanation"*. Instead, *"being an explanation"* needs to characterize a bi-directional interactive activity in which even the explainer may sometimes ask questions to the explainee to better facilitate the explanation process. As such, a human-centric explanation facility *"must possess (or create) a model of the learner's mental model"* [50]. Building on research on intelligent tutoring systems has been proposed as promising pathway for further research.

Accounting for Dynamic Explanation Needs Further, the explanation needs of users are not static. They evolve *"as one builds understanding and trust during the interaction process"* [69]. Previous research describes a differential impact of explanations on novice users compared to experienced users. Novice users are more likely to adhere to predictions as they are lacking the domain knowledge. In contrast, experts require strong domain-oriented arguments to be convinced to adhere to a prediction [106]. A user's experience with the target domain is represented by the cognitive chunks they can effectively process and understand [24]. Making these chunks the unit of explanation and adjusting it over time may be a promising pathway to explore.

Protocol for XAI Evaluation Our work in [P4] and [P5] highlights the importance of human-subject evaluation to account for human factors in their explanation evaluation. In [P8], we outline how XAI systems may be evaluated through user studies. However, there are further pitfalls in interpreting XAI user studies that researchers should be aware of. A common assumption in XAI is that good measure of performance during an XAI evaluation is simultaneously an indicator for a complete mental model [51]. However, research indicates that subjective evaluation with measures such as trust and preference may not correspond to the ultimate performance with the system [14]. Also, think-aloud studies may not convey how people make decisions with XAI in realistic setting. Some researchers argue that human subject evaluations imply a strong bias towards simpler but more inaccurate explanations due to implicit human biases. This poses the risk to create persuasive explanations instead of accurate ones [47]. Thus, human subject evaluation can only be one part of the evaluation chain in XAI. Looking forward, Jesus et al. [56] outline an application-grounded evaluation protocol that relies on users' performance metrics. This allows them to statistically compare explanation methods

in terms of their efficiency and accuracy. Research aiming to integrate functional evaluations and human-subject evaluations while accounting for human biases in evaluation may help the interdisciplinary research communities to develop an end-to-end evaluation protocol for XAI.

4.2 Concluding Remarks

"Given enough eyeballs, all bugs are shallow." Linus's Law, 1999 [96]

Explainability and interpretability may not be demanded for predictions with limited consequences, such as music or movie recommendations on Spotify or Netflix. However, if citizens' freedom is rated by AI-infused systems that are not understood by the judges nor audited by independent parties it causes societal protests and discussions[1]. More recently, massive protests by students, teachers, and other civic bodies against automated A-levels grading predictions during the COVID-19 pandemic in the United Kingdom showed that understanding the behavior of predictions is no longer a matter of engineers[2].

As such, AI-infused systems deployed into contexts with high-stakes decisions need to be inclusive to audiences beyond their developers and foster a pragmatic level of understanding. More information workers will be confronted with predictions by AI-infused systems in the future. Allowing non-expert end users, who may be affected or held accountable for predictions of an AI-infused systems, to engage with its prediction behavior in an accessible way is increasingly demanded. Building on the idea of Linus's Law [96], I believe that *given a large enough base of users with a satisficing level of interpretability, almost every AI problem will be characterized quickly and the fix obvious to someone*. Equipping AI-infused systems with means of transparency and explainability is a prerequisite for this. However, it does not necessarily lead to interpretability. They must take users' mental models into account, offer the right kind of user interaction, and be evaluated with human subjects to ensure their effectiveness. I hope that my work raises the awareness of XAI system designers to the human aspects in their quest for pragmatic interpretability.

[1] https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/

[2] https://www.wired.co.uk/article/alevel-exam-algorithm

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Association for Computing Machinery, 2018, pp. 1–18. DOI: 10.1145/3173574.3174156 (cited on pp. 3, 11, 12).

[2] Amina Adadi and M. Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160 (cited on pp. 8, 12, 18).

[3] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, P. Bennett, Kori Inkpen Quinn, J. Teevan, Ruth Kikin-Gil, and E. Horvitz. "Guidelines for Human-AI Interaction". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019) (cited on pp. 1, 7, 16).

[4] A. Arrieta, Natalia D'iaz-Rodr'iguez, J. Ser, Adrien Bennetot, S. Tabik, A. Barbado, Salvador Garc'ia, Sergio Gil-L'opez, D. Molina, Richard Benjamins, R. Chatila, and F. Herrera. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *ArXiv* abs/1910.10045 (2020) (cited on pp. 8, 9).

[5] K. Balog, Filip Radlinski, and Shushan Arakelyan. "Transparent, Scrutable and Explainable User Models for Personalized Recommendation". In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2019) (cited on p. 10).

[6] V. Belle and I. Papantonis. "Principles and Practice of Explainable Machine Learning". In: *ArXiv* abs/2009.11698 (2020) (cited on p. 9).

[7] R. Bellman. "An Introduction to Artificial Intelligence: Can Computers Think?" In: 1978 (cited on p. 7).

[8] Umang Bhatt, Alice Xiang, S. Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and P. Eckersley. "Explainable machine learning in deployment". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020) (cited on p. 9).

[9] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Association for Computing Machinery, 2018, pp. 1–14. DOI: 10.1145/3173574.3173951 (cited on p. 9).

[10] Or Biran and Courtenay Cotton. “Explanation and Justification in Machine Learning: A Survey”. In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. 2017, p. 1 (cited on p. 8).

[11] Harry Brignull. Dark Patterns. darkpatterns.org. 2010 (cited on p. 16).

[12] Harry Brignull. Dark Patterns: User Interfaces Designed to Trick People. http://talks.ui-patterns.com/videos/dark-patterns-user-interfaces-designed-to-trick-people, accessed November 28, 2018. 2014 (cited on p. 16).

[13] Peter Brusilovsky, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, John O’Donovan, Giovanni Semeraro, and Martijn C. Willemsen. “Interfaces and Human Decision Making for Recommender Systems”. In: *Fourteenth ACM Conference on Recommender Systems*. RecSys ’20. Association for Computing Machinery, 2020, pp. 613–618. DOI: 10.1145/3383313.3411539 (cited on p. 10).

[14] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. “Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Association for Computing Machinery, 2020, pp. 454–464. DOI: 10.1145/3377325.3377498 (cited on pp. 12, 23).

[15] Adrian Bussone, S. Stumpf, and D. O’Sullivan. “The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems”. In: *2015 International Conference on Healthcare Informatics* (2015), pp. 160–169 (cited on p. 12).

[16] Loren J. Chapman. “Illusory Correlation in Observational Report”. In: *Journal of Verbal Learning and Verbal Behavior* 6.1 (1967), pp. 151–155. DOI: https://doi.org/10.1016/S0022-5371(67)80066-5 (cited on p. 7).

[17] K. Cheverst, H. Byun, D. Fitton, C. Sas, C. Kray, and N. Villar. “Exploring Issues of User Model Transparency and Proactive Behaviour in an Office Environment Control System”. In: *User Modeling and User-Adapted Interaction* 15 (2005), pp. 235–273 (cited on p. 10).

[18] A Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”. In: *Big data* 5 2 (2017), pp. 153–163 (cited on p. 1).

[19] Dennis Collaris, Leo M. Vink, and Jarke J. van Wijk. “Instance-Level Explanations for Fraud Detection: A Case Study”. In: *ArXiv* abs/1806.07129 (2018) (cited on p. 15).

[20] Devleena Das and Sonia Chernova. “Leveraging Rationales to Improve Human Task Performance”. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI ’20. Association for Computing Machinery, 2020, pp. 510–518. DOI: 10.1145/3377325.3377512 (cited on p. 21).

[21] Xolani Dastile, Turgay Celik, and Moshe Potsane. “Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey”. In: *Applied Soft Computing* 91 (2020), p. 106263. DOI: https://doi.org/10.1016/j.asoc.2020.106263 (cited on p. 1).

[22] Amit Datta, Michael Carl Tschantz, and Anupam Datta. “Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination”. In: *Proceedings on privacy enhancing technologies* 2015.1 (2015), pp. 92–112 (cited on p. 1).

[23] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel K E Bellamy, and Casey Dugan. “Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Association for Computing Machinery, 2019, pp. 275–285. DOI: 10.1145/3301275.3302310 (cited on p. 9).

[24] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretability”. In: *CoRR* abs/1702.08608 (2017). arXiv: 1702.08608 (cited on pp. 8, 12, 18, 23).

[25] F. K. Došilović, M. Brčić, and N. Hlupić. “Explainable Artificial Intelligence: A Survey”. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 2018, pp. 0210–0215. DOI: 10.23919/MIPRO.2018.8400040 (cited on pp. 12, 18).

[26] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. “UX Design Innovation: Challenges for Working with Machine Learning As a Design Material”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI ’17. ACM, 2017, pp. 278–288. DOI: 10.1145/3025453.3025739 (cited on p. 16).

[27] John J Dudley and Per Ola Kristensson. “A Review of User Interface Design for Interactive Machine Learning”. In: *ACM Trans. Interact. Intell. Syst.* 8.2 (2018). DOI: 10.1145/3185517 (cited on p. 10).

[28] Jerrefey Dustin. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. 2018 (cited on p. 1).

[29] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. “The Role of Trust in Automation Reliance”. In: *Int. J. Hum. Comput. Stud.* 58 (2003), pp. 697–718 (cited on p. 20).

[30] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. “Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI ’19. Association for Computing Machinery, 2019, pp. 263–274. DOI: 10.1145/3301275.3302316 (cited on p. 21).

[31] Malin Eiband, H. Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and H. Hußmann. “Bringing Transparency Design into Practice”. In: *23rd International Conference on Intelligent User Interfaces* (2018) (cited on pp. 3, 11, 12).

[32] EU High-Level Expert Group on Artificial Intelligence. Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Tech. rep. 2020 (cited on p. 2).

[33] Anthony W. Flores, K. Bechtel, and Christopher T. Lowenkamp. “False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There’s Software Used across the Country to Predict Future Criminals. and It’s Biased against Blacks"”. In: *Federal Probation* 80 (2016), p. 38 (cited on p. 1).

[34] Saul I. Gass. “Model World: The Great Debate-MAUT Versus AHP”. In: *Interfaces* 35.4 (2005), pp. 308–312. DOI: 10.1287/inte.1050.0152 (cited on p. 20).

[35] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut Learning in Deep Neural Networks. 2020 (cited on pp. 1, 2).

[36] Joey F. George, Kevin Duffy, and Manju K. Ahuja. “Countering the anchoring and adjustment bias with decision support systems”. In: *Decis. Support Syst.* 29 (2000), pp. 195–206 (cited on p. 12).

[37] N. Gill, P. Hall, Kim Montgomery, and N. Schmidt. “A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing”. In: *Inf.* 11 (2020), p. 137 (cited on p. 8).

[38] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d’Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux. “Human-Centred Machine Learning”. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’16. ACM, 2016, pp. 3558–3565. DOI: 10.1145/2851581.2856492 (cited on p. 16).

[39] Bryce Goodman and S Flaxman. “European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"”. In: *AI Mag.* 38 (2017), pp. 50–57 (cited on p. 2).

[40] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. “The Dark (Patterns) Side of UX Design”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. ACM, 2018, 534:1–534:14. DOI: 10.1145/3173574.3174108 (cited on p. 16).

[41] H. Grice. “Logic and Conversation”. In: *Syntax and Semantics* 3 (1975), pp. 41–58 (cited on p. 6).

[42] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. “A Survey of Methods for Explaining Black Box Models”. In: *ACM Computing Surveys* 51.5 (2018). DOI: 10.1145/3236009 (cited on pp. 8, 9).

[43] David Gunning. “DARPA’s Explainable Artificial Intelligence (XAI) Program”. In: Association for Computing Machinery (ACM), 2019. DOI: 10.1145/3301275.3308446 (cited on pp. 8, 13).

[44] Kristian Bondo Hansen. “The Virtue of Simplicity: On Machine Learning Models in Algorithmic Trading”. In: *Big Data & Society* 7.1 (2020). DOI: 10.1177/2053951720926558 (cited on p. 1).

[45] Wim Hardyns and Anneleen Rummens. “Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges”. In: *European Journal on Criminal Policy and Research* 24 (2018), pp. 201–218 (cited on p. 1).

[46] G. Harman. “The Inference to the Best Explanation”. In: *The Philosophical Review* 74 (1965), p. 88 (cited on p. 7).

[47] Bernease Herman. The Promise and Peril of Human Evaluation for Model Interpretability. 2019. arXiv: 1711.07414 [cs.AI] (cited on p. 23).

[48] D. Hilton. “Conversational Processes and Causal Explanation.” In: *Psychological Bulletin* 107 (1990), pp. 65–81 (cited on p. 6).

[49] Michael Hind, Dennis Wei, Murray Campbell, Noel C. F. Codella, Amit Dhurandhar, Aleksandra Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. “TED: Teaching AI to Explain Its Decisions”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’19. Association for Computing Machinery, 2019, pp. 123–129. DOI: 10.1145/3306618.3314273 (cited on p. 9).

[50] Robert R. Hoffman, William J. Clancey, and Shane T. Mueller. Explaining AI as an Exploratory Process: The Peircean Abduction Model. 2020. arXiv: 2009.14795 [cs.AI] (cited on p. 23).

[51] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects. 2019. arXiv: 1812.04608 [cs.AI] (cited on p. 23).

[52] Kasper Hornbaek and Antti Oulasvirta. "What Is Interaction?" In: (2017). DOI: 10.1145/3025453.3025765 (cited on p. 19).

[53] Hsieh-Hong Huang, J. Hsu, and Cheng-Yuan Ku. "Understanding the Role of Computer-mediated Counter-Argument in Countering Confirmation Bias". In: *Decis. Support Syst.* 53 (2012), pp. 438–447 (cited on p. 12).

[54] Introducing AI Explainability 360. https://www.ibm.com/blogs/research/2019/08/ai-explainability-360/. 2019 (cited on p. 2).

[55] Anthony Jameson and John Riedl. "Introduction to the Transactions on Interactive Intelligent Systems". In: *ACM Trans. Interact. Intell. Syst.* 1.1 (2011). DOI: 10.1145/2030365.2030366 (cited on p. 7).

[56] S. Jesus, Catarina Bel'em, Vladimir Balayan, J. Bento, Pedro Saleiro, P. Bizarro, and João Gama. "How can I Choose An Explainer?: An Application-grounded Evaluation of Post-hoc Explanations". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021) (cited on p. 23).

[57] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Association for Computing Machinery, 2020, pp. 1–14. DOI: 10.1145/3313831.3376219 (cited on pp. 15, 21).

[58] F. Keil. "Explanation and Understanding." In: *Annual review of psychology* 57 (2006), pp. 227–54 (cited on pp. 5, 6, 16).

[59] Been Kim, O. Koyejo, and Rajiv Khanna. "Examples are not Enough, Learn to Criticize! Criticism for Interpretability". In: *NIPS*. 2016 (cited on p. 8).

[60] Tae-Nyun Kim and Hayeon Song. "The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020) (cited on p. 12).

[61] René F Kizilcec and Hansol Lee. "Algorithmic Fairness in Education". In: (2020) (cited on p. 1).

[62] Bart P. Knijnenburg, Svetlin Bostandjiev, J. O'Donovan, and A. Kobsa. "Inspectability and Control in Social Recommenders". In: *RecSys '12*. 2012 (cited on p. 10).

[63] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. "Principles of Explanatory Debugging to Personalize Interactive Machine Learning". In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. IUI '15. Association for Computing Machinery, 2015, pp. 126–137. DOI: 10.1145/2678025.2701399 (cited on pp. 8, 10).

[64] Mike Kuniavsky, Elizabeth Churchill, and Molly Wright Steenson. The 2017 AAAI Spring Symposium Series Technical Reports: Designing the User Experience of Machine Learning Systems. Tech. rep. Technical Report SS-17-04. Palo Alto, California, 2017 (cited on p. 16).

[65] A. Lau and E. Coiera. "Research Paper: Can Cognitive Biases during Consumer Health Information Searches Be Reduced to Improve Decision Making?" In: *Journal of the American Medical Informatics Association : JAMIA* 16 1 (2009), pp. 54–65 (cited on p. 12).

[66] Rebecca Lawson. "The Science of Cycology: Failures to Understand how Everyday Objects Work". In: *Memory and Cognition* 34 (2006), pp. 1667–1675 (cited on p. 7).

[67] M. Lent, W. Fisher, and M. Mancuso. "An Explainable Artificial Intelligence System for Small-unit Tactical Behavior". In: *AAAI*. 2004 (cited on p. 8).

[68] D.K. Lewis. Counterfactuals. Harvard University Press, 1973 (cited on p. 6).

[69] Q Vera Liao, Daniel Gruen, and Sarah Miller. "Questioning the AI: Informing Design Practices for Explainable AI User Experiences". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2020, pp. 1–15 (cited on p. 23).

[70] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. "Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Association for Computing Machinery, 2009, pp. 2119–2128. DOI: 10.1145/1518701.1519023 (cited on p. 20).

[71] Zachary Chase Lipton. "The Mythos of Model Interpretability". In: *Queue* 16 (2018), pp. 31–57 (cited on p. 8).

[72] Scott M. Lundberg, G. Erion, Hugh Chen, Alex J. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and Su-In Lee. "From Local Explanations to Global Understanding with Explainable AI for Trees". In: *Nature Machine Intelligence* 2 (2020), pp. 56–67 (cited on p. 9).

[73] J. McGuirl and N. Sarter. "Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information". In: *Human Factors: The Journal of Human Factors and Ergonomics Society* 48 (2006), pp. 656–665 (cited on p. 12).

[74] Christian Meske, Enrico Bunde, J. Schneider, and Martin Gersch. "Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities". In: *Information Systems Management* (2020), pp. 1–11 (cited on p. 9).

[75] D. Miller. "The Norm of Self-Interest." In: *The American psychologist* 54 12 (1999), pp. 1053–60 (cited on pp. 6, 16).

[76] Tim Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. 2019. DOI: 10.1016/j.artint.2018.07.007 (cited on pp. 3, 5, 6, 8, 11, 12, 18).

[77] Tim Miller, Piers Howe, and Liz Sonenberg. "Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences". In: *CoRR* abs/1712.00547 (2017). arXiv: 1712.00547 (cited on pp. 3, 6, 12, 18).

[78] Candice M. Mills and Frank C. Keil. "Knowing the Limits of One's Understanding: The Development of an Awareness of an Illusion of Explanatory Depth." In: *Journal of experimental child psychology* 87 1 (2004), pp. 1–32 (cited on p. 7).

[79] T.M. Mitchell. Machine Learning. McGraw-Hill International Editions. McGraw-Hill, 1997 (cited on p. 7).

[80] V Mnih, K Kavukcuoglu, D Silver, Andrei A Rusu, J Veness, Marc G Bellemare, A Graves, Martin A Riedmiller, Andreas K Fidjeland, Georg Ostrovski, S Petersen, C Beattie, A Sadik, Ioannis Antonoglou, H King, D Kumaran, Daan Wierstra, S Legg, and Demis Hassabis. "Human-level Control Through Deep Reinforcement Learning". In: *Nature* 518 (2015), pp. 529–533 (cited on p. 1).

[81] Johanna D Moore and Cécile Paris. "Requirements for an Expert System Explanation Facility". In: *Computational Intelligence* 7 (1991) (cited on pp. 3, 13, 20, 21).

[82] Shane T. Mueller, Robert R. Hoffman, William J. Clancey, Abigail Emrey, and Gary Klein. "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI". In: *CoRR* abs/1902.01876 (2019). arXiv: 1902.01876 (cited on p. 15).

[83] Shane T. Mueller, Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. Principles of Explanation in Human-AI Systems. 2021. arXiv: 2102.04972 [cs.AI] (cited on p. 13).

[84] Bonnie M Muir. "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems". In: *Ergonomics* 37.11 (1994), pp. 1905–1922 (cited on p. 20).

[85] G. Murphy and D. Medin. "The Role of Theories in Conceptual Coherence." In: *Psychological review* 92 3 (1985), pp. 289–316 (cited on p. 6).

[86] K Murphy. "Machine Learning - A Probabilistic Perspective". In: *Adaptive computation and machine learning series*. 2012 (cited on pp. 2, 7).

[87] Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. "Automated Classification of Skin Lesions: From Pixels to Practice". In: *Journal of Investigative Dermatology* 138.10 (2018), pp. 2108–2110. DOI: https://doi.org/10.1016/j.jid.2018.06.175 (cited on p. 1).

[88] National Institute of Standards and Technology. Four Principles of Explainable Artificial Intelligence. Tech. rep. 2020 (cited on p. 2).

[89] An-phi Nguyen and M. R. Martínez. "MonoNet: Towards Interpretable Models by Learning Monotonic Features". In: *ArXiv* abs/1909.13611 (2019) (cited on p. 8).

[90] Don Norman. The Design of Everyday Things: Revised and Expanded Edition. Basic books, 2013 (cited on p. 6).

[91] M. Nourani, Donald R. Honeycutt, Jeremy E. Block, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, and V. Gogate. "Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020) (cited on p. 12).

[92] Antti Oulasvirta and Kasper Hornbæk. "HCI Research as Problem-Solving". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM Press, 2016. DOI: 10.1145/2858036.2858283 (cited on pp. 2, 12, 13).

[93] Andrés Páez. "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)". In: *Minds Mach.* 29.3 (2019), pp. 441–459. DOI: 10.1007/s11023-019-09502-w (cited on pp. 3, 12).

[94] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. “Manipulating and Measuring Model Interpretability”. In: *CoRR* abs/1802.07810 (2018). arXiv: 1802.07810 (cited on p. 18).

[95] I. Rahwan, Manuel Cebrian, Nick Obradovich, J. Bongard, J. Bonnefon, C. Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, M. Jackson, N. Jennings, Ece Kamar, Isabel M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. Parkes, Alex ‘Sandy’ Pentland, Margaret E. Roberts, A. Shariff, J. Tenenbaum, and Michael P. Wellman. “Machine Behaviour”. In: *Nature* 568 (2019), pp. 477–486 (cited on p. 10).

[96] E. Raymond. “The Cathedral and the Bazaar - Musings on Linux and Open Source by an Accidental Revolutionary”. In: 1999 (cited on p. 24).

[97] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. Association for Computing Machinery, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778 (cited on pp. 2, 8).

[98] L Righetti, R Madhavan, and R Chatila. “Unintended Consequences of Biased Robotic and Artificial Intelligence Systems [Ethical, Legal, and Societal Issues]”. In: *IEEE Robotics Automation Magazine* 26.3 (2019), pp. 11–13. DOI: 10.1109/MRA.2019.2926996 (cited on p. 1).

[99] F Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.” In: *Psychological review* 65 6 (1958), pp. 386–408 (cited on p. 1).

[100] A Ross, M Hughes, and Finale Doshi-Velez. “Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations”. In: *IJCAI*. 2017 (cited on p. 2).

[101] Leonid Rozenblit and Frank C. Keil. “The misunderstood limits of folk science: an illusion of explanatory depth”. In: *Cognitive science* 26 5 (2002), pp. 521–562 (cited on pp. 7, 15).

[102] C Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1 (2018), pp. 206–215 (cited on p. 1).

[103] Cynthia Rudin. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215 (cited on p. 8).

[104] Thomas L. Saaty. Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World. New 3rd ed. Pittsburgh, Pa., RWS Publications, 2001 (cited on p. 20).

[105] A L Samuel. “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM J. Res. Dev.* 3 (1959), pp. 210–229 (cited on p. 1).

[106] James Schaffer, J. O’Donovan, James Michaelis, A. Raglin, and Tobias Höllerer. “I Can Do Better Than Your AI: Expertise and Explanations”. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019) (cited on pp. 12, 23).

[107] Ben Shneiderman. “Bridging the Gap Between Ethics and Practice”. In: *ACM Transactions on Interactive Intelligent Systems* 10.4 (2020), pp. 1–31. DOI: 10.1145/3419764 (cited on p. 11).

[108] D Silver, Aja Huang, Chris J Maddison, A Guez, L Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, S Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, T Lillicrap, M Leach, K Kavukcuoglu, T Graepel, and Demis Hassabis. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529 (2016), pp. 484–489 (cited on p. 1).

[109] L. Skitka, K. Mosier, M. Burdick, and B. Rosenblatt. "Automation Bias and Errors: Are Crews Better Than Individuals?" In: *The International Journal of Aviation Psychology* 10 (2000), pp. 85–97 (cited on p. 12).

[110] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. "No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Association for Computing Machinery, 2020, pp. 1–13. DOI: 10.1145/3313831.3376624 (cited on p. 10).

[111] Kacper Sokol and Peter A. Flach. "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020) (cited on pp. 8, 15).

[112] Jacob Solomon. "Customization Bias in Decision Support Systems". In: *CHI '14*. 2014 (cited on p. 12).

[113] Mina Son, Hyeonju Lee, and Hyejung Chang. "Artificial Intelligence-Based Business Communication: Application for Recruitment and Selection". In: (2019) (cited on p. 1).

[114] The European Parliament and Council. "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)". In: *Official Journal of the European Union* (2016) (cited on p. 16).

[115] N. Tintarev and J. Masthoff. "Explaining Recommendations: Design and Evaluation". In: *Recommender Systems Handbook*. 2015 (cited on p. 10).

[116] Nava Tintarev. "Explanations of Recommendations". In: *Proceedings of the 2007 ACM Conference on Recommender Systems*. RecSys '07. Association for Computing Machinery, 2007, pp. 203–206. DOI: 10.1145/1297231.1297275 (cited on p. 18).

[117] Richard Tomsett, Dave Braines, D. Harborne, A. Preece, and S. Chakraborty. "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems". In: *ArXiv* abs/1806.07552 (2018) (cited on pp. 8, 9).

[118] Chun-Hua Tsai and Peter Brusilovsky. "Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance". In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. UMAP '19. Association for Computing Machinery, 2019, pp. 22–30. DOI: 10.1145/3320435.3320465 (cited on p. 18).

[119] Michelle Vaccaro and Jim Waldo. "The Effects of Mixing Machine Learning and Human Judgment". In: *Communications of the ACM* 62 (2019), pp. 104–110 (cited on p. 12).

[120] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. "Designing Theory-Driven User-Centric Explainable AI". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Association for Computing Machinery, 2019, pp. 1–15. DOI: 10.1145/3290605.3300831 (cited on pp. 10, 12).

[121] Adrian Weller. “Challenges for Transparency”. In: *ArXiv* abs/1708.01870 (2017) (cited on pp. 9, 16).

[122] J. Wobbrock and J. Kientz. “Research Contributions in Human-Computer Interaction”. In: *Interactions* 23 (2016), pp. 38–44 (cited on p. 22).

[123] Wei Xu. “Toward Human-Centered AI: A Perspective from Human-Computer Interaction”. In: *Interactions* 26.4 (2019), pp. 42–46. DOI: 10.1145/3328485 (cited on p. 2).

[124] Fan Yang, Mengnan Du, and Xia Hu. “Evaluating Explanation Without Ground Truth in Interpretable Machine Learning”. In: *CoRR* abs/1907.06831 (2019). arXiv: 1907.06831 (cited on p. 8).

[125] Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman. “Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020) (cited on p. 16).

[126] Qian Yang, Nikola Banovic, and John Zimmerman. “Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI ’18. ACM, 2018, 130:1–130:11. DOI: 10.1145/3173574.3173704 (cited on p. 16).

[127] Ming Yin, Jennifer Wortman Vaughan, and H. Wallach. “Understanding the Effect of Accuracy on Trust in Machine Learning Models”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019) (cited on p. 12).

APPENDIX: ORIGINAL PUBLICATIONS

This appendix includes all contributing publications of this thesis in their original format without any modifications (except page numeration) in chronological order.

Making SHAP Rap: Bridging Local and Global Insights through Interaction and Narratives

Michael Chromik

LMU Munich, Munich, Germany
michael.chromik@ifi.lmu.de

Abstract. The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation-generating methods. In practice, Shapley explanations are widely used. However, they are often presented as visualizations and thus leave their interpretation to the user. As such, even ML experts have difficulties interpreting them appropriately. On the other hand, combining visual cues with textual rationales has been shown to facilitate understanding and communicative effectiveness. Further, the social sciences suggest that explanations are a social and iterative process between the explainer and the explainee. Thus, interactivity should be a guiding principle in the design of explanation facilities. Therefore, we (i) briefly review prior research on interactivity and naturalness in XAI, (ii) designed and implemented the interactive explanation interface *SHAPRap* that provides local and global Shapley explanations in an accessible format, and (iii) evaluated our prototype in a formative user study with 16 participants in a loan application scenario. We believe that interactive explanation facilities that provide multiple levels of explanations offer a promising approach for empowering humans to better understand a model's behavior and its limitations on a local as well as global level. With our work, we inform designers of XAI systems about human-centric ways to tailor explanation interfaces to end users.

Keywords: explainable AI · explanation interface · interactivity.

1 Introduction

Many decisions in our lives are influenced or taken by intelligent systems that leverage machine learning (ML). Whenever their predictions may have undesired or consequential impacts, providing only the output of the black box may not be satisfying to their users. Even if the prediction is accurate in regard to the underlying training data, users may distrust the system, have different beliefs regarding the prediction, or want to learn from individual predictions about a given problem domain. Thus, a need for understanding the ML model behavior arises [2]. The field of *explainable artificial intelligence (XAI)* develops novel methods and techniques to make black-box ML models more interpretable. Current XAI research mostly focuses on the *cognitive* process of explanation, i.e.,

identifying likely root causes of a particular event [21]. As a result, some notion of explanation is generated that approximates the model's underlying prediction process. Explanations may be textual, visual, example-based, or obtained by simplifying the underlying prediction model [3]. An approach widely used in practice is *explanation by feature attribution* [3]. Especially local explanations based on *Shapley values* [27] are widespread [4]. Feature attribution frameworks, such as SHAP[1], merely provide visual explanations and leave their interpretation entirely to the user. As such, they are targeting mostly ML experts, such as developers and data scientists. However, Kaur et al. [17] observed in their studies that even experts have an inaccurate understanding of how to interpret the visualizations provided by SHAP. Even if they are correctly interpreted by ML experts, they may still remain opaque to end users of XAI due to their technical illiteracy [6]. This applies especially to end users and subject-matter experts, who often have little technical expertise in ML. Thus, their interpretability needs require even more guidance and attention.

The main idea of this paper is to explore how to improve the accessibility of Shapley explanations to foster a pragmatic understanding [23, 11] for end users in XAI. We believe that an important aspect required to address the call for *"usable, practical and effective transparency that works for and benefits people"* [1] is currently not sufficiently studied: providing end users of XAI with means of interaction that go beyond a single static explanation and that are complemented by explicit interpretations in natural language. As the human use of computing is the subject of inquiry in HCI [22], our discipline *"should take a leading role by providing explainable and comprehensible AI, and useful and usable AI"* [34]. In particular, our community is well suited to *"provide effective design for explanation UIs"* [34]. Our work contributes to the HCI community in two ways: First, we present and describe the interactive explanation interface artifact *SHAPRap* that targets non-technical users of XAI. Second, we report promising results from a formative evaluation that indicates that our approach can foster understanding. With this work, we put our design rationales up for discussion with our fellow researchers.

2 Related Work

We base our work in the interdisciplinary research field of XAI. It aims to make black-box ML models interpretable by generating some notion of explanation that can be used by humans to interpret the behavior of an ML model [31]. An ML model is considered a black-box if humans can observe the inputs and outputs of the model but have difficulties understanding the mapping between them. However, most works focus on computational aspects of generating explanations while limited research is reported concerning the human-centered design of the explanation interface. The social sciences suggest that the explanation process should resemble a social process between the explaining XAI system (sender of an explanation) and the human explainee (receiver of an explanation)

[1] github.com/slundberg/shap

forming a multi-step interaction between both parties, ideally leveraging natural language [21]. Especially, in situations where people may be held accountable for a prediction-informed decision, they may have multiple follow-up questions before feeling comfortable to trust a system prediction. Abdul et al. emphasize that interactivity and learnability are crucial for the effective design of explanations and their visualization [1]. Widely used explainability frameworks, such as SHAP, present their explanations in the form of information-dense visualizations, however, they do not provide any interactivity nor guidance to support users in their interpretation process. As a consequence, even experienced ML engineers struggle to correctly interpret their output and often take them at face value [17]. Humans mostly explain their decisions with words [19]. Thus, it is intuitive to provide end users of XAI with explanations in natural language. We found first work that takes a human-centric perspective on XAI and encompasses interactivity and naturalness. Weld and Bansal [32] propose seven different follow-up and drill-down operations to guide the interaction. Liao et al. [18] compile a catalog of natural language questions that can technically be answered by current XAI methods. Covering multiple of them under a *"holistic approach"* allows users to triangulate insights. Reiter [24] discusses the challenges of natural language generation for XAI. Further, users have been shown to understand technical explanations better if they are complemented by narratives in natural language [9, 10, 13]. For instance, Gkatzia et al. improved users' decision-making by 44% by combining visualizations with statements in natural language [13]. Sokol and Flach [29] present *Glass-Box* an interactive XAI system that provides personalized explanations in natural language. Similarly, Werner [33] presents *ERIC* an interactive system that gives explanations in a conversational manner through a chat-bot like interface. Forrest et al. [12] generate textual explanations from feature contributions based on LIME [25].

3 SHAPRap

3.1 Scenario, ML Model, and XAI Method

Scenario. Our XAI system is centered in a decision-support situation in which the human decision-maker is accompanied by an intelligent and interpretable system. We put our study participants in the shoes of a private lender on a fictional crowd lending platform. We centered our study in a crowd lending domain because we assumed that the participants can relate to decisions about lending or investing personal money. Participants can see demographic information, loan details, and credit history of individuals that request a loan on the platform. Each request is accompanied by an "AI-based intelligent prediction" of the *default risk*, i.e., the probability that the borrower fails to service a loan installment some time during the loan period. The prediction is introduced as an "AI-based" feature that is based on machine learning from historic cases. We build on a tabular data set as many ML models deployed in practice build on this type of data [4, 20]. We used the *Loan Prediction*[2] data set which consists

[2] datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/

of 614 loan requests with 13 columns. We relabeled two columns of the data set to be consistent with our scenario[3].

ML Model. We calculated the default risk prediction using a *XGBoost classifier*. Tree-based ensembles, such as XGBoost, are widely used in many real-world contexts because of their practicability [20]. However, they are considered black-box ML models. To limit the cognitive load for participants we chose to train our model on a subset of columns. We used only the seven categorical columns (5 binary, 1 ternary, and 1 with four possible values). We trained a binary XGB classifier with 100 decision trees and class probabilities as outputs. Other than that, we used the default hyperparameters of the *xgboost* package. The accuracy of the predicted default risk on our stratified validation set was 0.83.

XAI Method. In this work, we use the *SHAP (SHapley Additive exPlanations)* [20] framework to compute the model's feature contributions on a local and global level. SHAP belongs to the class of *additive feature attribution methods* where the explanation is represented as a linear function of feature contributions towards an ML prediction. The contributions are approximated by slightly changing the inputs and testing the impact on the model outputs. The framework unifies the ideas of other feature attribution methods (such as LIME [25]) with *Shapley values*, which originate from game theory [27]. Shapley explanations quantify the contribution of individual features values towards a prediction. For a single observation, they uniquely distribute the difference between the average prediction and the actual prediction between its features [20]. For example, if the average prediction over all instances in a dataset is 50% and the actual prediction for a single instance is 75%, SHAP uniquely distributes the difference of 25 percentage points across the features that contributed to the instance's prediction. Despite their vulnerability to adversarial attacks [28] and potential inaccuracies [14], we consider Shapley explanations as relevant to end users for two reasons: (i) they can yield local and global insights because Shapley values are the atomic units of each explanation. As these units are additive, they may be aggregated over multiple predictions or features to learn about the model's global behavior, and (ii) the consistent and model-agnostic nature of Shapley values allows XAI designers to offer a uniform explanation interface to users even if the underlying data or ML model changes.

3.2 Explanation Interface

Local Explanation View. The local explanation view resembles a spreadsheet-like user interface that is overlaid with a heat map of Shapley values for each feature of an instance. We support users' rapid visual estimation of feature contributions through preattentive processing based on a cell's hue [15]. Each cell is shaded depending on their direction and magnitude of contribution towards the prediction (red increases the loan request's risk of defaulting, while green decreases it).

[3] we re-framed the *Loan_Status* column to represent the default risk and the *Credit_History* column to represent a negative item on a credit report.

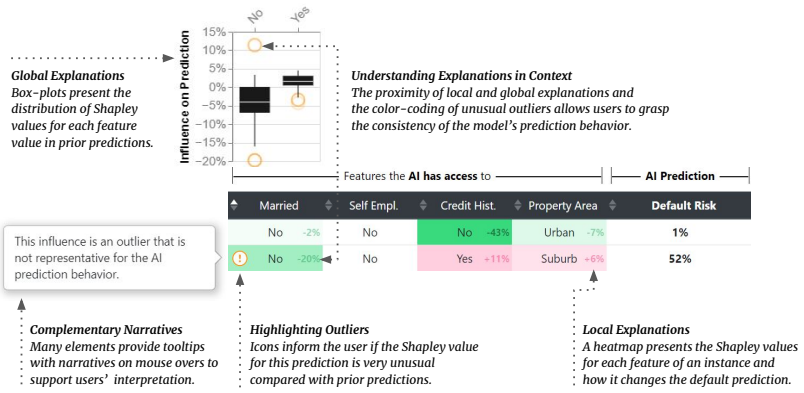


Fig. 1. The components of the *SHAPRap* explanation interface

The local explanation view is *contrastive* [21] as it allows comparing variances *between* feature contributions for individual instances (*horizontal axis*). Further, as we show multiple local explanations next to each other, users can compare variances or regularities *within* feature values across multiple instances (*vertical axis*). To support this, users can sort each column by value to contrast instances with identical feature values.

Global Explanations View. Local explanations yield how an ML model derives its prediction for a single data instance. In contrast, global explanations help users to get an intuition how a model derives its predictions over multiple instances or an entire dataset (*global sample*). For each feature value, we provide a box-plot of how it contributed to the prediction for all instances in the global sample. A narrow box-plot indicates a more consistent prediction behavior, while a wider box-plot indicates that the contributions vary for the same feature value. These variances result from interactions with other features and may require additional judgment (see next paragraph). The distribution of Shapley values in the global view depends on the chosen global sample. If the sample is representative for the population that the ML model will be confronted with in a particular domain, the global view helps users understanding when its predictions are consistent and therefore predictable and when they are not. In practice, the global sample may be the entirety of predictions of an ML model after its deployment across all users, or (if data sparsity requirements apply) a sample of predictions that an individual user has previously been exposed to. Further, it would be possible to let users customize the global sample (e.g., only instances above a certain prediction threshold or instances with a particular feature value). In our prototype, we displayed the distributions of the training and validation sets.

Highlighting Outliers. A post-hoc *explanation by feature attribution* approach, such as SHAP, is always an approximation of the actual prediction behavior of an ML model. Identifying inconsistent contributions and communicating them to the user can improve their interpretation by making it easier to identify ex-

planations that are more representative for the global model behavior. We built around the concept of role-based explanations [5]. We classify each instance's feature value contribution into the roles *normal* (within the *inter quartile range (IQR)* of the global sample), *unusual* (beyond IQR but within whiskers as defined by $\pm 1.5 \times IQR$), and *very unusual* (outliers beyond the whiskers). We highlight *very unusual* contributions in the global and local views as orange warning circles prompting the users to not generalize from these instances to the typical prediction behavior of the model. Further, these outliers may serve as starting points for analyzing feature value interactions. When hovering over an outlier, we highlight features of this instance that are *unusual* and thus provide hints which feature values may be interacting with each other.

Complementing Narratives: It is not easy to understand the concepts of additive Shapley explanations just by looking at plots [17]. It might take some time to interpret a plot, and the user is likely to be overwhelmed at first. Thus, we automatically created textual explanations from Shapley values using a template-based approach and to support their interpretation of the local and global views. We provide users with on-demand textual explanations in form of tooltips on mouseovers for each feature box-plot, instance cell, outlier highlight, and column header. Further, we provided background information about the local and global

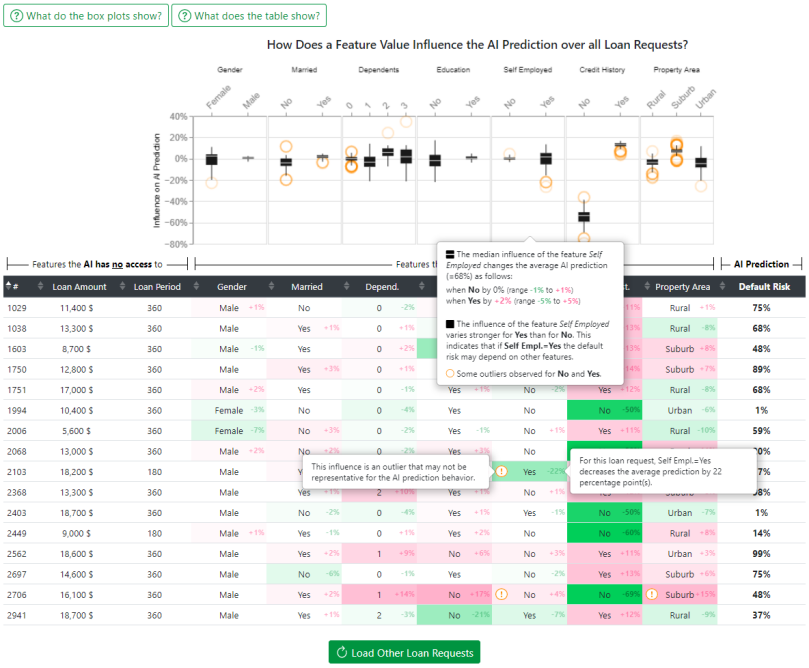


Fig. 2. The explanation interface that participants were exploring.

views during onboarding and accessible through help buttons during interaction. This way, information redundancy can be avoided following the *progressive disclosure* paradigm [30].

4 Formative Evaluation

Method. We conducted a formative evaluation with 16 participants recruited through the online platform *Prolific*. We recruited participants with at least a graduate degree, English fluency, and an approval rate of 100%. 8 participants self-identified as female, 8 as male and were in the age groups 18-24 (3), 24-35 (9), and 35-54 (4). 11 participants agreed to use spreadsheets at least weekly, 6 knew how to read box-plots, and 4 had practical experience with ML. After introducing their role in the crowd lending scenario and the explanation views, users were asked to freely explore *SHAPRap* for 10 to 15 minutes. Then, they rated their level of understanding on a 7-point scale[4] [8]. Afterwards, they completed a *forward prediction* quiz [7]. Participants had to simulate the AI prediction for 6 pre-selected loan requests with the help of the global explanation view. We randomly chose 6 instances with unique feature value combinations and at most two *unusual* contributions to assess participants' understanding of the typical prediction behavior. In the end, they rated the *explanation satisfaction scale* [16] and answered three open questions. On average, participants took 28.1 minutes (SD=10.4 minutes) to complete the study and were compensated £5 per completion (=£10.67/hour).

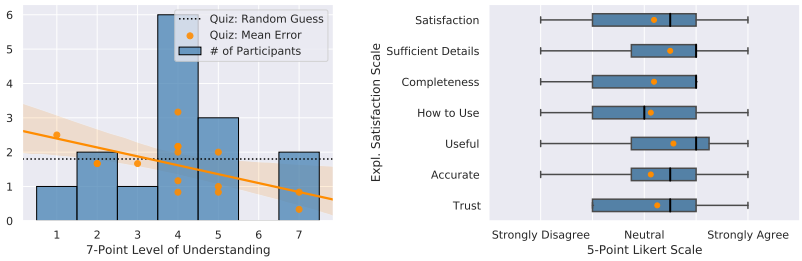


Fig. 3. *(left)* 11 participants perceived they understood at least which features were important for the prediction. 6 of them objectively proved their understanding via a lower than random mean error in a forward prediction quiz. *(right)* Results from the *explanation satisfaction scale*. The orange dots indicate the respective mean.

[4] Level 1: *I understand which features the AI has access to and what the AI predicts as an output.*, Level 4: *I understand which features are more important than others for the AI prediction.*, Level 7: *I understand how much individual feature values influence the AI prediction and which feature values depend on others.*

Results. Overall, our results indicate mixed reactions but show effective gains of pragmatic understanding for some participants. The explanation facility felt overwhelming at first, but the complementary elements of global, local, and textual explanations were considered as somewhat useful and sufficiently detailed to get a general idea about the typical prediction behavior. After exploring *SHAPRap*, participants on average rated their understanding as *"I understand which features are more important than others for the AI prediction"* (mean=4.07, SD=1.67). However, applying this understanding in the quiz turn out to be challenging for 6 participants as they scored worse than random guess (expected error for a random guess was 1.8). For example, P5 *"understood what the box representations meant but found it hard to actually apply this data to the applicants. It might just require practice."* On a positive end, 6 participants rated their gained understanding as at least level 4 and proved this with low mean errors in the quiz (cf. Fig. 3). Participant P6 (no ML experience, mean error of 0.8) *"found the explanations quite complicated to follow but after studying the table and explanations it became clearer as to which factors were being used to measure the likelihood of defaulting on the loan."* P3 (extensive ML experience, mean error of 0.33) found *"the explanations were detailed, and it was interesting to see that credit history was the leading variable for default risk."* Multiple participants appreciated the complementary nature of the natural language explanations. Without them *"the graph was quite difficult to understand on its own"* (P6). P13 liked *"that the [textual] explanations are written simply, everyone would understand it"* and P9 appreciated that the *"language was simple"*. However, it seemed that narratives on a more aggregated or abstract level were missing to understand the bigger picture. P4 found *"this kind of explanations useful just to people who already have studied this but for people with different educational background this kind of explanations are not enough."* P5 suggested adding an executive summary for each loan request and the overall global view. Further, some participants were overwhelmed by the non-linear behavior and interactions of the ML model and seemed to expect to figure them out. P5 found *"the green and red increase/decrease for risk seemed simple and helpful at first, but there seemed to be very random correlations between different aspects."* Similarly, P10 stated: *"I am guessing there are so many intersecting correlations it's hard to read for a non-numbers person."*. This resonates with Rudin [26] that the term *explanation* is misleading as it suggests a full understanding can be reached even if we merely provide pragmatic approximations.

5 Summary

This paper presents the explanation interface *SHAPRap*, which supports end users in interpreting local Shapley explanations in the global context of *normal* and *unusual* model behavior. Further, it provides narratives using a template-based approach. With our work, we contribute to the development of accessible XAI interfaces that enable non-expert users to get an intuition about the probabilistic decision behavior of black-box ML models.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. CHI '18 (2018). https://doi.org/10.1145/3173574.3174156
2. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access (2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion (2020). https://doi.org/10.1016/j.inffus.2019.12.012
4. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., Eckersley, P.: Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (2020). https://doi.org/10.1145/3351095.3375624
5. Biran, O., McKeown, K.: Human-centric justification of machine learning predictions. IJCAI '17 (2017). https://doi.org/10.24963/ijcai.2017/202
6. Burrell, J.: How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society (2016). https://doi.org/10.1177/2053951715622512
7. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H.: Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. CHI '19 (2019). https://doi.org/10.1145/3290605.3300789
8. Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A.: I think i get your point, ai! the illusion of explanatory depth in explainable ai. IUI '21 (2021). https://doi.org/10.1145/3397481.3450644
9. Das, D., Chernova, S.: Leveraging Rationales to Improve Human Task Performance. IUI '20 (2020). https://doi.org/10.1145/3290605.3300789
10. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., Riedl, M.O.: Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. IUI '19 (2019). https://doi.org/10.1145/3301275.3302316
11. Eiband, M., Schneider, H., Buschek, D.: Normative vs. pragmatic: Two perspectives on the design of explanations in intelligent systems. In: IUI Workshops (2018)
12. Forrest, J., Sripada, S., Pang, W., Coghill, G.: Towards making nlg a voice for interpretable machine learning. In: INLG (2018). https://doi.org/10.18653/v1/W18-6522
13. Gkatzia, D., Lemon, O., Rieser, V.: Natural language generation enhances human decision-making with uncertain information. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/P16-2043
14. Gosiewska, A., Biecek, P.: Do not trust additive explanations. ArXiv (2020), https://arxiv.org/abs/1903.11420
15. Healey, C.G., Booth, K.S., Enns, J.T.: High-speed visual estimation using preattentive processing. ACM Trans. Comput.-Hum. Interact. (1996). https://doi.org/10.1145/230562.230563
16. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. CoRR (2018), https://arxiv.org/abs/1812.04608

17. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. CHI '20 (2020). https://doi.org/10.1145/3313831.3376219
18. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI '20 (2020). https://doi.org/10.1145/3313831.3376590
19. Lipton, Z.C.: The mythos of model interpretability. ACM Queue (2016). https://doi.org/10.1145/3236386.3241340
20. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A.J., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence (2020). https://doi.org/10.1038/s42256-019-0138-9
21. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2019). https://doi.org/10.1016/j.artint.2018.07.007
22. Oulasvirta, A., Hornbaek, K.: HCI Research as Problem-Solving. CHI '16 (2016). https://doi.org/10.1145/2858036.2858283
23. Páez, A.: The Pragmatic Turn in Explainable Artificial Intelligence (XAI). Minds and Machines (2019). https://doi.org/10.1007/s11023-019-09502-w
24. Reiter, E.: Natural language generation challenges for explainable AI. In: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019) (2019). https://doi.org/10.18653/v1/W19-8402
25. Ribeiro, M.T., Singh, S., Guestrin, C.: "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 (2016). https://doi.org/10.1145/2939672.2939778
26. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence (2019). https://doi.org/10.1038/S42256-019-0048-X
27. Shapley, L.S.: A value for n-person games. Contributions to the Theory of Games (1953)
28. Slack, D., Hilgard, S., Jia, E., Singh, S., Lakkaraju, H.: Fooling lime and shap: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2020). https://doi.org/10.1145/3375627.3375830
29. Sokol, K., Flach, P.A.: One explanation does not fit all. KI - Künstliche Intelligenz (2020). https://doi.org/10.1007/s13218-020-00637-y
30. Springer, A., Whittaker, S.: Progressive Disclosure. ACM Transactions on Interactive Intelligent Systems (2020). https://doi.org/10.1145/3374218
31. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. CHI '19 (2019). https://doi.org/10.1145/3290605.3300831
32. Weld, D.S., Bansal, G.: The challenge of crafting intelligible intelligence. Communications of the ACM (2019). https://doi.org/10.1145/3282486
33. Werner, C.: Explainable ai through rule-based interactive conversation. In: EDBT/ICDT Workshops (2020), http://ceur-ws.org/Vol-2578/ETMLP3.pdf
34. Xu, W.: Toward human-centered AI: A perspective from human-computer interaction. Interactions (2019). https://doi.org/10.1145/3328485

Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces

Michael Chromik and Andreas Butz

LMU Munich, Munich, Germany
michael.chromik@ifi.lmu.de
butz@ifi.lmu.de

Abstract. The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation-generating methods. Although the social sciences suggest that explanation is a social and iterative process between an explainer and an explainee, explanation user interfaces and their user interactions have not been systematically explored in XAI research yet. Therefore, we review prior XAI research containing explanation user interfaces for ML-based intelligent systems and describe different concepts of interaction. Further, we present observed design principles for interactive explanation user interfaces. With our work, we inform designers of XAI systems about human-centric ways to tailor their explanation user interfaces to different target audiences and use cases.

Keywords: explainable AI · explanation user interfaces · interaction design · literature review.

1 Introduction

Intelligent systems based on machine learning (ML) are widespread in many contexts of our lives. Often, their accurate predictions come at the expense of interpretability due to their black-box nature. As consequential predictions of these systems may raise questions by those who are affected or held accountable, there is a call for *"explanations that enable people to understand the decisions"* [85]. Hence, much research is conducted within the emerging domain of explainable artificial intelligence (XAI) and interpretable machine learning (IML) on developing methods and interfaces that human users can interpret – often through some sort of explanation. Often there is not a single explanation to be conveyed [1]. Therefore, the DARPA XAI program describes the XAI process as a two-staged approach. It distinguishes between the explainable model and the explanation user interface [37] and, thus, disentangles analyzing the ML model behavior from communicating it to the user. We define an *explanation user interface (XUI)* as the sum of outputs of an XAI system that the user can directly interact with. An XUI may tap into the ML model or may use one or more explanation generating algorithms to provide relevant insights for

a particular audience. The design of interfaces that *"allow users to better understand underlying computational processes"* is considered a grand challenge of HCI research [86]. Shneiderman considers XUIs as a building block towards *human-centered AI* which aims *"to amplify, augment and enhance human performance"* instead of automating it [85].

However, most XAI research focuses on computational aspects of generating explanations while limited research is reported concerning the human-centered design of the XUI [89, 85, 102]. Similarly, resources targeting practitioners, such as UK's Information Commissioner's Office[1] , who aim to provide practitioners with *"guidance [that] is practically applicable in the real world"*, do not touch on explanation user interfaces nor how to present them to users and instead propose *"...to draw on the expertise of user experience and user interface designers"*. A notable exception is Google's *People+AI Guidebook*[2] which presents case studies of explanations integrated into mobile apps. As the human use of computing is the subject of inquiry in HCI [73], our discipline *"should take a leading role by providing explainable and comprehensible AI, and useful and usable AI"* [105]. In particular, our community is well suited to *"provide effective design for explanation UIs"* [105].

To follow this call and to understand the current practices in the field, we took an HCI perspective and conducted a systematic literature review. The overarching research question (ORQ) of our work is to **survey how researchers designed XUIs in prior XAI work.** From there, we analyze the user interactions offered by the XAI systems and describe observed design patterns. Our work is guided by the following more specific research questions:

- RQ1: How can the different concepts of interaction in XAI be characterized?
- RQ2: What design principles for interactive XUIs can be observed?

The increasing demand for interpretable systems also raises the question how to present this interpretability to users. The contribution of this paper is twofold: First, we provide a structured literature overview of how user interaction has been designed in XAI. Second, we outline design principles for human interaction with XUIs. Our work guides researchers and practitioners through the interdisciplinary design space of XAI from an HCI perspective.

2 Background and Related Work

2.1 Interaction in Surveys of Explainable AI

XAI is an umbrella term for algorithms and methods that extend the output of ML-based systems with some sort of explanation. The goal is *"to explain or to present [the ML-based system] in understandable terms to a human"* [27].

[1] ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/

[2] pair.withgoogle.com/chapter/explainability-trust/

Multiple reviews of the growing field of XAI exist. They formalize and ground the concept of XAI [1, 3], relate it to adjacent concepts and disciplines [1, 62], categorize methods [36, 57], analyze the user perspective [33], review evaluation practices [65], or outline future research directions [1, 3]. Most of these reviews acknowledge the importance of interaction for XAI only as a side note. For instance, Mueller et al. [65] consider an effective explanation to be *"an interaction"* and *"not a property of statements"*. Adadi et al. [3] state that *"explainability can only happen through interaction between human and machine"*. Abdul et al. [1] present research on interactive explanation interfaces as an important trajectory to advance the XAI research field. However, none of these reviews elaborates how this interaction could be described nor designed to inform researchers and practitioners. To our knowledge, none of the review look at XAI from an interaction design perspective.

On a broader level, there is a line of research on how to design the overall human interaction with AI-infused systems. For instance, Amershi et al. present guidelines for AI-infused systems [5]. While not explicitly addressing interpretability nor explanations, they point out the importance of making clear why the system did what it did in case of errors. However, their guidelines do not outline what this interaction could look like.

2.2 The XAI Pipeline and Explanation User Interfaces

The XAI process can be broken down into different steps. Murdoch et al. distinguish between the predictive accuracy, the descriptive accuracy, and the relevancy of an XAI system. *Predictive accuracy* is the degree to which the learned ML model correctly extracts the underlying data relationships. *Descriptive accuracy* (also referred to as fidelity) is the degree to which an explanation generation method accurately describes the behavior of the learned ML model. Both accuracies can be objectively measured. In contrast, the subjective *relevancy* describes if the outputs are communicated in a way that they provide insights for a particular audience into a chosen domain problem [67].

The DARPA XAI program illustrates the XAI process as a two-staged approach. It distinguishes between the explainable model and the explanation user interface [37]. The former addresses the predictive and descriptive accuracies, while the latter aims for relevancy. Such a two-staged approach disentangles the XAI process into analyzing the ML model behavior and communicating it to the user. Similarly, Danilevsky et al. [21] differentiate between explainability techniques and explainability visualizations. The former generates *"raw explanations"* typically proposed by AI researchers while the latter is concerned with the presentation of these *"raw explanations"* to users typically guided by HCI researchers. Most open-source methods for XAI provide a single explanation generation method. However, there is a growing number of explanation generation toolkits (e.g., AIX 360[3], Alibi[4], DALEX[5]) that combine multiple state-of-the-art

[3] https://aix360.mybluemix.net/
[4] https://docs.seldon.io/projects/alibi/en/latest/
[5] https://uc-r.github.io/dalex

methods in a uniform programming interface and thus enable rapid prototyping of XUI.

In this work, we define an explanation user interface (XUI) as the sum of outputs of an XAI process that the user can directly interact with. Shneiderman [85] outlines two modes of XUI. *Explanatory* XUIs aim to convey a single explanation (e.g., a visualization or a text explanation). In contrast, *exploratory* XUIs let users freely explore the ML model behavior. They are most effective when users have the power to change or influence the inputs. Arya et al. [7] distinguish between static and interactive explanations. A static explanation *"does not change in response to feedback from the consumer"*. In contrast, interactive explanations allow *"to drill down or ask for different types of explanations [...] until [...] satisfied"*.

3 Methodology

In line with our ORQ, our method for characterizing interaction in XAI was to collect a corpus of publications using the structured search approaches by Kitchenham and Charters [47]. We then analyzed the corpus regarding the interaction concepts followed by the authors as well as the design and interaction functionalities offered to users.

To collect a corpus of candidate publications, we conducted a systematic search in the *ACM Digital Library*. We limited our search to work that has been published at venues relevant to HCI (*Sponsor SIGCHI*). Through initial exploratory search, we obtained an initial understanding of relevant keywords, synonyms, and related concepts that helped us to construct the search query. Different terms are used to describe the field of XAI and XUI [1]. We focused on publications that include user-centered artefacts with explicit forms of explanation for the underlying intelligent behavior. Our primary focus was on research that builds on the potentials of current algorithmic explanation-generating XAI methods and thus often self-identifies as *"XAI"* or *"explainable AI"*. To account for the historic perspectives, we included *"explanation interface"* and *"explanation facility"*. These terms emerged in the 2000s from the recommender systems community and have often been used as a umbrella term for user interfaces covering different explanatory goals [92]. Further, we were interested in research that has a user focus and mentions some form of *"user interaction"*, *"user interface"*, or aspects of *"usability"* or *"interactive"*. We prepended the terms interaction and interface with *"user"* to distinguish them from feature interactions and system interfaces. While not covering the entire dynamic of this interdisciplinary field, this scoping resulted in a diverse set of works from multiple decades that put a focus on the user interface artefact. This resulted in the following search query:

> [[All: "xai"] OR [All: "explainable ai"] OR [All: "explanation facility"] OR [All: "explanation interface"]] AND [[All: "user interaction"] OR [All: "user interface"] OR [All: usability] OR [All: interactive]]

We conducted the search procedure in December 2020, which returned a total of 146 results. We then analyzed the full-text of all results. We excluded 13 results without a contribution (i.e., proceedings, keynotes, workshop summaries). Publications included in our analysis had to present results from *constructive* [73] research that involved an XUI artefact (n=57) or *conceptual* [73] research that addresses interaction in XAI (n=34). Consequently, we excluded 28 results that were not related to XAI and 14 results that were related to XAI but did not present an XUI nor describe interaction. The review was conducted by the first author. The second author was consulted for feedback. Our final set for analysis consisted of 91 publications. We analyzed the selected publications and coded information about the reported XUI and user interactions in a database.

4 Concepts of Interaction in XAI

Following Hornbæk and Oulasvirta [42], *interaction* describes the interplay between two or more constructs. They analyzed the interplay between the constructs human and computer that were discussed in HCI research. From this, they derived seven concepts of interaction: interaction as information transmission, as dialogue, as control, as experience, as optimal behavior, as tool use, and interaction as embodied action. More narrowly, Miller frames XAI as one kind of a human-agent interaction problem where an *"explanatory agent [is] revealing underlying causes to its or another agent's decision making"* [62]. As such, it is about the interplay between a human user and an AI agent that is mediated through an XUI. Tintarev and Masthoff [92] distinguish seven explanatory goals: transparency (answer how the system works), scrutability (allow to question and correct the system), trustworthiness (increase user confidence), persuasiveness (convince user), effectiveness (help user making good decisions), efficiency (help user making decisions faster), and satisfaction (increase usability). As these may be conflicting with one another, designers of XUI *"need to make trade-offs while choosing or designing the form of interface"* [93].

We build on the interaction concepts of Dubin and Hornbaek [42] and apply them to human-XAI interaction. To answer RQ1 (How can the different concepts of interaction in XAI be characterized?), we analyzed the primary interaction concept that authors (implicitly) applied as part of their work. In particular, we focus on the interplay between a user and an AI system that is facilitated through a UI that leverages some kind of explanation to reach an explanatory goal. We abstracted from the purpose that the researchers used the XUI for and instead looked at how a user could interact with it. As such, we approached the concepts of interaction with an *artefactist approach* [90]. Below, we introduce each concept and relate them to surveyed publications. Table 1 summarizes our analysis.

4.1 Interaction as (Information) Transmission

This concept centers around maximizing the throughput of information via a noisy channel. The interaction is about selecting the best message for transmis-

sion from a set of possible messages [42]. It follows the *Shannon-Weaver* [84] model of communication according to which the sender transmits information to the receiver but in between noise is added to the original message.

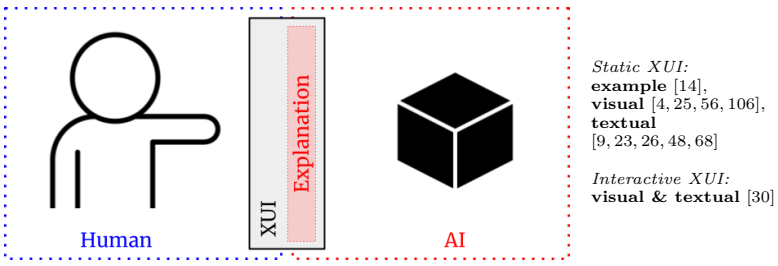


Fig. 1. XAI-interaction as (information) transmission is about presenting an accurate and complete explanation about the AI behavior.

Transfer to XAI: The goal of this interaction centers around presenting users with one complete explanation. Surveyed publications following this concept are mostly driven by the explanatory goal of transparency and acknowledge that *"algorithms should not be studied in isolation, but rather in conjunction with interfaces, since both play a significant role in the perception of explainability"* [25]. They emphasize either (i) the descriptive accuracy of an explanation to describe the underlying AI behavior [26, 30, 48, 56, 68] or (ii) the capacity of a single explanation style [4] or differences between explanation styles [9, 14, 23, 25, 106, 83] to convey information about the behavior to the human. The message is noisy because it may be difficult or even impossible to fully describe the complexity of the AI in a human understandable way, such as with deep neural networks. Unlike interaction as a dialogue, this interaction is mainly about unidirectional communication by presenting a single and static explanation. The XUI is mainly used as a medium for transmitting this explanation.

Examples: Ehsan et al. [30] present real-time explanations about the actions taken by an autonomous gaming agent in the form of natural language rationales. Alqaraawi et al. [4] study whether saliency maps convey enough information to enable users to anticipate the behavior of an image classifier. Cai et al. [14] compared how well two example-based explanation styles could promote user understanding of a sketch recognition AI. Dodge et al. [23] and Binns et al. [9] study how much different textual explanation styles convey about underlying fairness issues of an ML system. Yang et al. [106] study the differences in spatial layout and visual representation of example-based explanations.

4.2 Interaction as Dialogue

This concept describes a cycle of communication of inputs/outputs by the computer and perception/action by a human. The interaction happens in stages or

turns [42]. It tries to ensure a correct mapping between UI functions and the user's intentions and feedback by the UI to bridge the *gulf of execution* [69].

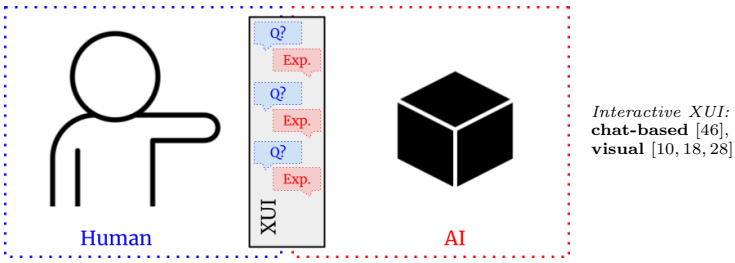


Fig. 2. XAI-interaction as dialogue is about facilitating an iterative communication cycle about the AI behavior.

Transfer to XAI: This concept acknowledges that a single explanation rarely results in a desired level of understanding [1]. Instead, it emphasizes the naturalness and accessibility of (often implicit or simplified) explanations. In contrast to interaction as embodied action, this concept is driven by the user, with the AI responding. Unlike interaction as control, this concept does not change the AI behavior. The goal of the interaction is to provide users with functionalities to gradually build a mental model of the AI behavior. We distinguish between inspection dialogues [10, 18, 28] and natural dialogues [46].

Inspection Examples: Exploratory dialogues allow the user to explore how (possibly hypothetical) changes in inputs lead to changes in the AI prediction or let the user inspect internals of the AI. The XUI is mostly about offering functionalities to iteratively request explanations of the same kind. Explanations have a high fidelity but are implicit. For instance, Cheng et al. [18] present an XUI that allows users to observe how the predictions of a university admission classifier change by freely adjusting the values of input features of applicants. Their exploratory approach was shown to improve users' comprehension although it required more of their time. Bock and Schreiber [10] present an XUI to inspect layers and parameters of deep neural networks in virtual reality. Similarly, Douglas et al. [28] visualize an AI agent's behavior in form of interactive saliency maps in virtual reality.

Natural Examples: Natural dialogues aim to *"lower the threshold of ability required to analyze data"* and thus make XUIs more accessible to end users of XAI. The XUI is about presenting functionalities to request different natural language explanations. The interaction is mostly driven by the human through questions. Explanations are explicit but simplified in the form of textual answers. Kim et al. [46] present an XUI that enables users to ask factoid questions about charts in natural language (e.g., *"What age had the lowest population of males?"*). The XUI provides the answer and an explanation how it was derived from the chart (e.g., *"I looked up 'age' of the shortest blue bar*).

4.3 Interaction as Control

This concept supports a rapid and stable convergence of the human-computer system towards a target state. Building on *control theory*, the interaction is aiming *"to change a control signal to a desired level and updating its behavior according to feedback"* [42].

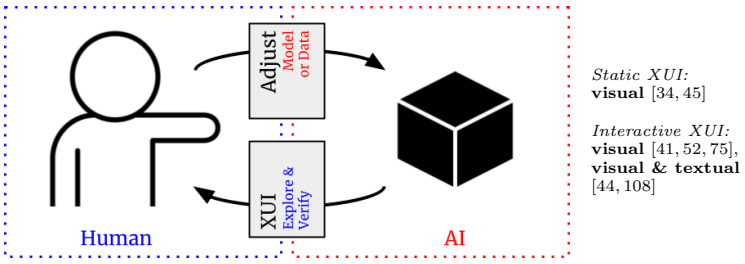


Fig. 3. XAI-interaction as control is about supporting a rapid convergence towards the desired AI behavior.

Transfer to XAI: This concept aligns with the ideas of interactive ML [29] and ML model tweaking. The XUI feeds control signals from the ML model to the human controller (feedback). These inform the controller how to change parameters of the ML model or its data so that the model adjusts its behavior (feedforward). The goal of the interaction is to reach the AI behavior desired by the controller. We found two streams of research that follow this paradigm. They can be distinguished by their targeted users: *AI experts* [41, 45, 52, 75, 78] or *AI novices* [44, 34, 108].

AI Expert Examples: Explanations are provided mainly on an abstract level as numbers and visualizations. The cycle of exploration and verification drives the process of understanding. The XUI is a standalone application facilitating this interaction while the actual model adjustments are performed in a separate UI (e.g., the development environment). For instance, [78] present an early XUI to debug rule-based expert systems by explaining why a rule was fired. Krause et al. [52] present the interactive visual analytics systems *Prospector*, that supports data scientists in understanding local predictions and deriving actionable insights on how to improve the ML model. They can (i) explore local predictions and simulate counterfactual changes by different ML models to support the formulation of tweaking hypotheses or (ii) verify how their implemented tweaking hypotheses change the prediction behaviour of the ML model. Hohman et al. [41] present *Gamut*, an XUI were *"interactivity was the primary mechanism for exploring, comparing, and explaining"*. User can link local and global explanations, ask counterfactual and compute similar instances. In contrast, Kaur et al. [45] show that the non-interactive XUIs of widely used explainability tools, such as InterpretML or SHAP, hinder experts to effectively control ML models.

AI Novice Examples: These XUI strive *"to effectively communicate relevant technical features of the [ML] model to a non-technical audience"* [108]. These XUIs provide explicit explanations to support the exploration. They also integrate controls for adjusting underlying the ML models without the need of a separate UI. Yu et al. [108] present an XUI for ML classification in the sensitive context of criminal justice. Their XUI enables designers and end-users to explore and understand algorithmic trade-offs based on an interactive confusion matrix and textual explanations. Further, it allows them to adjust model thresholds in a way that reflects their fairness beliefs (feedforward). Ishibashi et al. [44] present an XUI that synergetically combines low-level spectrograms with semantic thumbnails to interactively train a sound recognition AI. Fulton et al. [34] showcase how an XUI can be integrated into games for AI novices to generate usable data for AI experts.

4.4 Interaction as Experience

This concept considers human expectations towards a computer. It is closely related to *user experience* (UX) encompassing a person's emotions, feelings, and thoughts that may be formed before, during, or after interaction [53].

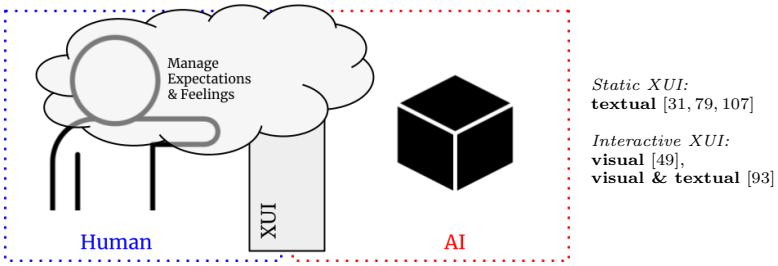


Fig. 4. XAI-interaction as experience is about managing expectations about the AI behavior.

Transfer to XAI: Applied to XAI, this interaction concept emphasizes managing the expectations and preferences of users about the AI. It centers around the explanatory goals of trust [49, 77, 79, 107], satisfaction [93], and persuasiveness [31].

Examples: Knijnenburg et al. show that letting users inspect a recommendation process through an interactive XUI increased their perceived understanding and satisfaction. Tsai et al. [93] investigate the relation of user preferences about explanation styles and user performance. Their results suggest that XUIs preferred by users *"may not guarantee the same level of performance"*. Yin et al. [107] show that a user's trust is impacted by upfront information on the AI's predictive accuracy even after repeated interactions. Pushing this interaction concept, Eiband et al. [31] show with their XUI that even empty (so-called placebic) explanations

can result in a soothing perceived understanding of users. As an intervention, Pilling et al. [77] outline a design fiction of an AI certification body that provides users with standardized AI quality marks (e.g., *"level 4: product is able to explain itself to users on request."*).

4.5 Interaction as Optimal Behavior

This concept centers around adapting the user behavior to better support their tasks and goals. It acknowledges that the interaction with the system is often constrained, and thus suboptimal. Users are trading off rewards and costs of an interaction. It builds around the idea of *bounded rationality* [87] according to which humans act as "satisficers" who strive for satisfying and sufficient solutions (instead of optimal ones) due to cognitive limitations.

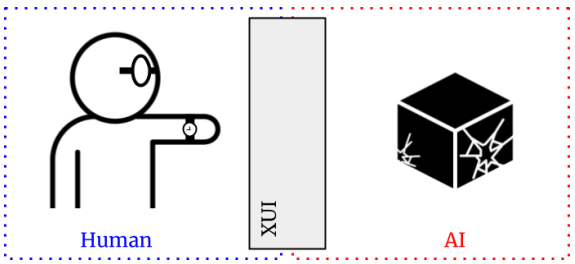


Fig. 5. XAI-interaction as optimal behavior is about adjusting the human behavior despite the cognitive or technical limitations of fully understanding the AI behavior.

Transfer to XAI: Applied to XAI research, the goal of the interaction is to guide users to reach a "satisficing" level of AI understanding for some downstream task. It focuses on providing explanations for *"training humans to have better interactions with AI"*, for example, when they face erroneous AI systems [99] or exhibit misconceptions caused by cognitive biases [97]. We distinguish between research that (i) examines limitations that occur during the interaction with an XAI [12, 13, 24, 60, 61, 70, 97] and (ii) designs interactions to better moderate these limitations [2, 16, 19, 59, 81, 98, 99].

Examples that Examine Limitations: Millecamp et al. [61] studied the impact of personal characteristics on the interaction and perception of XAI in a music recommender setting. They show that the perception and interaction with XUIs is influenced by a user's need for cognition (NFC) (i.e., their tendency to engage in and enjoy effortful cognitive activities). Nourani et al. [70] show that a user's first impression of an AI system influences their overall perception of the system. While a positive first impression may lead to automation bias, a negative first impression may result in a less accurate mental model. They call for XUIs that control a user's first impression and *"continually direct user attention to system strengths and weaknesses throughout user-system interactions"*. Similarly, Bucinca et al. [12] highlight that the effectiveness of XAI is impacted by the

design of the interaction itself. Thus, it is important to take *"into account the cognitive effort and cognitive processes that are employed [by the user]"* during their interpretation of explanations.

Examples that Moderate Limitations: Several of the works designed interactions that *"optimize the performance of the sociotechnical (human+AI) system as a whole"* [12]. For example, Wang et al. [98] provide confidence explanations to help users to gauge when or when not to trust an AI. Similarly, Schaekermann et al. [81] show that highlighting and textually explaining ambiguous predictions helps physicians to *"allocate cognitive resources and reassess their level of trust appropriately for each specific case"*. Abdul et al. [2] propose a visual explanation style that balances cognitive load and descriptive accuracy by limiting the visual chunks to be processed by the user. Further, they present a method to estimate users' cognitive load of explanations. Weisz et al. [99] teach users strategies to effectively interact with a limited capability chatbot in a banking and shopping context. Their interaction aims to explain to users why a chatbot may be unable to provide meaningful responses. For instance, explaining that the chatbot mapped the user's utterance to multiple low confidence intents because the utterance was poorly worded or ambiguous. Mai et al. [59] guide users through a military-inspired structured reflection process, called *after-action review* to understand the behavior of an AI agent. Accompanied by a visual explanation of AI decisions, the reflection process helped users to organize their cognitive process of understanding and kept them engaged.

4.6 Interaction as Tool Use

This concept centers around using computers to augment the user's capabilities beyond the tool itself. Following *activity theory*, the system influences the *"mental functioning of individuals"*. As such, AI can also be used as a tool for learning. For example, the social sciences use word embeddings as a diagnostic tool to quantify changes in society [35].

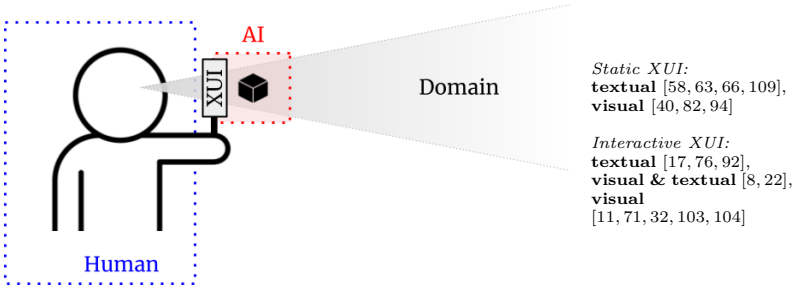


Fig. 6. XAI-interaction as tool use is about facilitating learning from the AI behavior about a given domain.

Transfer to XAI: Applied to XAI, this interaction concept helps humans to find hidden patterns and insights in domain-specific data. To facilitate this learning, some form of explanation is required. The XUI serves as a lens on a domain (beyond the AI behavior) that would otherwise be difficult to understand. In this way, the interaction contributes to augment human thinking.

Examples: Xie et al. [104] assist physicians analyzing chest x-rays of patients through an interactive mixed-modality XUI. Paudyal et al. [76] presents an interactive XUI for a computer-vision based sign language AI. The textual explanations provide learners with feedback on the location, shapes, and movements of their hands. Similarly, Schneeberger et al. [82] use an XUI to let users practice emotionally difficult social situations with a social AI agent. Das et al. [22] present an XUI which provides feedback on a chess player's intended moves. Their visual highlighting and textual explanations significantly improved the performance of chess players in a multi-day user study. They point out the importance of accompanying textual explanations for the AI reasoning. Only showing the visual explanation did not improve performance. Similarly, Feng et al. [32] support players by visually explaining evidences for each uncovered word of a quiz question. Xie et al. [103] use an interactive XUI with visual explanations to give game designers live-feedback on how challenging their created level designs are. Misztal-Radecka and Indurkhya [63] generate textual user stories for personas from large datasets to inform interaction designers about potentially relevant user groups.

Explainable Recommender Systems: In addition, most works on explainable recommender systems follow this interaction concept as their recommendations aim to give users insights about the recommender domain [40]. Some XUIs allow personalization by steering the recommendation behavior and thus, include aspects of the *interaction as control* concept. These user-initiated manipulations dynamically influence the recommendations and serve as a feedforward mechanism. However, users' focus is not about reaching an envisioned end state of AI behavior, but generating useful insights about the domain (or themselves). For example, O'Donovan et al. [71] present *PeerChooser*, an interactive movie recommender that enables users to provide "hints" about their current mood and needs by dragging movie genres closer or further away from their avatar. Bostandjiev et al. [11] use the XUI to explain a music recommendation process and to elicit preferences from users. Users can interactively adjust weights on the input and model level to explore the recommender. Chen et al. [17] present a preference-based recommender to increase users' product knowledge of high-investment products, such as digital cameras and laptops. Their XUI textually explains trade-offs within a set of recommended items.

4.7 Interaction as Embodied Action

This concept centers around collaboration and joint action with a computer. In 1960, Licklider formulated the vision of *man-computer symbiosis* in which

"men and computers [are] to cooperate in making decisions and controlling complex situations" [55]. Humans may be amplified through collaboration with AI. However, effective collaboration goes beyond interaction. In this way, this concept builds on theories from the computer-supported cooperative work (CSCW) community, such as mutual goal understanding, preemptive task co-management and shared progress tracking [96].

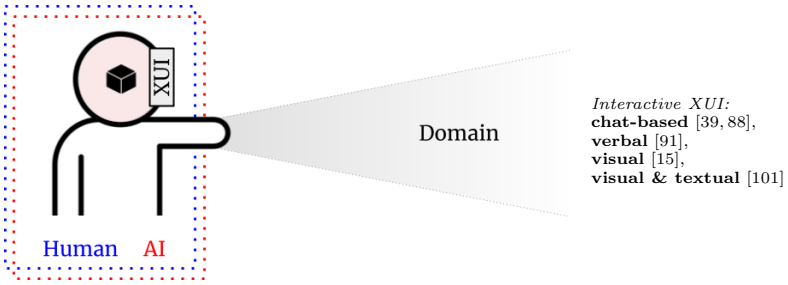


Fig. 7. XAI-interaction as embodied action is about establishing a joint understanding with the AI for an effective collaboration in a given domain.

Transfer to XAI: Applied to XAI, explanations are a crucial component for effective cooperation. A lack of explanatory communication resulted in dissatisfaction [38, 72]. In this way, XUIs contribute to the augmentation of human actions. A symbiotic relationship for which this is especially important involves *autonomous systems*. Autonomous systems in high-risk scenarios have a high degree of autonomy and thus *"need to explain what they are doing and why"* [39]. In such a setting, it is crucial for humans and agents alike to communicate each other's capabilities and intended next steps with respect to a common goal, often in real-time. We identified XUIs which are not only about understanding AI agents (*interaction as transmission*), but which enabled them to also influence the agents' actions – and vice versa [15, 39, 80]. Unlike *interaction as control* the interaction is not only driven by the human controller, but by both parties [6, 91, 101].

Examples: Tabrez et al. [91] present an AI agent that analyzes the game decisions of a human collaborator in a collaborative game setting and verbally interrupts the human in case the common goal becomes unattainable because of a wrong move. The AI agent dynamically constructs a *theory of mind* of the human collaborator and provides tailored explanations that aim to correct their understanding of the game situation. Chakraborti et al. [15] present an XUI that coordinates mission plans between a semi-autonomous search and rescue robot and a human commander who has an incomplete and possibly outdated map of the robot's environment. Visual explanations are embedded as changes in the commander map. The commander can either request (i) an optimal plan by the robot and explanations for this plan, or (ii) a potentially suboptimal

plan that is aligned with the commander's expectations. As such, the XUI reconciles potential mismatches about the plans between robot and commander. Hastie et al. [39] and Robb et al. [80] present an XUI that provides operators of autonomous underwater vehicles with why and why not explanations in real-time via a chat interface. Further, users can influence actions of the autonomous system through the XUI (e.g. setting reminders). Their XUI was reported to increase the situation awareness of operators and adjusted their mental model of system capabilities.

Table 1. Surveyed XAI publications categorized according to the different concepts of interaction by Hornbæk and Oulasvirta [42].

Interaction Concept	**Interaction Goal** *applied to XAI*	**References**
Transmission	Present users with accurate or complete explanation about AI behavior. *Explanatory goal: transparency*	[4, 9, 14, 23, 25, 26, 30, 48, 56, 68, 106]
Dialogue	Facilitate natural and iterative conversation about AI behavior. *Explanatory goals: transparency, scrutability*	[10, 18, 28, 46]
Control	Support rapid convergence towards desired AI behavior. *Explanatory goal: effectiveness*	[34, 41, 44, 45, 51, 52, 75, 78, 108]
Experience	Manage expectations about AI behavior. *Explanatory goals: satisfaction, trust, persuasiveness*	[31, 49, 77, 79, 93, 107]
Optimal Behavior	Adjust human behavior despite limitations of fully understanding the AI behavior. *Explanatory goal: efficiency*	[2, 12, 13, 16, 19, 24, 50, 59, 61, 60, 70, 81, 97–99]
Tool Use	Facilitate learning from AI behavior about a given domain. *Explanatory goals: effectiveness*	[8, 11, 17, 22, 71, 32, 40, 58, 63, 66, 76, 82, 92, 94, 103, 104, 109]
Embodied Action	Establish a joint understanding with the AI for an effective collaboration in a given domain. *Explanatory goal: effectiveness*	[15, 39, 80, 88, 91, 101]

5 Design Principles for Interactive XUI

In the last section, we described the general interplay between the XAI system and the user. Below, we will focus on the interactive qualities of the XUI itself. Vilone et al. define interactivity as *"the capacity of an explanation system to reason about previous utterances both to interpret and answer users' follow-up questions"* [95]. We expand this definition by building on the concept of *explanation facilities* that dates to the era of rule-based expert systems. Moore and Paris [64] proposed that a good explanation facility should, among others, fulfill the requirements of *naturalness* (explanations in natural language following a dialogue), *responsiveness* (allow follow-up questions), *flexibility* (make use of multiple explanation methods), and *sensitivity* (provided explanations should be informed by the user's knowledge, goal, context, and previous interaction). We analyzed our sample of XAI publications through the lens of these requirements to answer RQ2 (What design principles for interactive XUIs can be observed?). We found common interaction strategies and design recommendations [17, 45, 80, 104] that address aspects of these requirements. We unify and present them as *design principles*. In interaction design, design principles are *"guidelines for design of useful and desirable products"* [20].

5.1 Complementary Naturalness

> Consider complementing implicit explanations with rationales in natural language.

Why: Implicit visual explanations can accurately depict the inner workings of an AI but are often inaccessible to non-experts. In contrast, rationales in natural language are post-hoc explanations *"that are meant to sound like what a human [explainer] would say in the same situation"* [30]. Relaying facts through text may *"reassure users when system status might be uncertain or [...] obscure"* [80]. Combining visual cues with textual rationales can facilitate understanding and communicative effectiveness [30].

How: Kim et al. [46] outline a method that automatically generates explanations from visualizations through a template-based approach. Robb et al. [80] elaborate design recommendations on how to incorporate chat-based XUI for autonomous vehicle operators. For example, Yu et al. [108] provide users with a switch to change a visual explanation into verbose explicit sentences. Schaekermann et al. [81] complement quantitative low-confidence predictions with arguments in natural language to attract the attention of physicians. Sklar et al. [88] explain the reasoning behind an AI agent's actions through a chat-interface.

5.2 Responsiveness through Progressive Disclosure

> Consider offering hierarchical or iterative functionalities that allow follow-ups on initial explanations.

Why: Prior research indicated that there is a fine line between no explanation and too much explanation [61]. A user's individual need for cognition influences this threshold. Providing overly detailed explanations overwhelms users who may operate on a simpler mental model of the underlying AI.

How: Springer and Whittaker [89] recommend applying the interaction design pattern of *progressive disclosure*. It is about providing users only with high-level information and offering follow-up operations in case they are interested in further details[6] It resembles the *"progressive-step-by-step process"* demanded by [85]. As such, an XUI should (i) provide information on demand, (ii) hierarchically organize explanatory information, and (iii) keep track of the interaction with a user. For example, Millecamp et al. [61] provide a *Why?* button next to a recommendation. Clicking it provides a one-dimensional visual explanation in the form of a bar chart. If users are interested in additional details, they can click another button to receive a multi-dimensional visual explanation that compares multiple attributes of multiple recommendations in the form of a scatter plot. Krause et al. [52] use tooltips to summarize the most influential features and their sensitivity. If interested, users can drill down and freely explore these with partial dependence plots. Bock et al. [10] visualize a convolutional neural network in virtual reality. Progressive disclosure is realized through spatial distance. As the user approaches the network, more layers with finer granularity become visible. This design principle can also be implicitly implemented by enabling users to repeatedly adjust controls of the ML model [11, 108] or input parameters [18, 76] to progressively disclose local insights step-by-step.

5.3 Flexibility through Multiple Ways to Explain

> Consider offering multiple explanation methods and modalities to enable explainees to triangulate insights.

Why: Humans gain understanding in many ways. Paez [74] outlines them along a spectrum between understanding why (gained through observations and exemplifications) and objectual understanding (gained through idealizations and simplified models). In practice, there is often no best way to explain. For instance, a physician's *"differential diagnosis seldom relies on a single type of data"* [103]. In this way, explanation methods and modalities can complement each other.

How: This principle builds around the interaction design pattern of *multiple ways*[7] , which is about *"providing an opportunity to navigate [...] in more than one manner"*. Multiple publications recommend addressing local and global explanation paradigms within one XUI [24, 41, 104]. This enables users to get an overview of the overall AI behavior and scrutiny of individual cases at the same time. To facilitate this navigation, Liao et al. [54] present a catalog of natural

[6] nngroup.com/articles/progressive-disclosure/
[7] w3.org/tr/understanding-wcag20/navigation-mechanisms-mult-loc.html

language questions that can technically be answered by current XAI methods. Covering multiple of them under a *"holistic approach"* [54] allows users to triangulate insights. For example, Xie et al. [103] present a three-stage explanation workflow that supports physicians in top-down or bottom-up reasoning. Their XUI can *"connects the dots"* and highlight how explanations at each stage relate to one another. Wang et al. [97] present a XUI that provides feature attributions and counterfactual rules in parallel to support multiple ways of reasoning. Hohman et al. [41] provide highly interconnected visual model-level and instance-level explanations side by side to *"flexibly support people's differing processes"*. Chen et al. [17] provide different explanatory views that allow users to examine recommended products from different angles.

5.4 Sensitivity to the Mind and Context

> Consider offering functionalities to adjust explanations to explainees' mental models and contexts.

Why: Explanation needs of user evolve *"as one builds understanding and trust during the interaction process"* [54]. Further, prior beliefs and biases of users influence how they respond to different styles of explanations. This calls for *"a personalized approach to explaining ML systems"* [23].

How: This principle builds around the concept of *mixed-initiative interaction* [43], which emphasizes an interaction in which the human and the computer work towards the shared goal – fostering human understanding in the case of XAI. The timing of actions along the stages of grounding, listening, and interrupting is important for a successful interaction. To adapt its operations, an XUI needs to construct a computer model (or theory of mind [91]) of the user's mental model [65]. Despite its complexity, we found first examples. Tabrez et al. [91] estimate a human collaborator's beliefs in a collaborative game to identify explanation points. Other works [15, 19, 17], elicit preferences or beliefs to estimate a user's expected AI predictions (so called foils), e.g., so that counterfactual explanations can argument only regarding these. Wenskovitch et al. [100] present a method to infer user intent from interactions with visual explanations. Xie et al. [104] implement an *"urgent"* mode that can be toggled by physicians in a hurry to only see high confidence explanations with little system complexity.

6 Limitations and Outlook

Our review excluded publications outside the *ACM Digital Library* and the *SIGCHI* community. We are confident that our review covers many publications that emphasize the interaction design perspective of XAI. However, we probably have missed relevant applied research from adjacent XAI communities inside (e.g., FAccT) and outside (e.g., AIS) of *ACM*. Future work could extend our work with their learnings. Another promising direction for future research is constructive research that encompasses all presented design principles.

None of the survey publications considered all design principles in one XUI. This makes sense as researchers try to limit and control variables for a rigorous evaluation of their research questions. However, with the emergence of open-source explanation-generating toolkits it would be a logical next step to explore reusable and customizable XUI frameworks. These could integrate multiple explanation methods under a human-centric interaction concept.

7 Summary

Interaction design has been discussed as an important aspect for effective explainability in XAI. Yet, so far, it has not been systematically analyzed. Starting from a systematically obtained set of XAI publications that mention user interfaces or user interaction, we derived seven concepts of human-XAI interaction. Further, we analyzed the presented XUI and consolidated proposed recommendations as design principles encompassing four recurring themes: naturalness, responsiveness, flexibility, and sensitivity. We contribute a categorization to describe XAI work not only by the intended target audience or domain of application, but also through the pursued interaction concept. Our survey provides a starting point for researchers and practitioners planning and designing human-centric XAI systems.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and Trajectories for Explainable, Accountable and Intelligible Systems. CHI '18 (2018)
2. Abdul, A., von der Weth, C., Kankanhalli, M., Lim, B.Y.: COGAM: Measuring and Moderating Cognitive Load in ML Model Explanations. CHI '20 (2020)
3. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access (2018)
4. Alqaraawi, A., Schuessler, M., Weiss, P., Costanza, E., Berthouze, N.: Evaluating Saliency Map Explanations for Convolutional Neural Networks. IUI '20 (2020)
5. Amershi, S., et al.: Guidelines for human-AI interaction. CHI'19 (2019)
6. Andres, J., et al.: Introducing Peripheral Awareness as a Neurological State for Human-Computer Integration. CHI'20 (2020)
7. Arya, V., et al.: One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. arXiv (2019)
8. Barria-Pineda, J., Brusilovsky, P.: Explaining Educational Recommendations through a Concept-Level Knowledge Visualization. IUI '19 (2019)
9. Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N.: 'It's Reducing a Human Being to a Percentage'. CHI '18 (2018)
10. Bock, M., Schreiber, A.: Visualization of Neural Networks in Virtual Reality Using Unreal Engine. VRST '18 (2018)
11. Bostandjiev, S., O'Donovan, J., Höllerer, T.: TasteWeights: A Visual Interactive Hybrid Recommender System. RecSys '12 (2012)
12. Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L.: Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating XAI Systems. IUI '20 (2020)

13. Bunt, A., Lount, M., Lauzon, C.: Are Explanations Always Important? A Study of Deployed, Low-Cost Intelligent Interactive Systems. IUI '12 (2012)
14. Cai, C.J., Jongejan, J., Holbrook, J.: The Effects of Example-Based Explanations in a Machine Learning Interface. IUI '19 (2019)
15. Chakraborti, T., Sreedharan, S., Grover, S., Kambhampati, S.: Plan Explanations as Model Reconciliation: An Empirical Study. HRI '19 (2019)
16. Chen, L.: Adaptive Tradeoff Explanations in Conversational Recommenders. RecSys '09 (2009)
17. Chen, L., Wang, F.: Explaining Recommendations Based on Feature Sentiments in Product Reviews. IUI '17 (2017)
18. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H.: Explaining Decision-Making Algorithms through UI. CHI '19 (2019)
19. Chromik, M., Fincke, F., Butz, A.: Mind the (Persuasion) Gap: Contrasting Predictions of Intelligent DSS with User Beliefs. EICS '20 Companion (2020)
20. Cooper, A., Reimann, R., Cronin, D.: About Face 3: The Essentials of Interaction Design (2007)
21. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A Survey of the State of Explainable AI for Natural Language Processing. arXiv (2020)
22. Das, D., Chernova, S.: Leveraging Rationales to Improve Human Task Performance. IUI '20 (2020)
23. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K.E., Dugan, C.: Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. IUI '19 (2019)
24. Dodge, J., Penney, S., Hilderbrand, C., Anderson, A., Burnett, M.: How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games. CHI '18 (2018)
25. Dominguez, V., Messina, P., Donoso-Guzmán, I., Parra, D.: The Effect of Explanations and Algorithmic Accuracy on Visual Recommender Systems of Artistic Images. IUI '19 (2019)
26. Donkers, T., Kleemann, T., Ziegler, J.: Explaining Recommendations by Means of Aspect-Based Transparent Memories. IUI '20 (2020)
27. Doshi-Velez, F., Kim, B.: Towards A Rigorous Science of Interpretable Machine Learning. arXiv (2017)
28. Douglas, N., Yim, D., Kartal, B., Hernandez-Leal, P., Maurer, F., Taylor, M.E.: Towers of Saliency: A Reinforcement Learning Visualization Using Immersive Environments. ISS '19 (2019)
29. Dudley, J.J., Kristensson, P.O.: A Review of User Interface Design for Interactive Machine Learning. ACM Trans. Interact. Intell. Syst. (2018)
30. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., Riedl, M.O.: Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. IUI '19 (2019)
31. Eiband, M., Buschek, D., Kremer, A., Hussmann, H.: The Impact of Placebic Explanations on Trust in Intelligent Systems. CHI EA '19 (2019)
32. Feng, S., Boyd-Graber, J.: What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. IUI '19 (2019)
33. Ferreira, J.J., Monteiro, M.S.: What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. LNCS '20 (2020)
34. Fulton, L.B., Lee, J.Y., Wang, Q., Yuan, Z., Hammer, J., Perer, A.: Getting Playful with Explainable AI. CHI EA '20 (2020)
35. Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes (2018)

36. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Surveys (2018)
37. Gunning, D.: DARPA's XAI Program. IUI '19 (2019)
38. Guzdial, M., Liao, N., Chen, J., Chen, S.Y., Shah, S., Shah, V., Reno, J., Smith, G., Riedl, M.O.: Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators. CHI '19 (2019)
39. Hastie, H., Chiyah Garcia, F.J., Robb, D.A., Laskov, A., Patron, P.: MIRIAM: A Multimodal Interface for Explaining the Reasoning Behind Actions of Remote Autonomous Systems. ICMI '18 (2018)
40. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining Collaborative Filtering Recommendations. CSCW '00 (2000)
41. Hohman, F., Head, A., Caruana, R., DeLine, R., Drucker, S.M.: Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. CHI '19 (2019)
42. Hornbaek, K., Oulasvirta, A.: What Is Interaction? CHI '17 (2017)
43. Horvitz, E.: Principles of Mixed-Initiative User Interfaces. CHI '99 (1999)
44. Ishibashi, T., Nakao, Y., Sugano, Y.: Investigating Audio Data Visualization for Interactive Sound Recognition. IUI '20 (2020)
45. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting Interpretability. CHI '20 (2020)
46. Kim, D.H., Hoque, E., Agrawala, M.: Answering Questions about Charts and Generating Visual Explanations. CHI '20 (2020)
47. Kitchenham, B., Charters, S.: Guidelines for performing Systematic Literature Reviews in Software Engineering (2007)
48. Kleinerman, A., Rosenfeld, A., Kraus, S.: Providing Explanations for Recommendations in Reciprocal Environments. RecSys '18 (2018)
49. Knijnenburg, B.P., Bostandjiev, S., O'Donovan, J., Kobsa, A.: Inspectability and Control in Social Recommenders. RecSys '12 (2012)
50. Kocaballi, A.B., Coiera, E., Berkovsky, S.: Revisiting Habitability in Conversational Systems. CHI EA '20 (2020)
51. Koch, J., Lucero, A., Hegemann, L., Oulasvirta, A.: May AI? Design Ideation with Cooperative Contextual Bandits. CHI '19 (2019)
52. Krause, J., Perer, A., Ng, K.: Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. CHI '16 (2016)
53. Law, E.L.C., Roto, V., Hassenzahl, M., Vermeeren, A.P.O.S., Kort, J.: Understanding, Scoping and Defining User Experience. CHI '09 (2009)
54. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: Informing Design Practices for Explainable AI User Experiences. CHI '20 (2020)
55. Licklider, J.: Man-Computer Symbiosis (1960)
56. Lim, B.Y., Dey, A.K.: Weights of Evidence for Intelligible Smart Environments. UbiComp '12 (2012)
57. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods (2020)
58. Ludwig, J., Geiselman, E.: Intelligent Pairing Assistant for Air Operation Centers. IUI '12 (2012)
59. Mai, T., Khanna, R., Dodge, J., Irvine, J., Lam, K.H., Lin, Z., Kiddle, N., Newman, E., Raja, S., Matthews, C., Perdriau, C., Burnett, M., Fern, A.: Keeping It "Organized and Logical". IUI '20 (2020)
60. Mikhail, M., Roegiest, A., Anello, K., Wei, W.: Dancing with the AI Devil: Investigating the Partnership Between Lawyers and AI. CHIIR '20 (2020)

61. Millecamp, M., Htun, N.N., Conati, C., Verbert, K.: To Explain or Not to Explain: The Effects of Personal Characteristics When Explaining Music Recommendations. IUI '19 (2019)
62. Miller, T.: Explanation in Artificial Intelligence: Insights From the Social Sciences. Artificial Intelligence (2019)
63. Misztal-Radecka, J., Indurkhya, B.: Persona Prototypes for Improving the Qualitative Evaluation of Recommendation Systems. UMAP '20 Adjunct (2020)
64. Moore, J.D., Paris, C.: Requirements for an expert system explanation facility. Computational Intelligence (1991)
65. Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein Macrocognition, G.: Explanation in Human-AI Systems. arXiv (2019)
66. Muhammad, K.I., Lawlor, A., Smyth, B.: A Live-User Study of Opinionated Explanations for Recommender Systems. IUI '16 (2016)
67. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, Methods, and Applications in Interpretable Machine Learning (2019)
68. Musto, C., Lops, P., de Gemmis, M., Semeraro, G.: Justifying Recommendations through Aspect-Based Sentiment Analysis of Users Reviews. UMAP '19 (2019)
69. Norman, D., Draper, S.: User Centered System Design: New Perspectives on Human-Computer Interaction (1986)
70. Nourani, M., Honeycutt, D.R., Block, J.E., Roy, C., Rahman, T., Ragan, E.D., Gogate, V.: Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis. CHI EA '20 (2020)
71. O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., Höllerer, T.: PeerChooser: Visual Interactive Recommendation. CHI '08 (2008)
72. Oh, C., Kim, S., Choi, J., Eun, J., Kim, S., Kim, J., Lee, J., Suh, B.: Understanding How People Reason about Aesthetic Evaluations of AI. DIS '20 (2020)
73. Oulasvirta, A., Hornbaek, K.: HCI Research as Problem-Solving. CHI '16 (2016)
74. Páez, A.: The Pragmatic Turn in Explainable Artificial Intelligence (XAI). Minds and Machines (2019)
75. Patel, K., Bancroft, N., Drucker, S.M., Fogarty, J., Ko, A.J., Landay, J.: Gestalt: Integrated Support for Implementation and Analysis in ML. UIST '10 (2010)
76. Paudyal, P., Banerjee, A., Gupta, S.: On Evaluating the Effects of Feedback for Sign Language Learning Using Explainable AI. IUI '20 (2020)
77. Pilling, F., Akmal, H., Coulton, P., Lindley, J.: The Process of Gaining an AI Legibility Mark. CHI EA '20 (2020)
78. Poltrock, S.E., Steiner, D.D., Tarlton, P.N.: Graphic Interfaces for Knowledge-Based System Development (1986)
79. Pu, P., Chen, L.: Trust Building with Explanation Interfaces. IUI '06 (2006)
80. Robb, D.A., Lopes, J., Padilla, S., Laskov, A., Chiyah Garcia, F.J., Liu, X., Scharff Willners, J., Valeyrie, N., Lohan, K., Lane, D., Patron, P., Petillot, Y., Chantler, M.J., Hastie, H.: Exploring Interaction with Remote Autonomous Systems Using Conversational Agents. DIS '19 (2019)
81. Schaekermann, M., Beaton, G., Sanoubari, E., Lim, A., Larson, K., Law, E.: Ambiguity-Aware AI Assistants for Medical Data Analysis. CHI '20 (2020)
82. Schneeberger, T., Gebhard, P., Baur, T., André, E.: PARLEY: A Transparent Virtual Social Agent Training Interface. IUI '19 (2019)
83. Schuessler, M., Wei\ss{}, P.: Minimalistic Explanations: Capturing the Essence of Decisions. CHI EA '19 (2019)
84. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal (1948)

85. Shneiderman, B.: Bridging the Gap Between Ethics and Practice. ACM Transactions on Interactive Intelligent Systems (2020)
86. Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., Diakopoulos, N.: Confessions: Grand challenges for HCI researchers. Interactions (2016)
87. Simon, H.A.: Models of bounded rationality: Empirically grounded economic reason, vol. 3. MIT press (1997)
88. Sklar, E.I., Azhar, M.Q.: Explanation through Argumentation. HAI '18 (2018)
89. Springer, A., Whittaker, S.: Progressive Disclosure. ACM Transactions on Interactive Intelligent Systems (2020)
90. Stolterman, E., Wiltse, H., Chen, S., Lewandowski, V., Pak, L.: Analyzing artifact interaction complexity (2012)
91. Tabrez, A., Agrawal, S., Hayes, B.: Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning. HRI '19 (2019)
92. Tintarev, N.: Explanations of Recommendations. RecSys '07 (2007)
93. Tsai, C.H., Brusilovsky, P.: Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. UMAP '19 (2019)
94. Vig, J., Sen, S., Riedl, J.: Tagsplanations: Explaining Recommendations Using Tags. IUI '09 (2009)
95. Vilone, G., Longo, L.: Explainable Artificial Intelligence: a Systematic Review. arXiv (2020)
96. Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., Wang, Q.: From Human-Human Collaboration to Human-AI Collaboration. CHI EA '20 (2020)
97. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing Theory-Driven User-Centric Explainable AI. CHI '19 (2019)
98. Wang, N., Pynadath, D.V., Hill, S.G.: Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations. HRI '16 (2016)
99. Weisz, J.D., Jain, M., Joshi, N.N., Johnson, J., Lange, I.: BigBlueBot: Teaching Strategies for Successful Human-Agent Interactions. IUI '19 (2019)
100. Wenskovitch, J., Dowling, M., North, C.: With Respect to What? Simultaneous Interaction with Dimension Reduction and Clustering Projections. IUI '20 (2020)
101. Wiegand, G., Schmidmaier, M., Weber, T., Liu, Y., Hussmann, H.: I Drive - You Trust: Explaining Driving Behavior Of Autonomous Cars. CHI EA '19 (2019)
102. Wolf, C.T.: Explainability Scenarios: Towards Scenario-Based XAI Design. IUI '19 (2019)
103. Xie, J., Myers, C.M., Zhu, J.: Interactive Visualizer to Facilitate Game Designers in Understanding Machine Learning. CHI EA '19 (2019)
104. Xie, Y., Chen, M., Kao, D., Gao, G., Chen, X.A.: CheXplain: Enabling Physicians to Explore and Understand Data-Driven Medical Imaging Analysis. CHI '20 (2020)
105. Xu, W.: Toward human-centered AI: A perspective from human-computer interaction. Interactions (2019)
106. Yang, F., Huang, Z., Scholtz, J., Arendt, D.L.: How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? IUI '20 (2020)
107. Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the Effect of Accuracy on Trust in Machine Learning Models. CHI '19 (2019)
108. Yu, B., Yuan, Y., Terveen, L., Wu, Z.S., Forlizzi, J., Zhu, H.: Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives. DIS '20 (2020)
109. Zanker, M.: The Influence of Knowledgeable Explanations on Users' Perception of a Recommender System. RecSys '12 (2012)

I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Malin Eiband
LMU Munich
Munich, Germany
malin.eiband@ifi.lmu.de

Felicitas Buchner
LMU Munich
Munich, Germany
felicitas.buchner@campus.lmu.de

Adrian Krüger
LMU Munich
Munich, Germany
adrian.krueger@campus.lmu.de

Andreas Butz
LMU Munich
Munich, Germany
butz@ifi.lmu.de

ABSTRACT

Unintended consequences of deployed AI systems fueled the call for more interpretability in AI systems. Often explainable AI (XAI) systems provide users with simplifying local explanations for individual predictions but leave it up to them to construct a global understanding of the model behavior. In this work, we examine if non-technical users of XAI fall for an illusion of explanatory depth when interpreting additive local explanations. We applied a mixed methods approach consisting of a moderated study with 40 participants and an unmoderated study with 107 crowd workers using a spreadsheet-like explanation interface based on the SHAP framework. We observed what non-technical users do to form their mental models of global AI model behavior from local explanations and how their perception of understanding decreases when it is examined.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; *User studies.*

KEYWORDS

explainable AI; Shapley explanation; cognitive bias; understanding

ACM Reference Format:
Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces (IUI '21), April 14–17, 2021, College Station, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3397481.3450644

1 INTRODUCTION

There is a growing awareness that machine learning-based *intelligent systems* (IS) need to be capable of explaining their behavior in human-understandable terms to prevent unintended consequences in sensitive contexts of society (e.g., credit scoring, recruiting, predictive policing, or criminal justice) [18]. Driven by this concern, the field of *explainable artificial intelligence* (XAI) develops models, methods, and explainable interfaces that are interpretable to human users by providing some notion of explanation [16]. Organizations aspire to deploy explainability techniques to wider non-technical audiences to comply with demands and regulations [5]. Such users of XAI, also referred to as *operators* or *executers* [56], consume machine learning (ML) predictions to inform their decisions. They are centered between the developers and the individuals affected by the predictions [56]. Because they may be accountable for their decisions, they utilize explanations to assure the underlying models is *trustworthy* (i.e., *"they can reasonably trust a model's outputs"* [5]) (*operator-interpretability* [56]).

Many empirical XAI studies limit their explanation approaches to *outcome explanations* [21] for individual ML predictions (*local* explainability) without examining if users build an accurate mental model of the overall ML model behavior (*global explainability*). Local explanations based on *Shapley values* [51] are widely used in practice [5]. For a single observation, they perfectly distribute the difference between the average prediction and the actual prediction between its features [30]. Thus, much of the inherent ML model complexity (e.g., feature interactions) is simplified into accessible Shapley values [20]. Relying on them alone might leave users with a false sense of understanding that is merely illusive. Further, the explainability of explanations is often assessed through subjective user ratings [41]. In this type of evaluation, users are asked to report their perceived understanding, trust, or other relevant mental factors through one-shot ratings with little to no incentives for self-reflection or self-calibration [34, 35]. It has been shown, however, that people are *"often miscalibrated about their own judgments"* [35]. Psychological research has demonstrated in many contexts that humans have a robust bias of overconfidence regarding their understanding of how complex concepts work [46]. After being asked to explicate and actively reflect on their understanding, people significantly reduce their estimation of their own knowledge.

In this paper, we argue that because of this *illusion of explanatory depth* (IOED) [46], XAI explanations (especially in the form of additive local explanations for individual predictions) may be misleading for non-technical XAI users. Rather than stipulating effective gains in human understanding, they might cause them

IUI '21, April 14–17, 2021, College Station, TX, USA

ACM ISBN 978-1-4503-8017-1/21/04…$15.00
https://doi.org/10.1145/3397481.3450644

to form false or incomplete beliefs about the explained ML model. Some researchers already speculated that an IOED could be at play when users are confronted with XAI explanations [14, 35, 54]. To our knowledge, however, this has not yet been systematically investigated. We examine if end users fall for an IOED when consuming XAI explanations in a decision-support scenario. In particular, we focus on the effect of *local post-hoc* explanations using *Shapley* values. We conducted human-grounded evaluations [16] in a *crowd lending* scenario using a tabular real-world data set. The scenario leverages a functional black-box ML model (a random forest classifier) and functional Shapley explanations generated by the widely-used explainability framework *SHAP* [30][1]. We followed a mixed methods approach. First, we moderated 40 participants through the study and observed their interactions. Second, we verified our hypotheses in an unmoderated study with 107 crowd workers. The studies has been approved by our internal IRB.

With our work, we follow the call to improve the user experience of XAI for a wider range of stakeholders [6]. The majority of current XAI research targets ML experts (e.g., data scientists) [25] or specific domain experts (e.g., physicians) [2, 15]. In contrast, we focus on the understanding of users with low expertise in AI. Our work contributes to the HCI community in three ways: First, we present *SHAPTable* an explanation interfaces targeting end non-technical users of XAI systems that embeds Shapley explanations in an accessible spreadsheet-like user interface (section 4). Second, based on an empirical examination we show that non-technical users fall for an IOED when relying on Shapley explanations (section 6).

2 BACKGROUND AND RELATED WORK

2.1 Explanations from Intelligent Systems

The research field of XAI aims to make black-box ML models interpretable by generating some notion of explanation that can be used by humans to interpret the behavior of an ML model [58]. An ML model is considered as a black-box if humans can observe the inputs and outputs of the model but have difficulties understanding the mapping between them. This may result from the model either being too complex, such as many deep neural networks, or being proprietary, such as with the COMPAS system [47]. Black-box models are often reported to yield a high *predictive accuracy* with less effort [47]. There are two broad categories of explainability approaches: *transparency-based* and *post-hoc* explainability [29]. *Transparency-based* approaches focus on how the model works and leverage model characteristics to explain it. This may involve using simpler models with intrinsic explainability that may yield a lower predictive accuracy. In contrast, *post-hoc* approaches ignore model characteristics. Instead, they observe the inputs and outputs of the ML model and try to detect regularities in its behavior in an inductive manner. Thus, post-hoc approaches have no impact on the predictive accuracy of a model but may oversimplify the true model behavior. The ability of an explanation method to accurately describe the behavior of an ML model is referred to as *descriptive accuracy* [36] or *fidelity* [47]. Human understanding in XAI can be fostered either by offering means of introspection or through explanations [7]. A large variety of methods exist for both approaches [21]. XAI research distinguishes two types of explanations - local and global [2, 21]. *Local* explanations of an ML model explain why an individual model prediction was made. In contrast, *global* explanations aim to convey the overall structure of the model by looking at model predictions on an aggregated level. Some definitions of explainability are rather *system-centric*. Doshi-Velez and Kim [16] describe it as a model's *"ability to explain or to present in understandable terms to a human."* Miller [32] takes a more *human-centered* perspective calling it *"the degree to which an observer can understand the cause of a decision"*. For an explanation to be effective, it does not only need to have a sufficient level of fidelity but must *"provide insight for a particular audience into a chosen domain problem"* [36].

2.2 Illusion of Explanatory Depth

Insights emerge when humans gain *"a clear, deep [...] understanding of a complicated problem or situation"*[2]. Human understanding, however, is often impacted by various cognitive biases. Research in cognitive sciences showed that people often form an inaccurate understanding of complex systems and often overrate the depth of their knowledge [35]. Rozenblit and Keil coined this type of overconfidence bias as the *illusion of explanatory depth (IOED)* [46]. They observed that laypeople consistently reduced the estimation of their own knowledge of different phenomena or devices after they were inquired to provide explanations about them or apply their understanding. Furthermore, people are often surprised by their limited explanations [4]. The IOED is more pronounced for *explanatory knowledge*, i.e., knowledge that involves complex causal patterns, than it is for *descriptive knowledge*, i.e., knowledge about facts (names of capitals), procedures (baking), or narratives (movie plots) [28, 46]. The IOED has first been demonstrated for people's understanding of causally complex systems in mechanical (bicycles, crossbows) [28, 33, 46] and natural (tides, rainbows) [46] domains. Subsequent work reproduced the IOED for social and policy domains (voting, mental disorder) [4, 60].

The illusion is believed to be caused by the way humans build their *conceptual knowledge*. Conceptual knowledge refers to the entirety of a person's *concepts* that are causally related to each other. According to the *theory-based* approach, people form *theories* about all their concepts, not just for those that they use regularly [46]. For instance, people form their own theories of what causes volcanic eruptions or how AI-based systems derive their predictions even though they were never confronted with one. These theories often consist of vague explanations that are not necessarily accurate nor coherent with each other [37]. When inquired to explicate parts of our conceptual knowledge to ourselves or others, we fall for the illusion to think we know more about a system than we actually do. Four factors are believed to influence the emergence of an IOED [46]: (i) *Representation/recovery confusion:* We overestimate our abilities to remember what we have observed. People tend to store observations as mental images. If the stored mental images do not correspond to the original facts, the IOED occurs. (ii) *Label/mechanism confusion:* Most complex systems are hierarchical with various levels of sub components. If we can name and describe individual parts on the first level of the hierarchy, we often assume

[1] https://github.com/slundberg/shap

[2] https://dictionary.cambridge.org/dictionary/english/insight

to understand how the overall system works, even though we have little insight into the levels further down the hierarchy. (iii) *Undefined end states*: Because of the hierarchical and related structure of complex topics, we have difficulties to imagine what constitutes a good and complete understanding or explanation. The end states for descriptive knowledge about facts or procedures are much clearer (e.g., naming the capital of a country or reverse engineering how to book a flight). (iv) *Lack of practice*: in everyday life most people regularly retrieve facts or reconstruct procedures. However, many people lack the practice of giving an explanation of complex topics. Just because we consume or make up explanations does not mean that we can produce effective explanations when needed.

2.3 IOED and Cognitive Biases in XAI and IS

Building on the IOED theories, it can be assumed that users of XAI systems form their own theories about the global behavior of the underlying ML model during interaction with the explanation facility. These also overlap with the widespread HCI concept of mental models. According to Norman, people form theories about how objects and systems work to explain what they observe [39]. A *mental model* refers to a person's understanding of how a system works and how the person's behavior affects it. People form mental models for all kinds of systems including objects, people, and services. The respective mental model is adjusted with every interaction (e.g., exposure to an XAI explanation) and helps the person to reflect on their belief about the system (e.g., the ML model behavior) [39].

Little research has been published on a potential IOED in the context of XAI or IS. Some researchers speculated that an IOED may be at play when users deal with explanations from XAI systems [14, 35, 54]. Collaris et al. observed during their XAI evaluation that their users did not question the validity of local explanations, even when provoked to do so [14]. Sokol and Flach call for an XAI validation protocol that addresses the IOED [54]. Kaur et al. observed that even data scientists and ML engineers took visual explanations of interpretability tools at face value and missed to effectively use them to uncover data or model issues. The provided XAI explanations encouraged the users to apply their heuristic thinking instead of activating their analytical thinking [25]. Even though the IOED itself received little attention in the context of intelligent systems and XAI, there is prior research on cognitive biases of explanations from intelligent systems investigating automation [9, 31, 40, 49, 52], anchoring [19, 27, 57], framing [26], and confirmation biases [23, 55]. A cognitive bias related to the IOED is the Dunning-Kruger effect of illusory overconfidence, which states that people with low competence at a given task tend to overestimate their task performance [49]. It occurs only with individuals with low competence while the IOED affects almost everyone.

3 RESEARCH QUESTIONS AND HYPOTHESES

Our work investigates the formation and the accuracy of operators' understanding of the ML model behavior from Shapley based local explanations. **The overarching research question (ORQ) of our work is to examine whether non-technical users of such XAI systems are prone to an IOED**. It is driven by the following research questions:

- **RQ1**: How robust is a self-reported global understanding gained from local explanations when examined?
  - H1: When participants are exposed to local explanations, this leads to an increased perception of understanding how the XAI system works (compared to no explanations)
  - H2: The participants' perception of understanding decreases after they have been examined for their understanding (IOED applies)
- **RQ2**: What do non-technical XAI users do to construct a global understanding from local explanations?

We focus on Shapley based explanations because, despite their vulnerability to adversarial attacks [53] and potential infidelity [20], we consider them as relevant for end users for two reasons: (i) enabled by the mathematical properties of accuracy and consistency, multiple local explanations can be combined to be contrastive and counterfactual [43] as well as interactive [13], (ii) the *SHAP* framework is widely used by XAI practitioners[3] and thus end users will likely come across Shapley based explanations, (iii) model agnostic approaches allow system designers to offer uniform explanation interfaces even when the underlying ML models differ.

However, human cognition is biased towards simple explanation [11]. Thus, if users' expectations are not properly calibrated, we hypothesize they may be prone to an IOED for two reasons: (i) *Representation/recovery confusion through abstraction of local insights:* User that are provided with local justifications of an XAI system may perceive to understand why those explanations were chosen by the system. However, under the influence of prior beliefs and misconceptions about AI, they may abstract their local insights into higher-level anecdotal evidence that may not be consistent with the predictions of other observations. End users may only become aware of these inconsistencies when they recall their abstractions to self-explain their understanding of the global ML model behavior [22]. (ii) *Label/mechanism confusion through subtle interactions:* Shapley explanations hide much of the model's complex behavior behind accessible feature value attributions [20]. Knowing what features a model has access to and the effect of feature values for some observations might results in the impression that the user understands how the model comes to its predictions for all observations. However, especially in state-of-the-art black box ML models, feature values may interact with one another in non-linear ways and significantly influence the predictions for some observations while having little effect on others.

4 SHAPTABLE

We outline the exemplary XAI system *SHAPTable* that serves as the apparatus for our user studies. First, we describe the setting and implementation details. Second, we provide details on the used explanation-generation method and the rationale for our explanation interface.

4.1 Scenario, Data Set, and ML Model

Scenario. Our scenario resembles a decision-support situation in which the human decision-maker is accompanied by an intelligent

[3] compared with other open-source XAI frameworks (such as LIME, AIX360, or DALEX), SHAP has the most engagements on GitHub

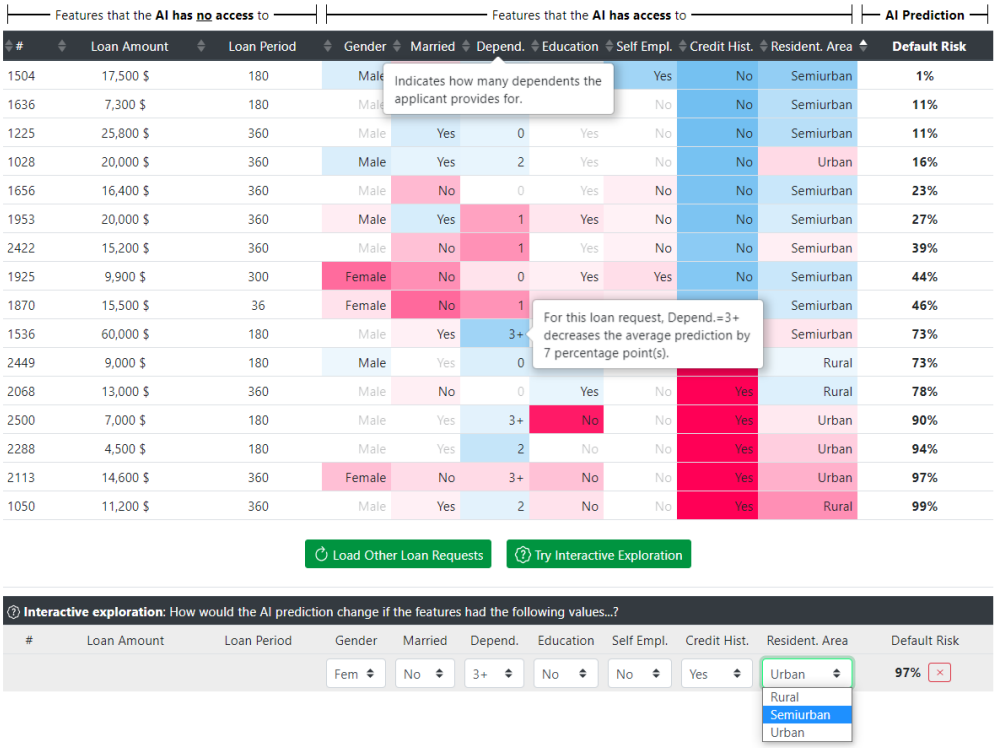


Figure 1: Overview of the explanation interface. Participants were presented a representative sample of 16 loan requests and their respective default risk prediction. Feature values of our ML model were shaded depending on their their Shapley values.

and interpretable system. Following [56], we take an XAI *operator* perspective in a loan application scenario. In such a scenario, the operating user of the XAI system is centered between the developer of the system and a decision-subject individual affected by the decision. We put our study participants in the shoes of a private lender on a fictional crowd lending platform[4]. Participants can see demographic information, loan details, and credit history of individuals that request a loan on the platform. Each request is accompanied by an "AI-based intelligent prediction" of the *default risk*, i.e., the probability that the borrower fails to service a loan installment some time during the loan period. The prediction is introduced as an "AI-based" feature that is based on machine learning from historic cases. As part of the scenario, participant evaluated a novel feature that explains the default risk prediction for each lending request through Shapley explanations. People utilize explanations for learning [32]. Thus, participants were instructed to give feedback to the platform if the provided explanation facility supports them in learning about the behavior of the default risk prediction feature (*operator interpretability* [56]).

Dataset. We chose a tabular data set for our user studies as many ML models deployed in practice build on this type of data. This applies especially to regulated domains such as healthcare, finance, and public services [5, 30]. Tabular data is often characterized by individually meaningful features and, unlike images or time series, lacks strong temporal or spatial structures [30]. Thus, each feature represents a distinct concept of a person's conceptual knowledge (e.g., gender, education, credit history). We built on the *Loan Prediction*[5] data set that is widely used for educational purposes. It consists of 614 loan requests with 13 columns. We relabeled two columns of the data set to be consistent with our scenario[6].

ML Model. We calculated the default risk prediction via a *random forest classifier (RFC)*. RFCs are widely used in many real-world contexts because of their practicability. They often yield competitive performances even without extensive ML engineering efforts. Especially for tabular data, tree-based models often outperform other black-box models [30]. However, random forests are considered black-box ML models. They consist of many decision trees. Each tree is trained on a random selection of features. The classifications

[4]a platform that facilitates the matchmaking between private lenders and borrowers over the internet

[5]https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/ or https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

[6]we re-framed the *Loan_Status* column to represent the default risk and the *Credit_History* column to represent a negative item on a credit report.

of individual trees are then combined into a final classification by a majority vote. Although individual decision trees are interpretable, it is unfeasible to understand the prediction behavior of their ensemble. To limit the cognitive load for participants we chose to train our model on a subset of columns. We used only the seven categorical columns (5 binary, 1 ternary, and 1 with four possible values). We trained a binary RFC with 100 decision trees using a 80:20 split for the training and validation sets. The split was stratified to have the same distribution of binary predictions between training and test sets. Other than that, we used the default hyperparameters of the *scikit-learn* package. The accuracy of the predicted default risk on the validation set was 0.83.

4.2 Explanation Facility

Explanation-generating Method. To algorithmically generate explanations for the default risk predictions, we build on the widely used post-hoc explanation framework *SHAP (SHapley Additive exPlanations)* [30]. SHAP belongs to the class of *additive feature attribution methods* where the explanation is represented as a linear function of feature contributions towards an ML prediction. It trains a surrogate model by slightly changing the inputs and testing the impact on the model outputs. The SHAP framework unifies the ideas of other feature attribution methods (such as LIME [44]) with *Shapley values*, which originate from game theory [51]. Applying Shapley values to XAI, an ML prediction can be modelled as a cooperative game between the features to produce a prediction. As the features may influence one another through interactions, the game is a cooperative one. With Shapley values we can assign a unique and fair contribution to each feature over all possible coalitions of features despite the presence of interactions. SHAP assigns a number for each input feature (the Shapley value) that is guaranteed to be consistent under mathematical guarantees: (i) *local accuracy* ensures that the sum of the feature contributions matches the ML prediction of an instance, (ii) *missingness* ensures that feature values that have no effect on the model prediction (e.g., because they are constant) have a Shapley value of zero, (iii) *consistency* ensures that changes in the contribution of an individual feature value in the black-box model result in a consistent change of the respective Shapley value. Consistency is interesting because it allows users to compare contributions between multiple observations, groups of observations, or even models. All contributions are relative to the *expected value*. The expected value equals the percentage of defaulted loan requests in the data set (32% for our data set). As such, it serves as a base value for all requests. The Shapley value for a feature value describes the direction and strength of the contribution relative to the expected value.

Explanation Interface. The SHAP framework provides information-dense visualizations of local and global feature attributions out-of-the-box. However, prior research showed that even ML experts face challenges to interpret them correctly without assistance [25]. Thus, for our explanation facility, we borrowed ideas from these visualizations but worked with the raw Shapley values. We assumed that most explanation-seeking end users in the decision-support context are familiar with spreadsheets. Thus, our explanation interface resembles a spreadsheet-like user interface that is overlaid with a heat map of Shapley values. We show 16 loan requests from the data set with their respective default risk prediction in percentage (i.e., 0%=no risk and 100%=highest risk of defaulting). The initial loan requests were sampled according to the confusion matrix to represent a representative range of default risk probabilities.[7] Each loan request is depicted as a table row. For each request, we show its column values in a separate cell. For columns that were used for the default risk prediction, the corresponding cell is shaded depending on their effect on the prediction. We chose a heatmap-like representation as it supports counterfactual reasoning through comparison of loan requests [58]. The direction and strength of the effect is given by the Shapley value. A red shading indicates a positive effect (increases the expected value) while a blue one a negative effect (decreases the expected value) on the ML prediction. The opacity of the shading indicates the strength of the effect. Details about the strength are provided in a tooltip when the user hovers the cell. For example in Figure 1, the fact that request #1536 has 3+ dependents decreases the expected value of 32% by 7 percentage points. We reviewed research on explanation design approaches that foster user understanding. In general, the design of explanation facilities should follow the guidelines of *contrastive*, *selective*, and *interactive* explanations [32]. Our explanation style is similar to the *input influence* explanations in [6] where each feature value is accompanied by the direction and strength of its effect on the prediction. Prior work reported that providing users with interactive explanation facilities improved their subjective and objective model understanding [12]. These mechanisms informed the designs of our explanation facilities as follows: (i) *contrastive*: we show multiple instances and their respective explanations at once so that users can contrast a local explanation with local explanations of other instances. Further, users can sort the data by columns to contrast instances with equal values to spot regularities; (ii) *selective*: we excluded neglectable feature values with absolute effects of less than one percentage point from the explanation; (iii) *interactivity*: following the call for more interactive explanation interfaces that *"allow users to explore the system's behavior freely"* [1], we provided participants with two basic interactive functionalities: (a) to *resample* a different set of 16 loan requests to get a more holistic understanding of the ML model behavior[8] and (b) to *simulate* a prediction for a hypothetical loan request with user-defined features values [12]. Figure 1 shows the final explanation interface from a participant's perspective.

5 METHODS

We pre-tested and iterated our scenario, apparatus, and procedure with 10 people to ensure they are comprehensible from a participant perspective. We applied a mixed methods approach. First, we moderated 40 participants through the study (6 of them followed a think aloud protocol to not bind cognitive capacities). Second, we conducted an unmoderated study with 107 crowd workers. Following [45], we describe our participants as educated lay users of XAI. We used a combination of moderated and unmoderated studies to account for dual process model of human reasoning [24, 58]. For the moderated study, the presence of a moderator motivated

[7] 4 requests for the 4 different combinations of predicted and actual values, i.e., *true positives*, *false positives*, *true negatives*, and *false negatives*
[8] again sampled according to the confusion matrix

participants to invest more resources and apply high-effort rational thinking throughout the procedure (*system 2 thinking*). There, we used a slightly shorter procedure to qualitatively investigate what users do to form their mental model of the global ML model behavior. In contrast in the unmoderated study, we assumed participants to be guided more by low-effort heuristic thinking (*system 1 thinking*).

5.1 Participants

5.1.1 Moderated Study (N=40). We recruited 40 participants via our internal university mailing list. All participants were supervised by a moderator during the study to ensure participants understand and follow the instructions. We randomly selected a subset of 6 participants to additionally follow a think aloud protocol. We used a subset as the think aloud puts additional cognitive load on the participants and might *"impact how people perform on cognitively-demanding tasks"* [8]. 17 participants self-identified themselves as female and 23 as male. Of these, at the time of the study 65% aged 18-24 years, 32.5% aged 25-34 years and 2.5% in the age of 35-44 years. Among the participants, 20 (50%) hold a high school degree, 10 (25%) an undergraduate degree, 8 (20%) a graduate degree, while 2 had other educational backgrounds. On average, participants took 37.8 minutes (SD=10.1 minutes) to complete the study and were compensated 10 EUR per completion. 29 (72.5%) participants disagreed and rather disagreed to have practical knowledge of AI (e.g. application of statistical learning methods or training of machine learning models), 8 agreed or rather agreed, while 3 were undecided. 29 (72.5%) agreed to or rather agreed to frequently explain complex things to other people (e.g. seminar contents to fellow students or smartphone features to friends), 11 were undecided. 19 (47.5%) participants stated they use spreadsheet applications at least weekly, while 21 used them once a month or less.

5.1.2 Unmoderated Study (N=107). We recruited participants via the crowd sourcing platform *Prolific*. The posting included a short description about the study, the expected duration, and the compensation. We only recruited workers with a 100% approval rate and at least 10 previous submissions. Further, we required all participants to hold at least an undergraduate degree. 116 participants started the study of which 8 only partly finished it. We screened the answers of all completed sessions and excluded 1 participant due to low quality verbalization that was most likely generated by a bot. Participants' demographics were quite diverse. 48 participants self-identified themselves as female and 59 as male. Participants were located in the United Kingdom (42), Portugal (13), the United States (10), and other countries (42). At the time of the study, 22.5% of participants were aged 18-24 years, 49.5% aged 25-34 years, 18.4% aged 35-44 years, and 9.6% 45+ years. Among the participants, 57.2% stated they hold an undergraduate degree, 35.9% a graduate degree, 2.9% a PhD, and 3.8% stated other as highest educational level. On average, participants took 28.5 minutes (SD=15.8 minutes) to complete the study and were compensated £3.75 per completion (=£7.09/hour). 68 (63.5%) participants disagreed and rather disagreed to have practical knowledge of AI, 25 agreed or rather agreed, while 14 were undecided. 81 (75.6%) agreed to or rather agreed to frequently explain complex things to other people, 12 (11.2%) were undecided, and 14 disagreed or rather disagreed (13.2%). 65 (60.7%) participants stated they use spreadsheet applications at least weekly, while 42 used them once a month or less.

5.2 Procedure

The goal of our user studies is to investigate if an IOED can be observed when end users are exposed to Shapley explanations of an ML model. For this purpose, we query the participants' model comprehension through different tasks and repeatedly measure their self-assessment of perceived understanding using a uniform scale. Our procedure was inspired by the study designs of the initial IOED studies [46] but adjusted to the XAI context. Participants used the apparatus described in section 4 to complete the five tasks illustrated in Figure 2. The moderated user studies were conducted via video conferencing to observe how users interact with the apparatus. We describe the stages below:

Introduction. After consenting with the participation and data processing information, participants reported their demographics (i.e., age, gender, and educational background), their frequency of use of spreadsheet applications, their frequency of giving explanations about complex topics to others, and their level of practical experience in the field of AI, (the last three questions were illustrated with example statements and rated on 5-point Likert scales). Next, we explained in multiple steps the crowdlending scenario, the "AI-enabled prediction" of the default risk, the explanation facility, and the scale self-rating scales. In the moderated study, participants were encouraged to ask clarifying questions to the moderator.

Task 1: Exploration of Black Box (only used in unmoderated study). Participants were presented with a table of 16 observations. For each observation the ML prediction was presented without any explanation. Participants were asked to spend 5 minutes and *"try to understand how the AI forms its default risk predictions"*. Afterwards, they were asked to rate their perceived understanding. To give them an indication, a timer showed how much time they already spent on this task. We used this task in the unmoderated study to ensure that our explanation interface was perceived to improve understanding of participants.

Task 2: Exploration of Explanation Facility. Next, we provided participants with the explanation facility presented in section 4.2. We asked them to freely explore the decision-making behavior of the prediction model for no longer than 10 minutes and re-rate their gained understanding. To give them an indication, a timer showed how much time they already spend on this task.

Task 3: Verbalization of Understanding. According to psychological research deliberate self-explanation results in a more realistic assessment of a user's own understanding and may potentially refine it [22, 35]. It does not matter whether the self-explanation is self-motivated or prompted by an instructor [22]. Further, retrospection techniques such as (self)-explanation, can provide rich information about a user's mental model [22]. Thus, as a next step, participants had to write a detailed explanation of their global understanding of the ML model's prediction behavior. Their explanation was to be between 50 and 100 words long and address three guiding questions. After the participants verbalized their understanding, they re-rated their perceived understanding.

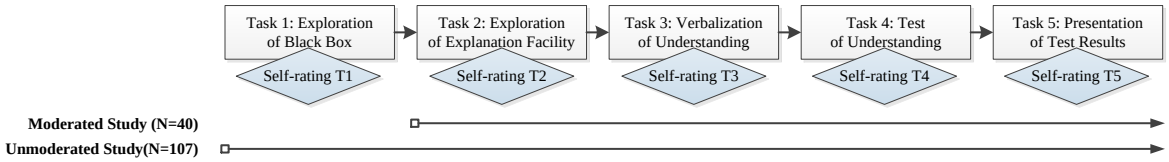


Figure 2: Stages of the procedure in the moderated and unmoderated study. First, we observe in task 1 (only in unmoderated study) and task 2 what end users do to form their mental models of the global ML model behavior. Second, we assess in tasks 3 to 5 what end users think they know about the behavior of an ML model in relation to what they actually know. Through multiple tests of comprehension, we assess how stable their self-reported understanding is if users need to put it into action.

Task 4: Test of Understanding. For the diagnostic questions, we based our questions on prediction tasks, where the participants had to simulate the prediction of the ML model for given sets of features. Afterwards, participants re-rated their perceived understanding.

Task 5: Presentation of Test Results. Rozenblit and Keil confronted their participants after the diagnostic questions with an expert statement [46]. In our case, we showed the participant's answers and contrasted them with the default risks predicted by the ML model. Further, we showed the Shapley explanations for each observation. We summarized their results as *"You predicted <n> out of 8 loan requests like the AI"*. This allowed the participants with incorrect predictions to re-examine the ML model behavior. Afterwards, participants re-rated their perceived understanding. Each session ended with a short questionnaire.

5.3 Dependent Variables

Self-Rating of Perceived Understanding. We used a uniform 7-point Likert scale that measures each participants' perceived understanding at multiple points throughout the study. We adopted the scale from the original IOED experiments and fitted it to the XAI context. To calibrate participants' usage of the scale, we demonstrated the scale during the introduction and provided explanations for levels 1, 4, and 7. On level 1, respondents think they can name features that the ML model has access to and what it predicts. On level 4, they think they understand the relative importance of individual features. At the highest level, level 7, they think they understand the absolute importance of individual feature values as well as possible interactions between them.

Objective Understanding. Following [12] and [59], a user "understands" an ML model *"if the human can see what attributes cause the algorithm's actions and can predict how changes in the situation can lead to alternative algorithm decisions"*. We built upon two question types from the explanation evaluation framework proposed by [12] to measure participants' objective model understanding. In total, we asked 8 questions (6x forward simulation, 2x relative simulation). For the first question type, we presented them with an observation and asked *"What do you think will the ML predict?"* (*forward simulation task*). We selected the observations according to the default risk predicted by the ML model: two at the extremes (0%, 100%), two with low risks (11%, 29%) and two with high risks (68%, 69%). We provided participant with five answer options of prediction ranges (from *0-20%* to *81-100%*). Following [22], participants had to rate their confidence for each prediction on a 5-point Likert scale (1=*very unconfident* to 5=*very confident*). As a second question type, we asked them to select the loan request with the highest (lowest in the second question) predicted default risk from a set of three given requests (*relative simulation task*). We offered three loan applications that differed in three (five in the second question) of the seven features that had on average a medium to low effect. The ML prediction of the correct option differed by at least 30 (66 in the second question) percentage points from the other options. Again, they had to rate their confidence in their simulation. We counted the number of correct answers and the mean deviation from the correct answer.

Demographics, Literacy, and Interaction. We asked participants on their age, gender, and level of education. Subject to participants' approval, we screen recorded their interactions in the moderated study. Further, we measured how much time participants spent at each step and logged their interactions with the explanation facilities (e.g., number of resamples and simulations). We used those measures as additional levels of control for analysis.

5.4 Design and Analysis

Both studies used a within-subjects design. Following the analysis in the original IOED experiments, we analyzed the differences in self-ratings through a *repeated measures ANOVA* [46]. None of the self-ratings of understanding were normally distributed. As a paired Student's t-test is not valid in such a case, we used a Wilcoxon signed-rank (WSR) test to analyze the planned linear contrasts for T1<T2, T3<T2, T4<T3, T5<T4 and T5<T2. If not stated otherwise, we based our significance at α=.05.

6 RESULTS

6.1 Robustness of Perceived Understanding

To answer RQ1, we present the distribution of participants' self-ratings throughout the moderated and unmoderated studies (see Figure 3). For comparison with the original IOED studies, the differences in the reported understandings were significant across the stages (repeated measures ANOVA: $F(4,424)=28.260$, $p<.001$, $\eta_p^2=.21$)

Shapley Explanations Increased Self-Ratings (T1<T2). In the unmoderated online study, participants reported on average rather high understanding levels even without explanations of the ML model (median=4, mean=4.33). 53 participants increased their understanding by at least one level after being exposed to Shapley

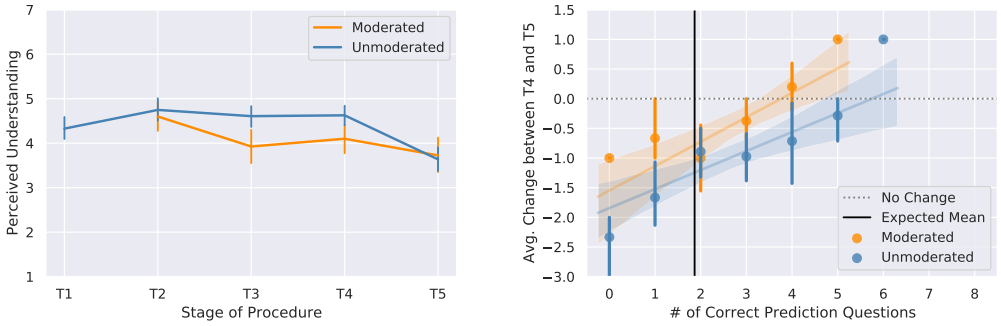


Figure 3: ***(left)*** **The means of self-ratings throughout the procedure for the moderated and unmoderated studies. In the moderated setting, we observed two large drops, one after the verbalization in T3 and another after the presentation of test results in T5. In the unmoderated setting, the drops remained insignificant until the last stage.** ***(right)*** **The average change in self-ratings between T4 and T5. After the participants saw their test results, most of them downgraded their perceived understandings.**

Table 1: The left side shows the mean and standard deviation of participants' self-ratings of understanding in the moderated and unmoderated studies. The right side presents the number of participants that *decreased* (*increased* for T1<T2) their self-rating by at least one level (#) and the results of our hypotheses tests using non-parametric Wilcoxon signed-rank test (w). The significance levels are reported as following: * p<.05; ** p<.01; *** p<.001

		T1	T2	T3	T4	T5		T1<T2	T3<T2	T4<T3	T5<T4	T5<T2
Moderated Study	Mean		4.60	3.95	4.10	3.73	#		19	7	18	25
(*N = 40*)	SD		1.06	1.21	1.03	1.28	w		6.5***	150.5	66.0**	76.0***
Unmoderated Study	Mean	4.33	4.75	4.61	4.63	3.64	#	53	27	20	72	77
(*N = 107*)	SD	1.30	1.33	1.25	1.12	1.38	w	2055.5**	408.5	384.0	358.0***	564.0***

explanations. Across all participants, the average reported understanding increased significantly (median=5, mean=4.75, w=2055.5, p<.01). Thus, **H1** was supported and our explanation interfaces was at first perceived as valuable to participants.

Examination Decreased Participants' Self-Ratings (T5<T2). Most participants in both studies significantly (p<.001) decreased their perception of understanding over the course of the procedure: 63% of participants in the moderated study and 72% in the unmoderated. Thus, **H2** stating that participants fell for an IOED was supported. Below, we report the changes in the self-ratings at individual stages of the procedure. *Verbalization (T3<T2):* In the original IOED studies, deliberate self-explanations decreased the perceived understanding. In our moderated studies, 48% of participants decreased their rating at this stage. The drop was significant. In the unmoderated online study, we observed a drop for only 25%. The drop was not significant. *Test of Understanding (T4<T3):* Participants remained confident in their understanding during the prediction tasks. Only, 19% decreased their rating in the unmoderated setting, compared to 18% in the moderated study. The drops were not significant. Contrary to our expectations, the prediction tasks increased the perceived understanding in the moderated study. *Test Results (T5<T4):* Confronting participants with their results of the prediction tasks caused a significant drop in understanding in both studies. In the unmoderated study, 67% decrease their understanding compared to the previous stage. In the moderated study, 45% did so. The drops in both studies were significant.

Moderated Participants Devoted More Resources. Participants in the moderated setting spent significantly more time on the study tasks than in the unmoderated setting. In the moderated study, participants spent on average 9.8 minutes (SD=4.9) exploring SHAP-Table, 10.9 minutes verbalizing their understanding (SD=4.2), and 7.1 minutes solving the prediction tasks (SD=3.0). In contrast in the unmoderated study, they spent only 3.8 minutes (SD=2.6), 6.7 minutes (SD=4.9), and 3.3 minutes (SD=1.9).

Moderated Participants Performed Better in Test of Understanding. We analyzed the number of correct predictions and the mean error of participants' predictions. The mean error describes the average number of bins between the participant prediction and the AI prediction over all questions (e.g., error between "0-20%" and "41-60%" is 2). On average, participants answered 2.85 (SD=1.05) questions correctly in the moderated and 2.66 (SD=1.20) questions in the unmoderated study. Both are significantly better than a random guess (expected mean) that would result in 1.86 correct questions. Further, on average, the mean error of participants in the moderated study (1.07, SD=0.29) was significantly lower compared to participants in the unmoderated study (1.22, SD=0.37). Both are significantly better than a random guess (expected mean) that would result in a mean error of 1.7 (see Figure 4).

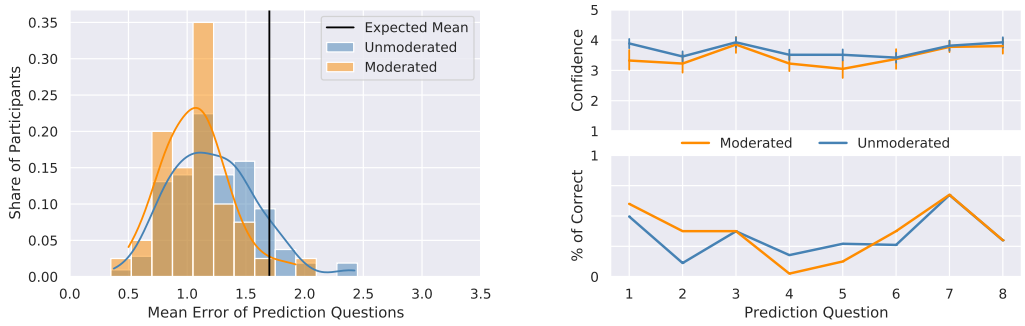


Figure 4: *(left)* The proportion of participants by their average distance to the correct answer (*mean error*). Participants in the moderated study were significantly closer to the correct answers than their unmoderated counterparts. The average confidence *(top right)* and share of correct answers *(bottom right)* for each prediction question in T4. On average, participants in the unmoderated study were more confident throughout the procedure.

6.2 How Users Formed Their Understanding

To answer RQ2, we report observations gained from the 6 think aloud protocols. We revisited the screen and audio recordings and openly coded recurrent themes during participants' interaction with SHAPTable. ***Orientation***: Participants used the coloring in SHAPTable to gain a first overview. They visually looked for inconsistencies in the heatmap. To calibrate their understanding of the coloring and the associated feature contributions, participant TA1 studied multiple tooltips. TA2 looked for *"global heuristics that always apply"* by shifting the attention from one feature to another (*column-wise comparison*). TA4 used a combination of sorting and rapidly resampling *"to look for [visual] patterns"*. Soon he stated that *"credit history correlates with the prediction without dependencies"*. After some resamples, the participant spotted an outlier that violated this hypothesis. TA1 identified an outlier in the heatmap where an effect was unusually strong. By this the participant realized that there are interactions in place. This discovery served as a starting point for deeper analysis. ***Analysis***: Single outliers guided the reasoning process of most participants. After they visually spotted one in the heatmap, they often replicated an observation in the simulation feature to *"live edit"* single feature values to understand their contributions. Further, participants often performed *pairwise comparisons* between two observations to understand differences. ***Abstraction***: All participants realized that interactions are present, but often over- or underestimated their impact. If they stumbled upon effects that violated their prior beliefs (e.g., that fewer dependents decrease the default risk), they searched for anecdotal memory aids for what they saw. Sometimes these were built from fragmented insights consisting of few features (e.g. *"self-employed in rural areas are high risk. That does not make sense."*) and missed that another feature (e.g. gender) had an impact too. Some participants stated it was difficult for them to assess when they should generalize from outliers and when not. Also, some participants assumed monotonic features effects (e.g., the effects of 0 vs. 2 dependents). If they found cases that violated this assumption, they judged the AI behavior *"as illogical"*. During verbalization, some participants recovered the effects of feature values from their memory aids and from the colors they remembered. ***Additional functionalities***: Some participants wished for aggregated *"scenarios"* consisting of similar observations (e.g. combinations of feature values that have consistent effects) and examples that illustrate interactions between features for easier orientation. TA1 and TA2 wished for an improved sorting feature that allows sorting by SHAP values to group observations with similar effects close to each other to identify regularities. TA1 wished for multiple rows in the simulation feature to simultaneously explore multiple combinations at once. Further, he wished to duplicate one observation into the simulation feature for improved usability. ***Reflection***: Participants perceived the study procedure as valuable. For example, participant TA1 considered the study procedure as a feedback loop that helped *"to learn from mistakes and expose my misconceptions [about the ML model behavior]"*. TA4 would have liked to complete the cycle multiple times to refine their insights: *"If I were to do this task again, I would gain a much better understanding."*

7 DISCUSSION

With a moderated and unmoderated study, we examined if and why an illusion of explanatory depth (IOED) emerges when non-technical users of XAI are exposed to local Shapley explanations. Our results indicate that participants overrated their understanding of the ML model behavior after freely exploring it with SHAPTable. On average, participants in both studies significantly decreased their perceived understanding throughout the procedure. What differed were the stages at which the drops occurred. In the moderated setting, we observed two large drops. One after the self-explanation stage (48% decreased their self-rating by at least one level) and another after the presentation of test results (45%). In the unmoderated setting, the self-ratings of participants remained mostly unchanged until the last stage. After they had seen and analyzed their test results, 67% decreased their self-rating. The IOED was more pronounced for participants in the unmoderated study. They spend significantly less time at each stage and had a significantly narrower objective understanding according to our prediction tasks. Still, on average they were more confident about the correctness

of their prediction questions. The magnitude in the decrease in self-ratings in the last stage depended on the number of correct predictions and was stronger in the unmoderated setting. It seems that participants in the unmoderated study expected more correct answers of themselves. While moderated participants with 4 out of 8 correct answers refrained from downgrades, unmoderated participants downgraded it even with 5 out of 8 correct questions. We interpret that participants in the unmoderated setting were guided by heuristic thinking and did not realize the incompleteness of their understandings until they saw their test results. We believe, they were less aware of irregularities of feature values effects and feature interactions than participants in the moderated setting. Overall, 85% of participants in the moderated and 69% in the unmoderated study agreed or agreed completely that the study procedure *"helped me to better assess my own understanding of the AI prediction behavior"*.

Humans will most likely never be able to correctly predict the behavior of complex non-linear ML models. Our results highlight the importance of XAI systems to not only provide non-technical users with static justifications, but also guiding user interactions that support them in building an accurate mental model – even if this means exposing complexities and irregularities of the ML model behavior. Otherwise, providing them with seemingly simple local justifications of complex ML behavior (as with Shapley values) may leave them with an *"easiness effect"* [50]. Below we discuss implications for the design of XAI systems derived from our findings and outline its limitations.

Calibrating Understanding as Part of XAI Interaction: An effective XAI system need to capture a wrong or incomplete mental model of the user and adjust its explanations accordingly [48]. An implication for XAI designers is that calibrating user perception of understanding through a structured procedure, as outlined in our studies, might expose that the system is more complex than it seems at first. For example, Cai et al. [10] described the onboarding phase to an XAI system as a key phase that forms users' initial impressions of an XAI system. It is during the onboarding that users form their mental model of the capabilities and limitations of the XAI system. Deliberate self-explanation has been proposed as being an effective way to calibrate XAI understanding [22, 35]. However, our results indicate this is only the case if users are willing to devote the required cognitive capacities. Buccinca et al. [8] describe cognitive forcing strategies, such as forcing users to form an own prediction before being confronted with the AI prediction. Our multi-stage procedure extends this idea in a playful way. Future work could explore how to leverage the individual results of such procedures to automatically learn about the mental model of the user and personalize explanations during the interaction with the XAI.

Forming (Global) Rationales from Local Explanations: Like [3], our results indicate that participants had difficulties in abstracting their local insights to a global understanding. They understood the justifications provided for individual observations but struggled to assess how representative they were for the average model behavior. The properties of SHAP enable novel ways for interactivity [13, 43] to provide selective, contrastive, and interactive explanations [32] that might bridge the gap between local and global understanding [42]. Future research could explore how to condense multiple local explanations into accessible higher order explanations to contextualize them. Such novel ways of interactivity could support the interpretation strategies applied by participants in our studies. This resonates with the concept of *rationales* [15, 17]. These aim to provide end users with contextually appropriate reasons for an ML prediction in natural language.

Limitations. There are several limitations to our studies. First, we examined a simplified extrinsic [38] scenario around a tabular data set. Thus, the external validity beyond this scenario (i.e., different decision-making situation) and type of data (i.e., visual data or natural language data) is uncertain. Second, the emergence and strength of an IOED might highly depend on the target audience. Physicians and risk managers may have very different reasoning strategies than the educated lay users in our studies. Future work could investigate different extrinsic as well as intrinsic [38] scenarios with varying ML model complexities or XAI methods. Still, we are confident that our insights highlight the importance of keeping cognitive biases in mind when designing and deploying XAI.

8 CONCLUSION

With XAI systems expected to be deployed deeper into organizations and society, it is important to understand how non-technical users of XAI consume explanations. In this work, we examined how non-technical XAI users form their mental model of the global ML behavior. Our results indicate that users overestimate the understanding they gain because of the illusion of explanatory depth. Further, we describe reasoning and interaction strategies that users applied. Future work could investigate how these strategies can be included into interactive explanation facilities to make them aware of potential fallacies and to support their reasoning. We offer starting points for XAI designers on how to support non-technical users to form a more appropriate mental model of ML model behaviors.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

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

[3] Ahmed Alqaraawi, M. Schuessler, P. Weiß, E. Costanza, and N. Bianchi-Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020).

[4] Adam L. Alter, Daniel M. Oppenheimer, and Jeffrey C. Zemla. 2010. Missing the trees for the forest: a construal level account of the illusion of explanatory depth. *Journal of personality and social psychology* 99 3 (2010), 436–51.

[5] Umang Bhatt, Alice Xiang, S. Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and P. Eckersley. 2020. Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[6] Reuben Binns, M. V. Kleek, M. Veale, Ulrik Lyngs, Jun Zhao, and N. Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. *ArXiv* abs/1801.10408 (2018).

[7] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[8] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[9] Adrian Bussone, S. Stumpf, and D. O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics* (2015), 160–169.
[10] C. J. Cai, S. Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3 (2019), 1 – 24.
[11] N. Chater. 1999. The Search for Simplicity: A Fundamental Cognitive Principle? *Quarterly Journal of Experimental Psychology* 52 (1999), 273 – 302.
[12] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 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* (2019).
[13] Michael Chromik. 2020. reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP. In *Proceedings of 18th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).
[14] Dennis Collaris, Leo M. Vink, and Jarke J. van Wijk. 2018. Instance-Level Explanations for Fraud Detection: A Case Study. *ArXiv* abs/1806.07129 (2018).
[15] Devleena Das and S. Chernova. 2020. Leveraging rationales to improve human task performance. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (2020).
[16] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretability. *CoRR* abs/1702.08608 (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608
[17] Upol Ehsan, Pradyumna Tambwekar, L. Chan, B. Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
[18] Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *ArXiv* abs/2004.07780 (2020).
[19] Joey F. George, Kevin Duffy, and Manju K. Ahuja. 2000. Countering the anchoring and adjustment bias with decision support systems. *Decis. Support Syst.* 29 (2000), 195–206.
[20] Alicja Gosiewska and Przemyslaw Biecek. 2020. Do Not Trust Additive Explanations. arXiv:1903.11420 [cs.LG]
[21] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (aug 2018). https://doi.org/10.1145/3236009
[22] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* abs/1812.04608 (2018). arXiv:1812.04608 http://arxiv.org/abs/1812.04608
[23] Hsieh-Hong Huang, J. Hsu, and Cheng-Yuan Ku. 2012. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decis. Support Syst.* 53 (2012), 438–447.
[24] D. Kahneman. 2011. Thinking, Fast and Slow.
[25] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219
[26] Tae-Nyun Kim and Hayeon Song. 2020. The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[27] A. Lau and E. Coiera. 2009. Research Paper: Can Cognitive Biases during Consumer Health Information Searches Be Reduced to Improve Decision Making? *Journal of the American Medical Informatics Association : JAMIA* 16 1 (2009), 54–65.
[28] Rebecca Lawson. 2006. The science of cycology: Failures to understand how everyday objects work. *Memory and Cognition* 34 (2006), 1667–1675.
[29] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages. https://doi.org/10.1145/3236386.3241340
[30] Scott M. Lundberg, G. Erion, Hugh Chen, Alex J. DeGrave, J. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2 (2020), 56–67.
[31] J. McGuirl and N. Sarter. 2006. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors: The Journal of Human Factors and Ergonomics Society* 48 (2006), 656 – 665.
[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1 – 38. https://doi.org/10.1016/j.artint.2018.07.007
[33] Candice M. Mills and Frank C. Keil. 2004. Knowing the limits of one's understanding: the development of an awareness of an illusion of explanatory depth. *Journal of experimental child psychology* 87 1 (2004), 1–32.
[34] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *CoRR* abs/1811.11839 (2018). arXiv:1811.11839 http://arxiv.org/abs/1811.11839
[35] Shane T. Mueller, Robert R. Hoffman, William J. Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *CoRR* abs/1902.01876 (2019). arXiv:1902.01876 http://arxiv.org/abs/1902.01876
[36] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and B. Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116 (2019), 22071 – 22080.
[37] Gregory L. Murphy and Douglas L. Medin. 1985. The role of theories in conceptual coherence. *Psychological review* 92 3 (1985), 289–316.
[38] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1802.00682 (2018). arXiv:1802.00682 http://arxiv.org/abs/1802.00682
[39] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
[40] M. Nourani, Donald R. Honeycutt, Jeremy E. Block, Chiradeep Roy, Tahrima Rahman, Eric D. Ragan, and V. Gogate. 2020. Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).
[41] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0
[42] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines* (2019), 1–19.
[43] Shubham Rathi. 2019. Generating Counterfactual and Contrastive Explanations using SHAP. arXiv:1906.09293 [cs.LG]
[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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* (2016).
[45] Mireia Ribera and Àgata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops*.
[46] Leonid Rozenblit and Frank C. Keil. 2002. The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive science* 26 5 (2002), 521–562.
[47] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[48] Heleen Rutjes, M. C. Willemsen, and W. IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019*.
[49] James Schaffer, J. O'Donovan, James Michaelis, A. Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019).
[50] Lisa Scharrer, Yvonne Rupieper, M. Stadtler, and R. Bromme. 2017. When science becomes too easy: Science popularization inclines laypeople to underrate their dependence on experts. *Public Understanding of Science* 26 (2017), 1003 – 1018.
[51] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.
[52] L. Skitka, K. Mosier, M. Burdick, and B. Rosenblatt. 2000. Automation Bias and Errors: Are Crews Better Than Individuals? *The International Journal of Aviation Psychology* 10 (2000), 85 – 97.
[53] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and H. Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020).
[54] Kacper Sokol and Peter A. Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).
[55] Jacob Solomon. 2014. Customization bias in decision support systems. In *CHI '14*.
[56] Richard Tomsett, Dave Braines, Dan Harborne, A. Preece, and S. Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *ArXiv* abs/1806.07552 (2018).
[57] Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Commun. ACM* 62 (2019), 104 – 110.
[58] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.
[59] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62 (2019), 70 – 79.
[60] Andrew Zeveney and Jessecae Marsh. 2016. The Illusion of Explanatory Depth in a Misunderstood Field: The IOED in Mental Disorders. In *CogSci*.

ML for UX? - An Inventory and Predictions on the Use of Machine Learning Techniques for UX Research

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Florian Lachner*
LMU Munich
Munich, Germany
florian.lachner@ifi.lmu.de

Andreas Butz
LMU Munich
Munich, Germany
butz@ifi.lmu.de

ABSTRACT

Machine learning (ML) techniques have successfully been applied to many complex domains. Yet, applying it to UX research (UXR) received little academic attention so far. To better understand how UX practitioners envision the synergies between empathy-focused UX work and data-driven ML techniques, we surveyed 49 practitioners experienced in UX, ML, or both and conducted 13 semi-structured interviews with UX experts. We derived an inventory of ML's impact on current UXR activities and practitioners' predictions about its potentials. We learned that ML methods may help to automate mundane tasks, complement decisions with data-driven insights, and enrich UXR with insights from users' emotional worlds. Challenges may arise from a potential obligation to utilize data and a more restrictive access to user data. We embed our insights into recent academic work on ML for UXR and discuss automated UX evaluation as a promising use case for future research.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; *User studies*; *Usability testing*.

KEYWORDS

User experience research; UX research; machine learning

ACM Reference Format:
Michael Chromik, Florian Lachner, and Andreas Butz. 2020. ML for UX? - An Inventory and Predictions on the Use of Machine Learning Techniques for UX Research. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI '20), October 25–29, 2020, Tallinn, Estonia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3419249.3420163

1 INTRODUCTION

In recent years, many enterprises shifted their priorities from purely focusing on efficient production and distribution to creating memorable customer experiences. In this way, they hope to differentiate themselves from competitors and establish a competitive edge. This shift towards an *"experience economy"* [33] made the *user experience* (UX) a primary design goal. The term UX refers to *"a person's perceptions and responses that result from the use or anticipated use of a product, system, or service"* [4]. UX provides a holistic perspective and encompasses a person's emotions, feelings, and thoughts that may be formed before, during, or after the interaction [25, 35].

Building on this notion, the discipline of *UX research (UXR)* aims to understand and design people's experiences from end to end. UXR has emerged as an interdisciplinary field with influences from various disciplines such as cognitive science, psychology, and engineering. Each discipline contributes different terminologies, methods, and technologies to it. UX researchers frequently utilize qualitative methods, such as semi-structured interviews or surveys, while data-driven quantitative approaches are currently still less common [32]. The rare use of data-driven approaches by UX researchers is surprising, given the increasing data volumes in many contexts. Especially bigger enterprises increasingly compete in a data-driven environment and try to embrace the *"age of analytics"* [15].

Fueled by the availability of large data sets and affordable computing resources, *machine learning* (ML) methods have successfully been applied to complex problems in various domains. Historically, most academic research on ML within the HCI community had a technical focus on how to improve the interaction with systems (e.g., through adaptive interfaces) or develop new modes of interaction (e.g., gesture and voice interfaces) [6]. In the opposite sense, HCI academics have started to investigate how designers can enhance the user experience of ML-powered intelligent systems (*"human-centered machine learning"*) [11, 24] and how to address the distinct challenges of *human-AI interaction* [1, 46] from a *user-centric* perspective. However, there is little academic discourse that takes a *UX practitioner-centric* perspective and examines how ML methods could be leveraged to enhance the UX activities themselves. This lack of discourse may result from ML not yet being a standard part of the UX design practice as no relevant design patterns or prototyping tools have emerged yet [6]. Even if UX practitioners had previous exposure to ML, they often miss opportunities to use it in their design practice [45]. A review of HCI literature that employs ML observed that academics frequently resort to convenient interaction and design choices [44]. Thus, there may be a lack of awareness that the actual UX research and design processes may also benefit from ML.

To better understand the perception in the field, we have focused on practitioners to identify promising directions for the application of ML to UXR. We followed a two-pronged approach consisting of two independent studies to complement our insights from multiple angles. We surveyed 49 practitioners from the fields of ML and UX. Furthermore, we conducted 13 semi-structured interviews with

*now at Google

NordiCHI '20, October 25–29, 2020, Tallinn, Estonia

ACM ISBN 978-1-4503-7579-5/20/10…$15.00
https://doi.org/10.1145/3419249.3420163

UX practitioners who were educated or had experiences at the intersection between UX and ML. Our work contributes to the HCI research community in two ways: (1) We provide insights from two studies on ML's impact on current UX practices and ML's potentials for UXR. (2) We present data-driven UX evaluation using ML as a promising direction for future research and link it to recent academic work.

2 RELATED WORK

2.1 Terminology When using ML for UX

In the so-called *"experience economy"* [33] people use technology not only to accomplish a given task (i.e., for its *pragmatic quality*), but also to enjoy doing so (*hedonic quality*) [12]. The combination of both qualities constitutes the *user experience* (UX), i.e., the overall quality of a human's interaction with an interactive system. The field of UX covers an entire spectrum between the investigation to find user problems worth being addressed (*UX research*) and the creation of relevant interactions that provide a specific experience (*UX design*) [20].

ML refers to *"a set of methods that can automatically detect patterns in data [...] to predict future data, or to perform other kinds of decision making under uncertainty"* [34]. ML methods have successfully been applied to complex problems in a variety of domains such as spam detection, speech recognition, autonomous systems, and games. From a technical perspective, ML is typically split into *supervised* learning methods, which focus on predictions based on labeled training data, *unsupervised* learning methods, which find relationships in unlabeled data, and *reinforcement learning*, which optimizes some notion of reward by interacting with an environment. *Generative learning* methods create new contents such as texts or images. Approaching ML from a user-centered perspective, Yang et al. distinguish four channels of how it might generate value for users: inferring insights about an individual user, inferring insights about the context and external world (e.g., time, place, or social connections), inferring knowledge about how to optimize some arbitrary metric, and enabling entirely new user capabilities (utility) [44].

Combining the practices of UX and ML may yield positive effects in both directions: On one hand, knowledge in many domains is not only encapsulated in data, but also in the implicit expertise of human domain-experts. UX practice plays a key role in designing interfaces for those experts to effectively teach an ML model. In this way, UX decisions may have an impact on the model performance and robustness in the field (*interactive machine learning*) [7, 28]. On the other side, conversational UI and other forms of intelligent user interfaces offer new possibilities for UX design. Some observers claim that ML might become the most important design element to enhance user experiences by automatically personalizing systems to users and contexts (*"ML is the new UX"*) [47]. However, it has been observed that UX practitioners face challenges in understanding the data dependencies of ML and lack the tools to properly prototype with it [6, 43, 47].

2.2 Using ML for UX Research

Our work focuses primarily on *UX research* side of the spectrum. The goal of UXR is to systematically gather and analyze user data to understand a problem space and guide the entire design process. It is primarily applied at the generative and evaluative stages of the design process [9]. In the context of this paper, we subsume all empirical activities conducted by practitioners along the UX design spectrum as UXR. Building on the user-centered value channels of ML, using ML for UXR broadly falls under the *utility* channel [44]. ML indirectly benefits users through an improved UX if UXR practitioners can more effectively identify and validate user needs. A structured literature review by Yang et al. revealed that there is only little academic work at the intersection of UX and ML [44]. We found even less research that explicitly addresses ML for UXR. However, we noticed that the number of relevant publications has been increasing since 2015 and we expect that it will most likely continue to do so as ML is gaining popularity in many contexts. Below, we present some notable exceptions without claiming to be exhaustive.

Unlike conventional UXR approaches, that primarily generate new study data (e.g., through surveys or interviews), ML-based approaches were primarily used to enrich already collected user data. Most of this work analyzes *textual user data*. ML and natural language processing (NLP) methods have been used to semi-automate the coding of interview transcripts [29] and to extract UX-related problems from online review narratives through classification [13, 30, 40]. Data-driven learning approaches have also been used to construct *behavioral personas* derived from user clickstreams [48] and social media [19], and automatic real-time evaluation of usability and user experience via *emotional logging systems* using video-captured facial expressions in lab contexts [37] and on mobile devices [8], using acoustic data [36], and skin conductance signals [27]. Furthermore, ML was used for *selecting participants* for usability tests [10] and A/B tests [22].

2.3 Opposing Mindsets in UX and ML

Research on UX and ML originates from different academic communities. The relationship between the academic communities of HCI and AI has been discussed for decades. They tend to differ in their views of how humans and computers should interact with one another. These views can be roughly depicted along a spectrum of decreasing autonomy. While the HCI community values simplicity and user control, the sub symbolic fraction of the AI community favors the power of data-driven inference and convenience for the user. Winograd [42] argues that these views result from an opposing understanding of people and how technology is created for their benefit. He distinguishes two opposing approaches that exist in both communities. The *rationalistic approach* tries to depict the world through a quantitative or formal logic and tries to optimize the interaction accordingly. In contrast, the *design approach* acknowledges the complexities of the human world and tries to account for the limitations of modeling it. Instead, this approach focuses on the pragmatic interaction between a human and her or his environment.

Similarly, the UX mindset emphasizes the exploration of the desired future to be designed (design approach) while the ML mindset settles to accurately predict the future given data from the past (rationalistic approach) [43]. Opposing mindsets are also

prevalent on the UX practitioners' side. UX has become increasingly cross-functional. Nowadays, many enterprises consider UX an organization-wide priority. This blurs the disciplinary boundaries between designers, developers, and marketers. UX teams often consist of experts from different disciplines [18]. UXR activities are seldom bundled in one role but often shared across the UX team. In practice, these teams must often cater to the needs of stakeholders with different mindsets: Colleagues with a design focus appreciate deep qualitative insights generated through user involvement. Additionally, business counterparts request aggregated quantitative insights to confirm their strategic decisions [26].

3 ONLINE SURVEY

3.1 Participants, Data Collection, and Analysis

With our survey, we intended to illuminate the impact of opposing mindsets on product development with a broader audience of ML and UX professionals. We were specifically interested in the differences between UX practitioners with and without experience in ML. In the last part of the survey, we examined if and how UX practitioners envision ML can be leveraged specifically for UXR activities. We designed a non-probabilistic self-selected survey that targets practitioners who have at least experience in either UX or ML, ideally both (inclusion criteria). Because the boundaries of UXR are fluid along the UX spectrum, we addressed a broader audience of UX professionals. We also assumed that ML developers are often involved at some stages of the UX process and could thus contribute valuable perspectives. The questionnaire consisted of 6 closed questions with ordered response options, 16 closed questions with unordered options and 4 open-ended questions. To understand the practitioners' contexts, we inquired about their demographics, educational background, working position and experience, and the qualitative and quantitative methods they apply regularly. Furthermore, we asked them to express their interpretation of the intersection between ML and UX and potential use cases for it. This way, we implicitly examined whether they could imagine potentials for UXR use cases. In the last part, we explicitly asked how they assessed the applicability, feasibility, and value of applying ML to different UXR use cases. The survey was designed according to the guidelines of the local institutional review board (IRB).

We pre-tested the survey with a few subjects to eliminate potential ambiguities and design flaws. We evaluated and incorporated their feedback into the final survey design. The survey was distributed through UX- and ML-related mailing lists of academic institutions in the United States and Germany as well as practitioner-oriented social media groups. Survey participants were self-selected and submitted their responses anonymously. As a reward for their participation, all respondents had the chance to take part in a lottery of three e-commerce vouchers worth 150 USD and two vouchers worth 60 USD. The survey was open for 4 weeks. 124 participants started the survey during this period. 19 participants did not meet the inclusion criteria. 56 participants did not finish the survey. After cleaning the data, we obtained 49 complete responses that met the inclusion criteria.

Respondents' demographics were quite diverse. 14 respondents self-identified themselves as female and 35 as male. Respondents are located in Germany (28), the United States (12) and other countries (9). 36 are working in the industry, 4 in academia, and 9 at the intersection of both. Their average work experience was 5.8 years (min=1, max=23 years). 17 respondents self-reported they have working experience only in UX (*UX-only*), 23 in UX and ML (*UX+ML*), and 9 respondents only in ML (*ML-only*). Most of UX-only respondents described their primary role as UX designer or UX consultant, UX+ML respondents as product manager, UX designer or UX researcher, and ML-only respondents as ML engineer/developer. 13 of the 17 UX-only respondents assessed their knowledge of ML as *basic* (familiar with the term and basic concepts) while 16 of the 23 UX+ML respondents consider their knowledge of ML as *advanced* (basic practical experiences) or *expert-level* (applied experience in the field of ML). All ML-only respondents assessed their knowledge as advanced or expert-level. In total, 3 respondents stated they are unfamiliar with ML (all in UX-only).

3.2 Findings

Our analysis of responses indicates that UX practitioners with ML experience have a different take on UX and more often leverage quantitative methods as part of their daily work. Most of the respondents believe that ML and UX will increasingly overlap in the future. Lastly, respondents consider the data-driven evaluation of UX as a promising use case for applying ML to UX research.

3.2.1 Current Project Involvement and Research Methods. Most of the respondents are involved in the pre-deployment stages of product development. There, the respondents work mainly on the initial development (e.g., wireframing, low-fidelity prototyping) and final development (e.g., high-fidelity prototyping, final product) of a product. In our sample, we see a trend that UX-only respondents are more often responsible for the conceptual stages such as product vision development or needs research (88% for UX-only compared to 43% for UX+ML). In contrast, UX respondents with ML experience are slightly more often involved in the final development stages (87% compared to 71%). Only half of the respondents (27 out of 49) are regularly involved in the evaluation of a product after its launch.

Overall, 30 out of 49 (61%) respondents apply qualitative and quantitative methods equally often as part of their daily work. However, this is only the case for 8 out of 17 (47%) UX-only respondents. 7 of them are mainly qualitative researchers. In contrast, 16 out of 23 UX+ML respondents (70%) apply both types of methods equally often. This trend is also reflected in different opinions on how UX should be approached. We asked participants about their agreement using a 6-point Likert-scale (1=disagree very strongly, 6=agree very strongly). 76% of UX-only respondents agree or agree strongly that UX must be approached qualitatively (compared to 57% of UX+ML respondents). Furthermore, 65% of UX-only respondents agree or agree strongly that UX can be quantified (compared to 87% of UX+ML respondents).

Respondents mostly agree on when to use qualitative methods. For qualitative methods, we observe large differences between UX-only and UX+ML respondents. When talking to their project stakeholders, half of the respondents argue with qualitative insights. However, 61% of UX+ML respondents leverage quantitative data to back their arguments while only 29% of UX-only respondents

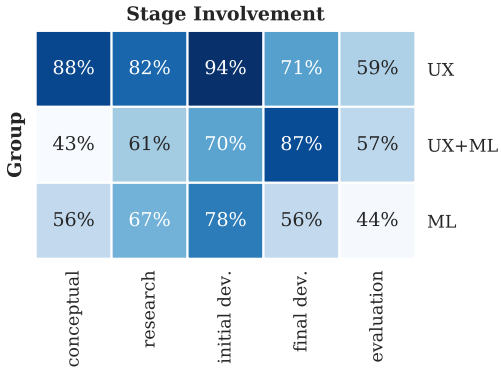


Figure 1: Respondents' involvements in the product development stages per group based on 49 respondents (17 UX-only, 23 UX+ML, and 9 ML-only). Each cell represents the percentage of respondents in the group that stated to be involved in the respective stage.

do so. Similarly, we observe differences when choosing between design options. Proportionally twice as many UX+ML respondents leverage quantitative methods to back their decisions in addition to qualitative methods (18% UX-only vs. 43% UX+ML). When it comes to individual research methods, such as semi-structured interviews or questionnaires, we see roughly equally often usage. Yet, we see a difference in leveraging user log data and online feedback. 35% and 41% of UX-only respondents apply these methods in at least half of their projects, respectively (compared to 74% and 74% of UX+ML respondents, respectively).

3.2.2 ML and UX Are Expected to Overlap in the Future. Respondents were asked to assess their current perception of the interplay between ML and UX, and how they predict it will evolve in the future. Assessing the status quo, only 9 respondents perceive ML and UX to overlap to some or great extent. However, 23 expect ML and UX to overlap at least to some extent in the future. In total, 35 out of 49 respondents think that the interplay between both disciplines will increase in the future. None of the respondents are expecting that the disciplines will drift apart (see Figure 2).

Next, we asked respondents to describe the perceived interplay in their words. We asked them to illustrate it based on a promising scenario from their daily work. We aimed to examine their perception of applying ML to UXR without directly asking them about it. We grouped the mentioned scenarios by use case: 19 respondents mentioned use cases that aim to improve the UX of products for users through ML features, 17 mentioned use cases that enhance UX research and design activities, 11 mentioned use cases about improving the UX of developing ML models, and 8 mentioned miscellaneous use cases (some respondents described multiple use cases). The UX research and design use cases included the use of ML *to reveal user insights* (6 mentions; e.g., trends in user behavior, analysis of user reactions, identifying plots in user study results), to *evaluate the UX of products* (6 mentions; e.g., automated measurement of UX, predicting the UX of new users, continuous UX monitoring) and *to augment the creation of UX artifacts* (5 mentions; e.g., producing variations of interaction flows, automating standard design tasks, informing design with historical data).

3.2.3 Leveraging ML for UX Research. Since we were interested in the opportunities for applying ML to UXR, we subsequently asked all participants about activities and use cases that are specifically related to UXR. We asked them to assess the potential per use case taking applicability, feasibility, and value into account. Respondents consider applying ML to yield insights from log and time series data, to remotely track user behavior over time, and user modeling as promising fields for future exploration (see Figure 3).

Furthermore, we asked which types of ML they had in mind when assessing the use cases: (1) prediction of an outcome based on the analysis of given data, (2) detection of patterns within structured or unstructured data, (3) generation of new outcomes or data, or (4) other. Most respondents think of scenarios for pattern detection and outcome prediction. UX-only respondents are more optimistic about the potentials of generative learning approaches. 41% of UX-only respondents consider them feasible and valuable. In contrast, UX+ML (13%) and ML-only (11%) respondents are much more conservative.

Next, we questioned for which stages of the product development process they perceive ML to be well-suited. We provided them with typical example activities for each stage. Respondents think that ML is especially applicable to later stages of the development process. 41 out of 49 believe ML can be applied to some or to a great extent to evaluate and test products after their development (e.g., UX evaluation of products on the market). On the other hand, few respondents can envision how ML can support the conceptual stage of product development (e.g., product vision or strategy development). The opinions tend to be divided for the stages of research (e.g., user research) and initial development (e.g., wireframing or prototyping) (see Figure 2). When comparing the results between the three groups, we observe that UX-only respondents have almost equal assessments for the first four stages. In contrast, UX+ML respondents have a more distinguished opinion. They see more potential in the initial as well as final development stages. The assessment of the UX+ML respondents is very much in unison to the assessment of the ML-only respondents.

4 EXPERT INTERVIEWS

4.1 Participants, Data Collection, and Analysis

We conducted semi-structured interviews with 13 UX experts from industry and academia to understand how they envision ML methods to enhance or influence their UX processes. We recruited experts from the fields of Human-Computer Interaction (HCI) and UX who are experienced with the concepts and applications of ML (either by professional collaboration with ML engineers or by education). As the intersection of UX and ML is a young field, we aimed for a mix of experienced senior professionals as well as young professionals (who were trained in both fields as part of their study program). Starting the recruiting through our academic network, we asked each participant to recommend experts who potentially meet our criteria for further interviews (*snowball sampling*). Our panel comprised mostly UX professionals from leading digital companies as

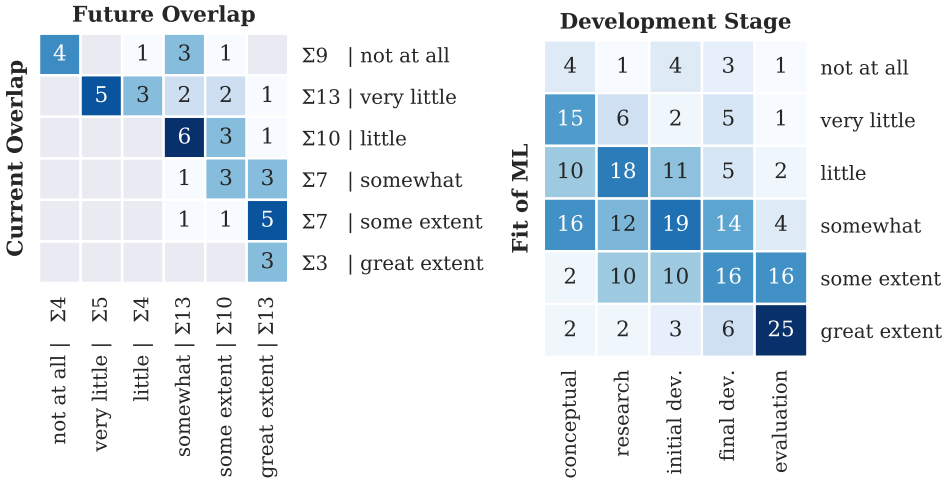


Figure 2: ***(Left)*** **Perceived overlap of ML and UX aggregated by the number of responses (N=49). Entries on the diagonal mean that no change is expected. Entries towards the upper right corner mean that the overlap is expected to increase (e.g., of the 13 respondents that currently see *very little* overlap, 8 respondents expect the overlap to increase in future).** ***(Right)*** **Respondents' assessment of how well ML techniques can be applied to the respective stages of the product development process (N=49 for each development stage).**

well as graduates from a relevant interdisciplinary study program at a renowned academic institution. Table 1 presents a summary of the participants' characteristics. The interviews were conducted in person or via video calls and lasted roughly forty-five minutes each. The sessions were recorded and transcribed to analyze them further (total recording time of 12 hours).

To understand the contexts of the participants, we asked them about their backgrounds, work routines as well as the importance of UX within their institution. We inquired about previous projects in which they had applied data logging to get a sense of their exposure to quantitative research methods and ML. Furthermore, we asked them to ideate how ML might enhance their UX method toolbox or enable novel ways of UX research. We asked them to ideate around a hypothetical ML system that automatically evaluates the UX of a user during interaction with a product based on usage data. Lastly, we asked them what challenges they thought stakeholders in the UX research process might face when applying ML-based methods, especially in terms of privacy and ethics.

For data analysis purposes, we transcribed the audio recordings from the expert interviews. Then we followed a *Grounded Theory*-inspired emergent coding approach, i.e., we analyzed without a guiding theory in mind. In a first step, one author extracted 120 UX and ML-related trains of thought from the interviews (each consisting of one to many sentences) and gave each observed phenomenon a distinctive name using mostly in-vivo codes. The author also identified connections between the codes and grouped them in multiple iterations into higher-level themes. Those themes are represented by the derived opportunities and challenges. In a second step, two authors independently coded the extracted trains of thoughts given the codebook of higher-level themes from step one. The inter-rater reliability was α = .8710 with 95% confidence in a CI of (0.8015, 0.9305). According to Krippendorff [23], values for α greater than .8 can be considered satisfactory. Typically, Grounded Theory (GT) starts from a set of empirical observations and aims to reverse-engineer a hypothesis from the observations in multiple iterations [38]. Our approach follows the GT approach in terms of open and axial coding. However, we are not formulating a (well-grounded) theory from our observations as we primarily aim to describe and group practitioners' opinions in terms of perceived challenges and opportunities.

4.2 Findings: Opportunities

Our analysis revealed 3 areas of opportunity along the dimensions of *automating*, *complementing*, and *enriching* the insight generation practices of UX researchers.

4.2.1 Automate the Mundane Part. ML is often perceived as a tool to free people from time-consuming and repetitive tasks of limited value. In this sense, our participants saw opportunities for ML to (semi)-automate parts of their current work routines. Furthermore, designing survey studies as a more engaging and personalized experience could result in richer user data and higher response rates.

Automated Transcription: ML-based speech-to-text services were hoped to significantly shorten the time between data collection and data analysis. This was considered particularly interesting for

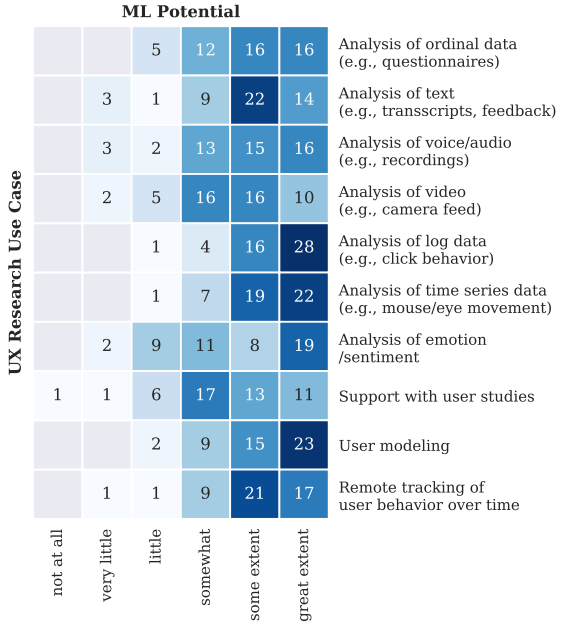


Figure 3: The distribution of respondents' assessments regarding the potentials for applying ML to UXR use cases taking applicability, feasibility, and value into account (N=49 for each use case).

the post-processing of contextual inquiries or interviews since *"you don't need to transcribe or code all the information"* (P11). Instead, *"you can give them [the ML systems] your audio file and they'll send you a typed-up version of it."* (P6). *"Instead of trying to scribble notes or record the conversation and then transcribe it, it's just doing it while you're in the field."* (P10). When ML takes over mundane tasks, UX researchers are enabled to focus on higher-value activities. *"AI can do things to make us faster at producing the kind of work that we want to do versus the kind of work we have to do"* (P1).

Engaging Surveys: Participants felt that ML can simplify survey studies for researchers and survey participants alike by leveraging the idea of adaptive user interfaces. Questionnaires might automatically be tailored in real-time to the individual survey participant based on their previous answers as well as answers of similar participants. The intend is *"to not give everybody 100 questions, but just give the 25 most important ones"* (P5). Also, advances in conversational and voice user interfaces were considered an opportunity for more empathetic survey studies. A questionnaire might be turned into engaging conversational experiences by *"acting like it's a person, giving it a personality"* (P10). Thereby researchers would receive *"a response based on a conversation rather than just filling out a survey question"* (P10). One participant described an industry project where the questionnaire mimicked a TV personality to better engage with teenagers. Voice user interfaces could furthermore enrich the responses with affective signals derived from speech.

4.2.2 Complement With Undrawn Data. ML methods excel at quickly analyzing vast amounts of existing data. Leveraging this capability, participants see opportunities for ML to identify subtle patterns in dispersed data silos as well as to inform UX decisions with data insights.

Linking Insights Across Data Silos: Participants believed that ML can augment UX researchers to *"understand the links between data sources"* (P13) and see *"if there are any behavioral patterns, [or] pain points that we overlooked during the quantitative analysis."* (P5). Participants envisioned that with the help of ML tools they could *"map people to other data sets that we have"* (P13), such as clickstreams, social media, similar interviews, or survey responses. Doing such analyses manually is often time-consuming and slows down the line of thought, thus their potentials remain currently untapped. Participants perceived that ML methods might broaden their scope while leaving the interpretation with the human. *"Whenever you look at information just from one data set - it's like shining the flashlight only in one corner. [...] ML can help us to illuminate more."* (P7). *"It's going to be helpful to understand the bigger picture. [...] It's going to be quicker. [...] At the moment I don't see much use of AI to help us to understand the why"* (P11).

Data-Driven Personas: Furthermore, many participants saw potential for unsupervised ML in supporting user segmentation. Clustering methods may automatically identify unique user groups from data logs. *"Don't make me do all the work to figure out what kind of user segments I have"* (P1). Instead, tools might provide analytical insights on how many clusters can be observed in the data and let the UX researcher fill in the details. This could also help to *"keep the user researcher unbiased"* (P1). Unsupervised ML methods often identify patterns for which *"there is not really a human word"* (P4) and challenge researchers' potentially biased assumptions. Additionally, a data-driven persona approach could enable UX researchers to monitor user shifts more closely over the product life cycle. Currently, personas are often created once in the beginning *"and maybe you do it again in a couple of years"* (P4). UX researcher might be notified when significant changes are observable in transaction data that require adaptation of personas.

Data-Driven Design: Supporting design decisions by evaluating and recommending UI options based on historical data of user behavior or user preferences was considered another field of interest. *"We can use ML and its potential to help make good decisions in design"* (P7). P3 would like to see data-driven design tools that back the UX design process with actual numbers, e.g., *"with this particular design this is the problem [...] because 50% of the users failed at this particular step."*

4.2.3 Novel Insights into Users' Affect. In addition to improving current practices and connecting existing data, participants envisioned ML to yield novel information about users' feelings and emotions. This would enable UX researchers to better *"understand the affective component"* (P9). *"The one thing I still don't have access to is sentiment. I don't know their emotional state. Often a system can figure out the emotional state [of a user] faster [than humans]"* (P1).

Generalizing Beyond the Lab: Intrusive methods, such as electroencephalography (EEG) sensors, could be used during real-time usability tests in lab contexts to record typical flows of interaction and their corresponding emotional responses. P4 states that their

behavioral and emotional responses could be used as labels for an ML model that could then be applied to *"other users in mass"*.

Non-intrusive and Remote: Affective signals could also be derived from remote test settings to provide UX researchers with richer context when studying product use in the field. Specifically trained ML models could provide *"a better window into the emotional state that people are actually feeling"* (P9) through *"non-intrusive measures of sentiment"* (P1). For example, P1 envisioned ML to continuously classify users' facial expressions from an *"accompanying web camera feed"* and reveal that a user *"was actually looking over here and was chatting with his wife"*.

Identifying User Inconsistencies: People are sometimes observed to provide inconsistent feedback, i.e., users may *"say one thing but do another"* (P9). Affective signals may be compared to the actual behavior and thus help UX researchers to identify inconsistencies. *"These mechanisms could help uncover some of the usability flaws that are very difficult to extract with manual methods"* (P2).

Virtual User Testing: Further down the road, P9 saw potential in conducting UX studies entirely in a virtual setting with the help of virtual reality (VR). A virtual world could resemble the physical world but enable researchers to stimulate responses that are hard to simulate in the physical world. In such a virtual environment, ML methods could be used to evaluate eye gaze, motion, and neurological activity when people are experiencing those situations.

4.3 Findings: Challenges

Furthermore, we identified 2 emergent areas of challenges. First, participants foresee changing expectations towards the UX profession and a shift in future responsibilities. Second, ML was seen to make it more difficult to recruit human subjects for UXR activities.

4.3.1 Changing Expectations and Responsibilities. Participants felt that the availability of data might oblige them to report quantified insights while not feeling entirely prepared for it. Furthermore, some participants perceived that ML changes how UX researchers will be involved in projects.

Peers Demand Numbers: Driven by the promises of ML, our participants felt that leveraging large-scale data for UXR might be increasingly demanded by their peers. The potential availability of data might make expressing insights through aggregated numbers mandatory. P5 described cases where it was necessary to use quantified insights *"to convince product managers or management because without numbers it's oftentimes very hard to get somebody to understand what is happening. We already have this but need numbers."* P7 explained that *"having numbers makes it feel more scientific, even though that's not necessarily the case. [...] It's kind of a pervasive problem in [the] industry that people think only numbers are true."* A key challenge in our participants' view was *"how to balance [those] different analytical needs"* (P3). While most people in an organization require an aggregated view to understand the bigger picture, UX researchers cherish to *"look at individual flows"* (P3) to address underlying problems. *"Any good [UX] researcher or good [UX] designer would start with a user need"* (P1). Data alone leaves many interpretations. So, it is mandatory *"to enrich it with qualitative insights"* (P1). Convincing internal peers of the need for resource-intensive low-number qualitative insights might become more challenging as ML is successfully applied to other parts of an organization. *"Qualitative data is only as powerful to those who participate in it and can see the actual results. [...] Most people aren't trained to understand this thing I call qualitative validity"* (P7). To advocate the validity of qualitative insights might become more challenging for UX practitioners when not supported by numbers.

Developing Confidence and Literacy in ML: Figuring out what to do in a world of data streams was perceived as a complex challenge for UX teams. *"The problem with data is that there is so much of it. The world [...] becomes even more complex because all those data streams don't go away"* (P1). Participants perceived that UX researchers *"are not completely educated about ML [...] and do not understand that the two can work together"* (P5). Participants admitted that a cultural change is needed among UX practitioners to foster a data-driven spirit in organizations. *"A lot of user researchers are essentially traditional qualitative researchers. There is a little bit of resistance [...], but that's becoming lesser and lesser given that management wants it to be both [qualitative and quantitative]"* (P5). On the technological side, participants observed that ML remains an inaccessible design material as the usability of ML tools is often a hurdle for UX practitioners. UX teams must blindly trust the default settings of tools as they do not understand the technical foundations. *"ML is totally inaccessible to anyone who has never coded. People just trust these out of the box models and try to get it to work. It's not [going to]."* (P4). P4 would appreciate less technical terms in ML tools. Instead, the participant would like to *"call it what it is, like tell what problem it is solving"*. In-browser ML environments and interactive ML approaches have been named as examples. Often there seems to be a common belief within organizations that people could simply run ML on a problem and would obtain a meaningful solution. *"A lot of problems aren't scoped in a way that ML can help"* (P4). Interpreting the potentials of ML and having the vocabulary and confidence to argue about it with stakeholders was perceived as an obstacle for current UX practitioners.

Changing Project Involvement: Overall, participants expressed little concern that applying ML methods to UXR would reduce the demand for human researchers. *"Qualitative methods [...] result in rich data, that is only truly understandable by [...] a human being. There is so much information [...] that a machine would have a very hard time truly understanding it. It requires actual empathy and cultural appreciations"* (P7). *"I don't think any machine will ever get to the point where we trust the AI more than we trust the [UX] person"* (P1). However, there was some disagreement among participants about how the skill set of UX researchers might change in the future due to ML. On one hand, some participants believed that the role of UX will likely stay the same. *"I don't think the skill set would change. You still need to do all the things [...] to understand human behavior"* (P11). In contrast, P4 believed that ML and data-driven methods are not only changing the mindset of UX researchers but *"how people are currently doing their jobs" (P4).* Working on ML-enabled products was considered an ongoing process that will involve researchers over longer periods before becoming effective for users. This contrasts with currently established design thinking approaches, where UX researchers tend to move on to the next project after few prototype iterations (P4). Some participants believed that UXR will become even more interdisciplinary. Other disciplines, such as anthropology and sociology, might increasingly

Table 1: Participants in the expert interviews including their role, country, institution size, and work experience in the field of UX. Participants P11 and P13 asked us to omit their work-related information.

Participant	Current Role	Country of Residence	Size of Institution	Work Experience in UX
P1	UX Researcher	USA	<50	20+ years
P2	Student	India	*did not disclose*	4 years
P3	Data Analyst/Scientist	USA	1,000+	1 year
P4	UX Designer	USA	1,000+	4 years
P5	UX Designer	USA	100+	1 year
P6	UX Designer	USA	100+	1 year
P7	UX Executive	USA	<50	25+ year
P8	UX Researcher	Germany	250+	3 year
P9	Academic Researcher	USA	1,000+	13 years
P10	UX Researcher	USA	50+	1 year
P11	*did not disclose*	*did not disclose*	*did not disclose*	*did not disclose*
P12	Academic Researcher	USA	*did not disclose*	4 years
P13	*did not disclose*	*did not disclose*	*did not disclose*	*did not disclose*

contribute to the study of complex human phenomena in collaboration with current disciplines. *"It's not one single skill set anymore that you apply to understand the users"* (P1). Instead, participant P1 envisioned UX researchers to *"become a translator"* between the stakeholders involved.

4.3.2 Access to Users and Their Data. To effectively leverage ML methods, access to large amounts of user data is necessary. However, ML's reliance on data resulted in an increasing number of regulations and increased sensitivity regarding user data usage. Participants saw challenges in interpreting these regulations, balancing data economy, and finding alternative means to incentify users to participate in UXR activities.

Interpreting Privacy Regulations: Getting access to users was considered a major constraint for UX researchers as it imposes legal, confidential, and financial requirements. *"Recruiting [users] will always be the golden key"* (P1). New privacy regulations, such as the European Union *General Data Protection Regulation (GDPR)* [39], aim to improve the control for users over their data. Interpreting those regulations and finding the right balance between advocating in favor of users versus pursuing organizational interests was considered a major challenge for the time ahead. Some participants believed that UX researchers *"need to err on the side of [data] protection"* (P1) while others felt that *"any data can be used for analysis as long as PII [personal identifiable information] data is not used"* (P5).

Dealing with the Principle of Data Economy: We observed different opinions on the importance of individual data in user behavior tracking. Some participants think that the principle of data economy may limit their access to user data. Others feel that having access to aggregated data might be enough for most use cases. To understand the big picture, P8 perceived it to be more important *"to see the behavior of one average user than to watch individual cases". "I want to know when it fails. That does not need to be tied to [...] username"* (P4). Instead of tracking everyone by default, UX researchers could also turn to selectively ask individual users for feedback, e.g., via pop-up surveys on a website (P5). Excluding demographic data from individual cases may even have positive effects in terms of bias avoidance. *"I had to keep telling myself that I can't be biased over some person's background since that kind of information is not available when we generalize"* (P5).

Incentivizing Users to Contribute: As an alternative way forward, some participants felt that UX researchers and companies should rethink their relationship with user data. *"We need to give a lot more credit to the producers of the data"* (P1). They envisioned ways to encourage users to contribute their data to UXR. P1 suggested some form of *"privacy currency"* that offers benefits, such as *"reduced number of ads"* or *"5% off the purchase price"*. Companies should be more honest about their need for usage data. *"Don't automatically opt everybody in. Give them the option. Make it easy. People appreciate that more than having to dig through layers and layers of UI to uncheck a box"* (P6). None of the participants reported practical experiences in this direction. While participants seemed positive about such alternatives to compensate for potentially fewer user data, these approaches also entail challenges in promoting and implementing them internally and externally.

5 DISCUSSION

The notion of the *fourth wave of HCI* [2, 3] speculates that HCI is converging towards a trans-disciplinary paradigm as new disciplines enter the stage. Each discipline adds new dimensions, such as ethics or creativity, to the interdisciplinary discourse. Our findings from the survey and the interviews suggest that the discipline of ML entered the UXR discipline even though it may not be effectively applied to UX practices yet. Furthermore, it shifts the mindsets and work practices of practitioners towards a more quantitative interpretation of UX. Our identified opportunities overlap with findings from prior research. [45] report how UX practitioners enrich their UXR toolkit through telemetry and data stories. Further, some of our themes resemble the user-centric perspective of ML by [44]. Their perspective describes how ML can provide direct value to the user by enabling them to understand themselves (e.g., through insights into their affect) or their surrounding (e.g., through insights across data silos). This assumes that this is done dynamically by the system without a UX researcher in the loop. Our findings suggest that ML may also be used to indirectly provide value to the user by informing UXR activities. Our identified challenges indicate that

Table 2: Summary of themes observed in the expert interviews.

Higher-level Theme	*Emergent Theme*
Opportunities	
Automate the Mundane Part	Automated Transcription Engaging Surveys
Complement With Undrawn Data	Linking Insights Across Data Silos Data-Driven Personas Data-Driven Design
Novel Insights Into Users' Affect	Generalizing Beyond the Lab Non-intrusive and Remote Identifying User Inconsistencies Virtual Testing
Challenges	
Changing Expectations and Roles	Peers Demand Numbers Confidence and Literacy in ML Changing Project Involvement
Access to Users and Their Data	Interpreting Privacy Regulations Dealing with Data Economy Incentivizing Users to Contribute

UX researchers' core skills of interpersonal communication are expected to advance beyond the focus on users. Instead, they translate between multiple stakeholders as well as privacy requirements. Weaving in insights derived from data-trails and ML techniques may be required to persuade stakeholders that their conclusions are valid and will solve a relevant problem. Based on the interpretation of our findings, we see three promising directions for further HCI research that have not yet been adequately addressed.

5.1 Data-driven UX Evaluation With ML

Our findings indicate that using ML for the evaluation of a product's UX may be a promising field for future research. Most of the respondents believe that ML can provide the biggest value at the evaluation stage. Traditional UX evaluation methods are often resource-intensive and not scalable. Often standardized questionnaires such as the *user experience questionnaire (UEQ)* or the *AttrakDiff questionnaire (AD)* are used [32]. ML techniques may offer a more resource-effective alternative. Connecting questionnaire results with log and time series data about user behavior may be used as labeled data for supervised ML. Furthermore, such approaches may allow to continuously monitor changes in users' UX and inform UX researchers when it might be worth to revisit parts of the product experience. We observed that fewer UX practitioners are involved in the evaluation of a product's UX after its launch. Thus, respondents' assessment' may be positively biased because they may not have a complete picture of potential obstacles in this field. However, we found recent academic work that explores the challenges of evaluating UX using multiple data sources and proposes ML-based approaches [17, 31]. We propose to explore sensitizing concepts for ML-supported continuous UX evaluation and UX monitoring in future work.

5.2 Ensuring Effectiveness of ML-based UXR

Insights from data-driven ML techniques have the potential for effective triangulation to ultimately yield a more complete picture [32]. This matches the opportunities identified by our study participants. However, they should be carefully evaluated in practice. Critical voices have been raised about the practical applicability of automated systems in the field. Previous works compared automated ML approaches for UX research with traditional (manual) methods [14]. Results indicated that issues extracted by algorithms might differ after deployment to the field – even though they looked precise during training. UX researchers need to be able to spot questionable predictions and develop an understanding of when to rely on automated methods and when to carefully supervise them. Building ML tools for UX activities around the guidelines for *interactive ML (IML)* [7] and *explainable artificial intelligence (XAI)* [41] may be promising directions to enable UX researchers to validate and maintain the effectiveness of such tools in the field.

5.3 Calibrating Expectations Regarding ML

UX practitioners have been confronted with many novel forms of technology and interaction. Multi-device experiences, voice interfaces, and unpredictable intelligent systems pose new challenges and opportunities in terms of UX research and design [6, 43, 47]. The HCI community already raised the question of whether current methods are keeping up with the technological advancements and user expectations [16]. In line with prior work, almost all our inquired UX practitioners experimented with the new design material ML at least on a basic level – even when their work practices may primarily be qualitative. However, we observed opposing mindsets between UX practitioners with and without ML work experience. The assessments of UX+ML respondents have often been in unison with ML-only respondents. We interpret this in a way that UX practitioners with work experience in ML have a sufficient understanding to realistically assess capabilities but also limitations of ML – even though they are no technical experts. In contrast, UX-only practitioners may envision more creative use cases, e.g., regarding generative approaches, because their knowledge about difficulties in practice is limited. This might imply that UX researchers would benefit from more distinguished educational material that also addresses ML's limitations. Recent academic work lets UX practitioners refine their mental models with tools for playful exploration [21] and metaphors [5]. We suggest that such educational materials also include case studies on how to apply ML to UXR use cases.

5.4 Limitations

We acknowledge that our findings are indications that can only be generalized to a limited extent. Our participants were not selected for demographically representative proportions. The studies recruited mainly participants from the United States and Germany and were limited in time. Further, we asked our participants to reflect on the potential of ML in the future. As 13 out of 49 participants (especially in the UX-only group) had only a basic understanding of ML, some future predictions might turn out to be too optimistic. Additional experts from adjacent disciplines should be interviewed and the derived insights should be related to our analysis. Still, we are confident that our studies capture up-to-date insights about

practitioners' understanding and serve as an informative first step for future work in the emerging research field of ML for UXR. We welcome other researchers to extend or amend our insights and interpretations. Eventually, we will only be able to draw a complete picture of the applicability and acceptance of ML for UX when we conceptualize, develop, and evaluate respective tools and methods in case studies and prototypes.

6 CONCLUSION

The disciplines of ML and UX are contesting each other's borders. There is ongoing research within the HCI and UX communities on how to improve the performance of ML models through UX as well as research on how to use ML to improve a product's UX. With our work, we add the intersection of *ML for UX research* to the discussion. We found promising academic work that already experimented at the intersection of ML for UXR. Based on these, we surveyed and interviewed UX and ML practitioners. We presented practitioners' experiences and visions derived from a snapshot of 49 survey responses from UX and ML practitioners as well as 13 interviews with UX experts. Our survey indicated that the disciplines of ML and UX are expected to overlap and that UX practitioners see promising use cases of applying ML to UXR. Further, they are anticipating these developments as they are experimenting with ML even though their work routines may primarily be qualitative. We learned from the interviews that ML methods may help to automate mundane tasks, complement decisions with data-driven insights, and enrich UXR with insights from users' emotional worlds. We link our interpretations to recent academic work on ML for UXR and discuss data-driven UX evaluation based on ML as a promising use case for future research.

ACKNOWLEDGMENTS

We thank all interview partners and survey participants who took their time to contribute their experiences and opinions. The authors gratefully acknowledge the grant provided by the German Federal Ministry for Education and Research (#01IS12057).

REFERENCES

[1] Saleema Amershi, Daniel S. Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi T. Iqbal, P. Bennett, Kori Inkpen Quinn, J. Teevan, Ruth Kikin-Gil, and E. Horvitz. 2019. Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[2] Eli Blevis, Kenny Chow, Ilpo Koskinen, Sharon Poggenpohl, and Christine Tsin. 2014. Billions of Interaction Designers. *interactions* 21, 6 (Oct. 2014), 34–41. https://doi.org/10.1145/2674931

[3] Susanne Bødker. 2015. Third-wave HCI, 10 Years Later—participation and Sharing. *interactions 22*, 5 (Aug. 2015), 24–31. https://doi.org/10.1145/2804405

[4] ISO DIS. 2009. 9241-210: 2010. Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO). Switzerland* (2009).

[5] Graham Dove and Anne-Laure Fayard. 2020. Monsters, Metaphors, and Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3313831.3376275

[6] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning As a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 278–288. https://doi.org/10.1145/3025453.3025739

[7] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (June 2018), 37 pages. https://doi.org/10.1145/3185517

[8] Jackson Feijó Filho, Thiago Valle, and Wilson Prata. 2015. Automated Usability Tests for Mobile Devices Through Live Emotions Logging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Copenhagen, Denmark) *(MobileHCI '15)*. ACM, New York, NY, USA, 636–643. https://doi.org/10.1145/2786567.2792902

[9] Interaction Design Foundation. 2020. *What is UX Research?* Retrieved April 27th, 2020 from https://www.interaction-design.org/literature/topics/ux-research

[10] Juan E Gilbert, Andrea Williams, and Cheryl D Seals. 2007. Clustering for usability participant selection. *Journal of Usability Studies* 3, 1 (2007), 40–52.

[11] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux. 2016. Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. ACM, New York, NY, USA, 3558–3565. https://doi.org/10.1145/2851581.2856492

[12] Marc Hassenzahl. 2008. User Experience (UX): Towards an Experiential Perspective on Product Quality. In *Proceedings of the 20th Conference on L'Interaction Homme-Machine* (Metz, France) *(IHM '08)*. ACM, New York, NY, USA, 11–15. https://doi.org/10.1145/1512714.1512717

[13] Steffen Hedegaard and Jakob Grue Simonsen. 2013. Extracting Usability and User Experience Information from Online User Reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 2089–2098. https://doi.org/10.1145/2470654.2481286

[14] Steffen Hedegaard and Jakob Grue Simonsen. 2014. Mining Until It Hurts: Automatic Extraction of Usability Issues from Online Reviews Compared to Traditional Usability Evaluation. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (Helsinki, Finland) *(NordiCHI '14)*. ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/2639189.2639211

[15] Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. 2016. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute* 4 (2016).

[16] Karen Holtzblatt, Ilpo Koskinen, Janaki Kumar, David Rondeau, and John Zimmerman. 2014. Design Methods for the Future That is Now: Have Disruptive Technologies Disrupted Our Design Methodologies?. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. ACM, New York, NY, USA, 1063–1068. https://doi.org/10.1145/2559206.2579401

[17] Md Sazzad Hossain, Amin Ahsan Ali, and M. Ashraful Amin. 2019. Eye-Gaze to Screen Location Mapping for UI Evaluation of Webpages. In *Proceedings of the 2019 3rd International Conference on Graphics and Signal Processing* (Hong Kong, Hong Kong) *(ICGSP '19)*. Association for Computing Machinery, New York, NY, USA, 100–104. https://doi.org/10.1145/3338472.3338483

[18] Christian Jetter and Jens Gerken. 2007. A simplified model of user experience for practical application. In *2nd COST294-MAUSE*. 106–111.

[19] Soon-Gyo Jung, Jisun An, Haewoon Kwak, Moeed Ahmad, Lene Nielsen, and Bernard J. Jansen. 2017. Persona Generation from Aggregated Social Media Data. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. ACM, New York, NY, USA, 1748–1755. https://doi.org/10.1145/3027063.3053120

[20] Ashley Karr. 2015. UX Research vs. UX Design. *Interactions* 22, 6 (Oct. 2015), 7–7. https://doi.org/10.1145/2834964

[21] Claire Kayacik, Sherol Chen, Signe Noerly, Jess Holbrook, Adam Roberts, and Douglas Eck. 2019. Identifying the Intersections: User Experience + Research Scientist Collaboration in a Generative Machine Learning Interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, Article CS09, 8 pages. https://doi.org/10.1145/3290607.3299059

[22] Eugene Kharitonov, Alexey Drutsa, and Pavel Serdyukov. 2017. Learning Sensitive Combinations of A/B Test Metrics. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, United Kingdom) *(WSDM '17)*. ACM, New York, NY, USA, 651–659. https://doi.org/10.1145/3018661.3018708

[23] Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology* (second ed.). Sage.

[24] Mike Kuniavsky, Elizabeth Churchill, and Molly Wright Steenson. 2017. *The 2017 AAAI Spring Symposium Series Technical Reports: Designing the User Experience of Machine Learning Systems.* Technical Report. Technical Report SS-17-04. Palo Alto, California.

[25] Effie Lai-Chong Law, Virpi Roto, Marc Hassenzahl, Arnold P.O.S. Vermeeren, and Joke Kort. 2009. Understanding, Scoping and Defining User Experience: A Survey Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 719–728. https://doi.org/10.1145/1518701.1518813

[26] Effie Lai-Chong Law, Paul van Schaik, and Virpi Roto. 2014. Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies* 72, 6 (2014), 526–541.

[27] Alexandros Liapis, Christos Katsanos, Dimitris Sotiropoulos, Michalis Xenos, and Nikos Karousos. 2015. Subjective Assessment of Stress in HCI: A Study of the Valence-Arousal Scale Using Skin Conductance. In *Proceedings of the 11th*

Biannual Conference on Italian SIGCHI Chapter (Rome, Italy) *(CHItaly 2015)*. ACM, New York, NY, USA, 174–177. https://doi.org/10.1145/2808435.2808450

[28] Martin Lindvall, Jesper Molin, and Jonas Löwgren. 2018. From Machine Learning to Machine Teaching: The Importance of UX. *Interactions* 25, 6 (Oct. 2018), 52–57. https://doi.org/10.1145/3282860

[29] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 348, 12 pages. https://doi.org/10.1145/3173574.3173922

[30] Marília Soares Mendes and Elizabeth Sucupira Furtado. 2017. UUX-Posts: A Tool for Extracting and Classifying Postings Related to the Use of a System. In *Proceedings of the 8th Latin American Conference on Human-Computer Interaction* (Antigua Guatemala, Guatemala) *(CLIHC '17)*. ACM, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/3151470.3151471

[31] Walter T. Nakamura, Elaine H. T. de Oliveira, and Tayana Conte. 2019. Negative Emotions, Positive Experience: What Are We Doing Wrong When Evaluating the UX?. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, Article LBW0281, 6 pages. https://doi.org/10.1145/3290607.3313000

[32] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A Bermuda Triangle?: A Review of Method Application and Triangulation in User Experience Evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 461, 16 pages. https://doi.org/10.1145/3173574.3174035

[33] B Joseph Pine, Joseph Pine, and James H Gilmore. 1999. *The experience economy: work is theatre & every business a stage*. Harvard Business Press.

[34] Christian Robert. 2014. Machine learning, a probabilistic perspective.

[35] V Roto, E Law, APOS Vermeeren, and J Hoonhout. 2011. U white paper-Bringing clarity to the concept of user experience. In *Outcome of Dagstuhl Seminar on Demarcating User Experience, Germany. http://allaboutux. org/uxwhitepaper*.

[36] Samaneh Soleimani and Effie Lai-Chong Law. 2017. What Can Self-Reports and Acoustic Data Analyses on Emotions Tell Us?. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. ACM, New York, NY, USA, 489–501. https://doi.org/10.1145/3064663.3064770

[37] Jacopo Staiano, María Menéndez, Alberto Battocchi, Antonella De Angeli, and Nicu Sebe. 2012. UX_Mate: From Facial Expressions to UX Evaluation. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. ACM, New York, NY, USA, 741–750. https://doi.org/10.1145/2317956.2318068

[38] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research*. Sage publications.

[39] The European Parliament and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* (2016).

[40] Alexandre N. Tuch, Rune Trusell, and Kasper Hornbæk. 2013. Analyzing Users' Narratives to Understand Experience with Interactive Products. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 2079–2088. https://doi.org/10.1145/2470654.2481285

[41] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.

[42] Terry Winograd. 2006. Shifting Viewpoints: Artificial Intelligence and Human–computer Interaction. *Artif. Intell.* 170, 18 (Dec. 2006), 1256–1258. https://doi.org/10.1016/j.artint.2006.10.011

[43] Qian Yang. 2017. The Role of Design in Creating Machine-Learning-Enhanced User Experience. In *2017 AAAI Spring Symposium Series*.

[44] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 130, 11 pages. https://doi.org/10.1145/3173574.3173704

[45] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. ACM, New York, NY, USA, 585–596. https://doi.org/10.1145/3196709.3196730

[46] Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020).

[47] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2016. Planning Adaptive Mobile Experiences When Wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. ACM, New York, NY, USA, 565–576. https://doi.org/10.1145/2901790.2901858

[48] Xiang Zhang, Hans-Frederick Brown, and Anil Shankar. 2016. Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5350–5359. https://doi.org/10.1145/2858036.2858523

Mind the (Persuasion) Gap: Contrasting Predictions of Intelligent DSS with User Beliefs to Improve Interpretability

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Florian Fincke
LMU Munich
Munich, Germany
florian.fincke@campus.lmu.de

Andreas Butz
LMU Munich
Munich, Germany
butz@ifi.lmu.de

ABSTRACT

Decision support systems (DSS) help users to make more informed and more effective decisions. In recent years, many intelligent DSS (IDSS) in business contexts involve machine learning (ML) methods, which make them inherently hard to explain and comprehend logically. Incomprehensible predictions, however, might violate the users' expectations. While explanations can help with this, prior research also shows that providing explanations in all situations may negatively impact trust and adherence, especially for users experienced in the decision task at hand. We used a human-centered design approach with domain experts to design a DSS for funds management in the construction industry and identified a strong need for control, personal involvement, and adequate data. To create an adequate level of trust and reliance, we contrasted the system's predictions with the values derived from an analytic hierarchical process (AHP), which makes the relative importance of our users' decision-making criteria explicit. We developed a prototype and evaluated its acceptance with 7 construction industry experts. By identifying situations in which the ML prediction and the domain expert potentially disagree, the DSS can identify a persuasion gap and use explanations more selectively. Our evaluation showed promising results and we plan to generalize our approach to a wider range of explainable artificial intelligence (XAI) problems, e.g., to provide explanations with arguments tailored to the user.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

decision support systems; decision-making; interpretability; explainable artificial intelligence; analytical hierarchical process

ACM Reference Format:
Michael Chromik, Florian Fincke, and Andreas Butz. 2020. Mind the (Persuasion) Gap: Contrasting Predictions of Intelligent DSS with User Beliefs to Improve Interpretability. In *ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '20 Companion), June 23–26, 2020, Sophia Antipolis, France.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3393672.3398491

EICS '20 Companion, June 23–26, 2020, Sophia Antipolis, France

ACM ISBN 978-1-4503-7984-7/20/06…$15.00
https://doi.org/10.1145/3393672.3398491

1 INTRODUCTION

The rapidly growing volume of data in many parts of the enterprise makes it necessary to structure and manage it in information systems. Those systems which help in decision-making are referred to as *decision support systems (DSS)* [30]. With the recent improvements in *machine learning (ML)* methods, DSS are becoming more and more intelligent. So-called *intelligent decision support system (IDSS)* augment the collected data with predictions that guide and (semi-) automate parts of the decision-making process [30]. However, these intelligent DSS also introduced new challenges because their rationale is often not interpretable and hence perceived as non-deterministic by their users.

The effectiveness of an intelligent DSS depends not only on the accuracy of its underlying ML model or algorithm. Instead, it is only effective if it serves the information needs of decision-makers and is also accepted and trusted by them. Jarvis describes DSS as the general idea of "*combining the computer's computational power with the decision maker's intuition and judgment in an interactive manner, [so that] better decisions will result than by either the computer or human taken separately*" [11]. To achieve such a symbiosis, we need to design user interfaces (UI) that communicate the rationale behind algorithmic predictions in human-understandable terms. The UI should help to calibrate the user's understanding of the system's capabilities and limitations to prevent both over-reliance (when users blindly trust system recommendations) and under-reliance (when users simply ignore system recommendations) [5].

We conducted a design study in the construction industry and asked decision-makers about their requirements regarding interpretability of a novel intelligent DSS module on addenda approval. We use the term *interpretability* to refer to measures provided by a DSS with the aim of enabling users to understand its inner workings. Interpretability is a broad concept that may imply other distinct ideas such as transparency, trust, and fairness [17]. It is often used to indirectly evaluate whether important requirements, such as reliability, trust, or control are met in a particular context [8]. Biran and Cotton consider intelligent systems interpretable "*if their operations can be understood by a human*" [3]. We followed a human-centered design process to understand how project managers and executives make decisions regarding validation and approval of budget addenda. Budget management in the construction industry is an interesting context to study for two reasons: First, the construction industry itself is one of the least digitized industries but digitization efforts (e.g., building information modeling (BIM)) are

gaining adoption despite decision makers skepticism [1, 4]. Secondly, the addenda approval process is a complex decision situation that requires decision makers to retrieve and interpret data from distributed sources and also consider their dependencies. To date this is a highly manual and subjective process

This paper investigates interpretability needs of human decision-makers in the field regarding an intelligent DSS. In particular, we propose an approach to align the level of trust and reliance by contrasting ML predictions with user beliefs. User beliefs can be extracted through multi-attribute decision making approaches such as the *analytic hierarchical process (AHP)*. Making the user beliefs explicit allows the system to better identify *persuasion gaps* [6], i.e., situations in which the system and user base their decision on different criteria. We think that this approach might be a valuable starting point for providing selective and personalized explanations to the field of explainable artificial intelligence (XAI). With this work, we put our suggested approach and formative evaluation up for discussion with our fellow researchers.

2 RELATED WORK

2.1 Intelligent Decision-Support Systems

Decision-making refers to the cognitive process of selecting a logical choice from many available alternatives. Decision-making problems are often structured into three phases: problem identification, model development and use, and action plan development [21]. In our work, we focus on the second phase that deals with eliciting user preferences and comparing alternatives in a consistent way. If a decision is rational it is typically based on facts instead of arbitrary choices. *Multi-attribute decision making (MADM)* describes approaches that leverage (potentially conflicting) attributes to select, compare, and rank a limited number of discrete alternatives [31]. The rationality of decision-making, however, is bounded as individuals are often not able to make optimal decisions in an economically rational way due to cognitive limitations and resource constraints [28]. Simon suggests that instead of maximizing (search for the best possible option), decision makers in the field are rather satisficing (stick to an option that is considered good enough) [28].

In many business-related contexts, *decision support systems (DSS)* organize relevant facts to assist users in decision-making and improve effectiveness of the decision outcome [30]. DSS can range from simple spreadsheets to complex data warehouse systems [30]. They are typically distinguished by their underlying technology, theory foundations, target users, and decision tasks [2]. So called, *intelligent decision support systems (IDSS)* use artificial intelligence methods to support the decision-making and exhibit some notion of "intelligent behavior" [30]. Such intelligent behavior may either by applied to the system's underlying data base (e.g., identifying relevant attributes), knowledge base (e.g., suggesting decision alternatives), or model base (e.g., choosing applicable formal decision-making methods) [22].

In our work, we focus on IDSS that recommend decision alternatives to the user (*model development and use* at the *knowledge base*). The first generation of IDSS (also called *knowledge-driven DSS*) leveraged domain knowledge encoded in rule-based reasoning modules [30]. Examples include MYCIN [27] for bacterial infection diagnosis or DENDRAL [16] for chemical analysis. In contrast, modern IDSS leverage machine learning (ML) methods that implicitly infer rules from observations and thus learn from experience. This implicit inference of rules may result in the *black-box problem* for decision-makers. A black box refers to a situation in which it is possible to observe the input and outputs of a model, but the internals remain disclosed or uninterpretable to humans. In ML, the black box behavior may arise either from complex algorithms (as with deep neural networks) or from proprietary models that may otherwise be interpretable (such as with the COMPAS recidivism model) [24]. As decision-makers were always considered an integral part of the DSS [22], special attention must be paid to the design of the user interaction. With ML-enabled DSS this interaction must include explanation facilities that result in usable interpretability for decision makers.

2.2 Interpretability and Task Expertise

Prior research shows that a lack of interpretability can lead to users that mistrust, misuse, or reject a system [15, 19]. Often these result from a perceived mismatch between users' expectations and the actual behavior of a system [9]. Chander et al. describe two reasons for the mismatch to occur in a business-related decision-making context [6]: (i) the system's underlying data lacks decision criteria relevant for this situation (*awareness gap*). For instance, the user might have relevant contextual information from a phone call with a client that the system has no access to; (ii) the system's prediction is not in line with the user's beliefs and the system fails to persuade the user to adjust their beliefs (*persuasion gap*). In such a situation, the user and the system have access to the same information but weight decision criteria differently. The gaps are even more pronounced in a business-related context, where domain experts often can draw upon rich prior knowledge and beliefs about the decision situation when assessing the system (*extrinsic setting*) [20]. Explanations about the factors that contributed to the system's prediction, e.g., in natural language or in the form of visualizations, are considered one way of addressing those gaps. However, in prior research, rational explanations were shown to be only effective for participants that are not familiar with a given task [26]. The effects of explanation drop as users' confidence with the task increases over time. As user get confident with the task and the system's prediction, they become less situation aware. Most explanation approaches assume that explanations are displayed with every system prediction.

3 USE CASE AND METHODOLOGY

In our work, we outline and probe an approach that provides system explanations only when a mismatch with the user's beliefs occurs (persuasion gap). Such an approach may increase the situation awareness of decision-makers. We focused on the use case of addenda approval and risk assessment in the construction industry. We cooperated with *Alasco*[1] and their clients. Alasco provides a web-based cost controlling system for the construction industry that connects stakeholders and digitizes processes around budgeting, reporting, addenda management, and payment. We followed a human-centered design process that consisted of three

[1] http://www.alasco.de

phases: (i) we interviewed executives about their current addenda-related decision-making and derived intelligibility needs for an (semi-)automated addenda approval process; (ii) we designed and developed an interactive prototype that reflects those intelligibility needs; (iii) we evaluated our prototype in a formative user study to understand the acceptance of the prototype workflow.

Use Case: Addenda Approval. Our use case targets project management (PM) executives in the construction industry. The PM is responsible for the fulfillment of the construction project and acts as a coordinator between contractors on behalf of the building owner [13]. During the initial budget planning, the overall budget is split into a hierarchy of cost groups (e.g., property or financing). Each cost group consists of one to many contract units. Each contract unit represents the budget planned for commissioning a contractor for a task. As a construction project advances, contract units might require budget addenda due to unforeseen incidents or flaws in the initial budget planning. After ensuring the plausibility of the addendum, PM executives need to redistribute budgets from other contract units to accommodate the addendum. While doing this, decision-makers need to take the addendum risk and cost forecast of the other contract units into account.

Phase I: Need-Finding. The goal of the first phase, was to identify decision criteria and interpretability needs for an intelligent DSS for the addenda approval process. To understand domain experts' current decision-making processes around addenda approval, we conducted semi-structured interviews with 3 project managers and 2 project controllers who are proficient users of the Alasco software. Their average industry experience were 2.8 years (min=1, max=6). The interviews were held in the regular work environments and took between 30 and 45 minutes. The interviews were recorded and transcribed. To enrich our qualitative insights, participants were surveyed after each interview with the *decision-making questionnaire (DMQ)*. The DMQ is a validated psychological questionnaire that aims to examine factors important to a decision-maker in the moment of decision-making in a specific context [7]. It consists of 14 questions which correspond to 3 factors (and 10 subfactors) that characterize a decision-making situation: (i) the nature of the decision or task, (ii) the cognitive and affective abilities of the decision maker, and (iii) the environment of the decision.

Phase II: Prototyping. We integrated our prototype as a separate module on top of the Alasco software. We reused the general structure and user interface of the software as participants were already familiar with it. The prototyping process was informed by the results of the need analysis as well as prior work on DSS and interpretability. Financial data has strict privacy regulations. Also, the production data of the participant's organizations varied greatly and was often incomplete. Thus, we centered our prototype around an addenda approval scenario based on a synthetically created data set so that all participants could be evaluated on the same scenario. The scenario consisted of an onboarding phase and an addenda approval phase. We developed a functional front-end prototype while the back-end was mostly static around the evaluation scenario.

Phase III: Formative Evaluation. After the design phase, we conducted a formative user study to evaluate the prototype's acceptance regarding participant's sense of control and sense of information. As the use case and required domain expertise limited the number of potential study participants, we adopted a qualitative evaluation approach. We recruited 5 project managers and 2 project controllers with an average industry experience of 3.7 years (min=0.5, max=15). The user study included 3 participants from phase I as well as 4 new participants. This reduced the risk of receiving biased feedback from participants who had already thought about (semi-)automating addenda approvals. The participants were presented with the scenario and asked to complete an addenda approval task including the onboarding task. While doing so, participants were encouraged to think aloud. Completing the task took approximately 10 minutes. After the tasks, participants were interviewed using open-ended questions about their experience. The user study was audio-recorded, transcribed. The results were qualitatively analyzed according to Kuckartz [14] by two coders (with a Kappa coefficient of 0.86). We used the driving factors resulting from the DMQ as categories and, following Kuckartz, their gradual levels as subcategories. Table 1 presents our final coding system after multiple iterations.

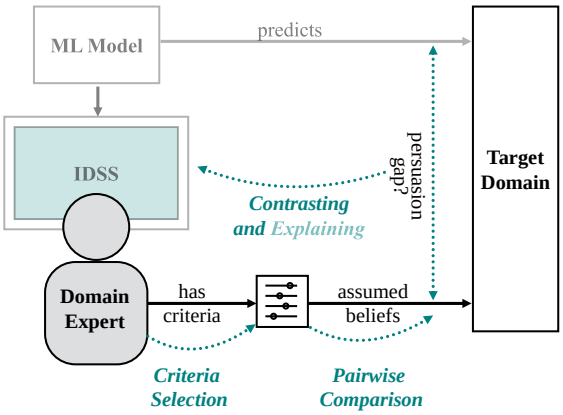


Figure 1: Explanations are triggered if there is a mismatch between the user's assumed beliefs (elicited through AHP) and the system's predictions. Blue parts relate to screens of our prototype. Muted parts relate to our proposed future work.

4 RESULTS

4.1 Interpretability Needs

The process for addendum validation was uniform for all participants. However, all participants agreed that there is no documented or formal way of deciding how to redistribute budgets. Instead, they base decisions on their personal experience and data derived from reporting features of the Alasco software. However, this approach has limitations. P2 asked for more structured workflow for addendum approval so that every stakeholder accomplishes the task in a predefined order to improve reporting. P1 would like have feedback on how well the initial budget distribution worked in

comparison to the final stage of a project. P1 and P5 asked for decision support that guides the user and recommends possible sources (e.g., based on the forecasted costs). P3 and P4 even suggested to (semi-)automate the allocation. The analysis of the DMQ indicated that control and personal involvement are important requirements for the participants. The most important subfactors were the need for *information and goals* (5 participants), *self-regulation* (4 participants), and *time/money pressure* (4 participants). It is important for the participants to have adequate and transparent data available that help them to plan, monitor, and evaluate results [7]. We leveraged these insights as guidelines for our prototype.

4.2 Prototype

We developed an IDSS interface with which participants could interact. The prototype consisted of two user flows. The first flow elicits the user's beliefs during the user onboarding through a widely accepted MADM approach. The second flow guides the user through the approval process once an addendum is requested and suggests options for budget transfer.

4.2.1 Belief Elicitation Flow. MADM approaches were used to make subjective user preferences explicit and, thus, make decision-making more transparent [21]. We leverage such an approach to elicit user beliefs about our target domain. A widely accepted and accessible MADM approach is the *analytic hierarchy process (AHP)* [10, 25]. AHP builds on a hierarchical representation of the decision problem. It leverages a user's judgments of the relative attribute importance to choose an alternative. The judgments and beliefs are elicited through pairwise comparisons of attributes. The decision criteria may be split into multiple hierarchy levels depending on the complexity. However, we limited our prototype to five decision criteria that are on the same level. We applied the wizard pattern to guide the decision maker through the steps of the AHP setup as part of a mandatory module onboarding [29]. First, users were introduced to the purpose of the flow and each step. Second, user had to select at least three criteria that they believe are important when withdrawing budget from a contract unit. Afterwards, they had to express the relative importance of each criteria through pairwise comparisons. We used the original AHP space consisting of a bidirectional Likert scale ranging from 9 (absolutely more important) to 1 (equally important) to 9. In a last step, we checked the judgments for inconsistencies and asked users to revise them if necessary. After the onboarding, users can revise their preferences anytime.

4.2.2 Intelligent Addenda Approval Flow. We enriched the manual approval flow with an intelligent overview that suggests contract units to withdraw budget from. First, the user is notified via email if a new addendum is to be reviewed. After confirming that the addendum is valid, the user sees an overview of possible contract units that may be used to accommodate the addendum. Each contract unit alternative is enriched with two types of information: (i) a score that reflects the user's beliefs. The score is calculated by AHP based on the user's relative importance of attributes as elicited during the onboarding; (ii) an intelligent suggestion that was said to take historical data into account. The suggestion may be derived through a machine learning model. Contrasting both information

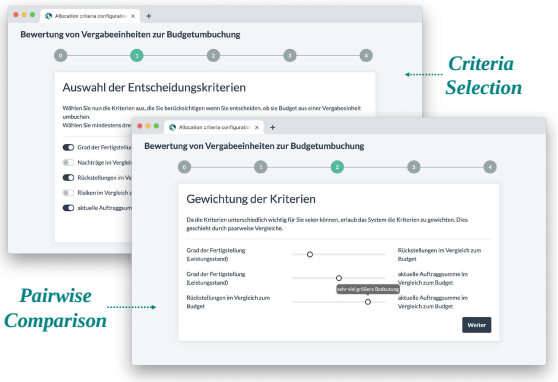


Figure 2: User's beliefs about the decision situation are elicited through AHP. In the first step, the user indicates which decision criteria are important for her. Afterwards, she compares those criteria pairwise express the relative importance.

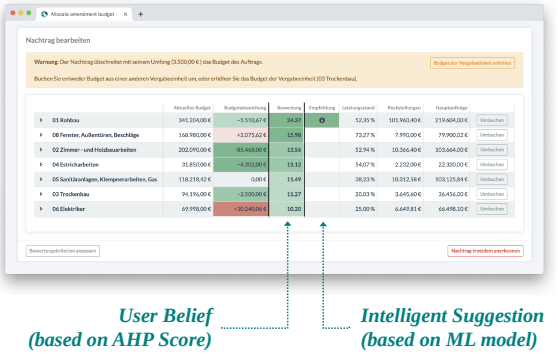


Figure 3: After an addendum is validated by the user, the IDSS gives an overview of contract units to transfer budget from. The alternatives are scored based on elicited user beliefs and contrasted with the system recommendation.

enables the user to grasp when their beliefs diverge from the system suggestion. Furthermore, it enables the system to identify and address a persuasion gap. Each column and the prediction have a tooltip that explains where the information is coming from. In our formative evaluation, the system suggestion and explanation were non-functional but based on static information. As participants were not provided with information about the underlying system logic, it resembles from a user perspective an IDSS.

4.3 Formative Evaluation

All participants were able to complete the given approval task. The results of the qualitative analysis show that all participants made

Table 1: (*Left*) Categories and subcategories derived from the results of the DMQ. (*Right*) Number of participants' statements during formative evaluation coded according to those (sub)categories.

Categories	*# of Statements*
Sense of Control	**36**
Full sense of control *(no doubts)*	12
Reinforced control *(feeling of guidance)*	8
Foreseeable behavior of the system *(no surprises)*	7
Expressed doubt/questioned system	8
Unclear statements regarding sense of control	1
Sense of Information	**23**
Improved experience due to information displayed	8
Satisfied with the amount of information	7
Desired additional information	7
Unclear statements regarding sense of information	1
Usability	**31**
Perceived increase in efficiency	7
User was hesitating/unclear	18
Expressed high mental effort	6

positive statements regarding their *sense of control* (relates to DMQ's self-regulation subfactor). 5 participants stated that their *sense of information* (relates to DMQ's information and goals subfactor) improved due to the information provided. However, 4 participants questioned the system at some point. 3 participants wished for additional information (e.g., emails or contract correspondences) or more detailed explanations (e.g., how their input affects the outcome). 4 participants perceived high mental efforts when choosing and comparing their relevant decision criteria during the onboarding. We attributed this to the fact, that they rarely had to articulate how they make addendum-related decisions before this study. However, these efforts paid off later on. 4 participants perceived increased efficiency during the addenda approval flow as they did not need to assess each alternative individually but instead could rely on the score and suggestion. Overall, we found that our prototype left the participants with an increased sense of control and information. However, the usability of the belief elicitation flow should be revised to reduce users' mental efforts. Table 1 presents a categorized summary of participants' statements.

5 LIMITATION AND FUTURE WORK

While our formative evaluation shows promising results, we acknowledge multiple limitations. Our work focuses on the limited use case of addenda approval in the construction industry. Our user studies were conducted under supervision in a controlled environment. Thus, actual user behavior and usage may be different in the field. Furthermore, our evaluation focused on the general acceptance of the approach by domain experts with a non-functional prototype. In future, we plan to conduct an experimental study that focuses on whether such an approach improves a user's understanding of an intelligent system. For this, we plan to transfer the approach to a human-grounded [8] evaluation scenario with lay users.

We believe that eliciting user beliefs and comparing them with intelligent predictions offers a promising basis for personalized explanations in XAI systems. ML algorithms take features and calculate their respective weights while optimizing a utility function. Post-hoc feature attribution methods, such as LIME [23] or SHAP [18], elicit the relative importance of a black box model's decision criteria. Similarly, decision-makers try to, explicitly or implicitly, optimize a utility function that is used to quantify their preferences regarding decision alternatives [12]. The difference is that decision-makers often do not know their utility function in advance and sometimes construct it ad-hoc during the decision-making situation. MADM methods, such as AHP, can make the user's beliefs explicit and accessible to explanation generating XAI systems. As part of our future work, we want to examine ways to relate the weights of post-hoc feature attribution methods to AHP's relative attribute importance. By this, XAI systems could adapt their explanation vocabulary (e.g., add or remove features to an explanation) or argumentation (e.g., argue with the user's expected outcome as the foil) based on the user's beliefs.

6 ACKNOWLEDGMENTS

We thank the participants who took their time to contribute their experiences and opinions from the field.

REFERENCES

[1] Rajat Agarwal, Shankar Chandrasekaran, and Mukund Sridhar. 2016. Imagining construction's digital future. *McKinsey & Company* (2016). https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/imagining-constructions-digital-future

[2] David Arnott and Graham Pervan. 2012. Design Science in Decision Support Systems Research: An Assessment using the Hevner, March, Park, and Ram Guidelines. *J. AIS* 13, 11 (2012), 1. http://aisel.aisnet.org/jais/vol13/iss11/1

[3] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[4] JL Blanco, S Fuchs, M Parsons, and MJ Ribeirinho. 2018. Artificial intelligence: Construction technology's next frontier| McKinsey. https://www.mckinsey.com/industries/capital-projects-and-infrastructure/our-insights/artificial-intelligence-construction-technologys-next-frontier

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

[6] Ajay Chander, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. 2018. Working with Beliefs: AI Transparency in the Enterprise. In *Joint Proceedings of the ACM IUI 2018 Workshops co-located with the 23rd ACM Conference on Intelligent User Interfaces (ACM IUI 2018), Tokyo, Japan, March 11, 2018.* http://ceur-ws.org/Vol-2068/exss14.pdf

[7] María Luisa Sanz de Acedo Lizarraga, María Teresa Sanz de Acedo Baquedano, María Soria Oliver, and Antonio Closas. 2009. Development and validation of a decision-making questionnaire. *British Journal of Guidance & Counselling* 37, 3 (2009), 357–373. https://doi.org/10.1080/03069880902956959

[8] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretability. *CoRR* abs/1702.08608 (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[9] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58 (2003), 697–718.

[10] Saul I. Gass. 2005. Model World: The Great Debate-MAUT Versus AHP. *Interfaces* 35, 4 (July 2005), 308–312. https://doi.org/10.1287/inte.1050.0152

[11] John J Jarvis. 1976. *Decision Support Systems: Theory.* Technical Report. Battelle Columbus Labs OH.

[12] Ralph Keeney, Howard Raiffa, and David Rajala. 1979. Decisions with Multiple Objectives: Preferences and Value Trade-Offs. *Systems, Man and Cybernetics, IEEE Transactions on* 9 (08 1979), 403 – 403. https://doi.org/10.1109/TSMC.1979.4310245

[13] Sascha Kilb and Markus Weigold. 2017. *Projektmanagement.* Springer Fachmedien Wiesbaden, Wiesbaden, 479–503. https://doi.org/10.1007/978-3-658-05368-0_20

[14] Udo Kuckartz. 2019. *Qualitative Text Analysis: A Systematic Approach.* Springer International Publishing, Cham, 181–197. https://doi.org/10.1007/978-3-030-

15636-7_8
[15] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. Association for Computing Machinery, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023
[16] Robert K Lindsay, Bruce G Buchanan, Edward A Feigenbaum, and Joshua Lederberg. 1993. DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial intelligence* 61, 2 (1993), 209–261.
[17] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages. https://doi.org/10.1145/3236386.3241340
[18] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions
[19] Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.
[20] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1802.00682 (2018). arXiv:1802.00682 http://arxiv.org/abs/1802.00682
[21] M. Pavan and R. Todeschini. 2009. 1.19 - Multicriteria Decision-Making Methods. In *Comprehensive Chemometrics*, Steven D. Brown, Romá Tauler, and Beata Walczak (Eds.). Elsevier, Oxford, 591 – 629. https://doi.org/10.1016/B978-044452701-1.00038-7
[22] Gloria Phillips-Wren, Manuel Mora, Guisseppi A. Forgionne, Leonardo Garrido, and Jatinder N. D. Gupta. 2006. *A Multicriteria Model for the Evaluation of Intelligent Decision-making Support Systems (i-DMSS)*. Springer London, London, 3–24. https://doi.org/10.1007/1-84628-231-4_1
[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778
[24] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[25] Thomas L. Saaty. 2001. *Decision making for leaders : the analytic hierarchy process for decisions in a complex world* (new 3rd ed ed.). Pittsburgh, Pa., RWS Publications.
[26] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308
[27] Edward Shortliffe. 1976. *Computer-based medical consultations: MYCIN*. Vol. 2. Elsevier.
[28] H.A. Simon. 1957. *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. Wiley. https://books.google.de/books?id=lEZdwwEACAAJ
[29] Jenifer Tidwell. 2011. Designing Interfaces (Vol. 2).
[30] Efraim Turban, Ting-Peng Liang, and Jay E Aronson. 2005. *Decision Support Systems and Intelligent Systems:(International Edition)*. Pearson Prentice Hall.
[31] Stelios H. Zanakis, Anthony Solomon, Nicole Wishart, and Sandipa Dublish. 1998. Multi-attribute decision making: A simulation comparison of select methods. *European Journal of Operational Research* 107, 3 (1998), 507 – 529. https://doi.org/10.1016/S0377-2217(97)00147-1

Michael Chromik (2020): reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP. In: Proceedings of the 18th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2020_p06

reSHAPe: A Framework for Interactive Explanations in XAI Based on SHAP

Michael Chromik
LMU Munich
michael.chromik@ifi.lmu.de

Abstract. The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation-generating methods. In this paper, we describe the need for interactive explanation facilities for end-users in XAI. We believe that interactive explanation facilities that provide multiple layers of customizable explanations offer promising directions for empowering humans to practically understand model behavior and limitations. We outline a web-based UI framework for developing interactive explanations based on SHAP.

Introduction

We have witnessed the widespread adoption of intelligent systems into many contexts of our lives. The perception of intelligence often results from their black-box behavior, which may manifest itself in two ways: either from complex machine learning (ML) architectures, as with deep neural networks, or from proprietary models that may intrinsically be white-boxes, but are out of the user's control (Rudin, 2019). As such black-box systems are introduced into more sensitive

contexts, there is a growing call by society that they need to be capable of explaining their behavior in human-understandable terms.

Much research is conducted in the growing fields of *interpretable machine learning (IML)* and *explainable artificial intelligence (XAI)* to foster human understanding. IML often refers to research on models and algorithms that are considered as inherently interpretable while XAI typically refers to the generation of (*post-hoc*) explanations for black-box models to make those systems comprehensible (Rudin, 2019; Biran and Cotton, 2017). Current XAI research mostly focuses on the cognitive process of explanation, i.e., identifying likely root causes of a particular event (Miller, 2018). As a result of this cognitive process, some notions of explanation, such as texts, annotations, or super-pixels, are generated that approximate the model's underlying prediction process.

We believe that an important aspect required to address the call for *"usable, practical and effective transparency that works for and benefits people"* (Abdul et al., 2018) is currently not sufficiently studied: providing users of XAI methods and systems with means of interaction that go beyond a single explanation.

Explanation as an Interactive Dialogue

XAI research often implicitly assumes that there is a single message to be conveyed through an explanation (Abdul et al., 2018). However, in decision-making situations that demand explainability, it is unlikely that a single explanation can address all concerns and questions of a user. This resonates with the social science perspective that considers explanation to be a social process between the *explainer* (sender of an explanation) and the *explainee* (receiver of an explanation) forming a multi-step dialogue between both parties (Miller, 2018). Especially, in situations where people may be held accountable for a particular decision, a user may have multiple follow-up questions before feeling comfortable to trust a system prediction. To model the notion of social explanation between an explanation-generating XAI system and a human decision-maker, we need means of interactivity. Related machine learning approaches, such as explanatory debugging (Kulesza et al., 2015) or interactive machine learning (Dudley and Kristensson, 2018), leverage explanations, interactivity, and human inputs to correct bugs or to improve model performance, respectively.

In our opinion, the social perspective of explanation is currently not sufficiently reflected in current XAI research that addresses decision-making situations. Weld et al. propose seven different follow-up and drill-down operations (Weld and Bansal, 2019). Olah et al. (2018) explore the design space of interpretability interfaces for neural networks and describe possible interaction operations. Recent tools, such as *Google's What-If*, focus primarily on developers and enable them to interactively inspect a ML model with minimal coding. However, they do not provide interactive explanations to end-users of XAI systems.

reSHAPe: Interactive SHAP Explanations

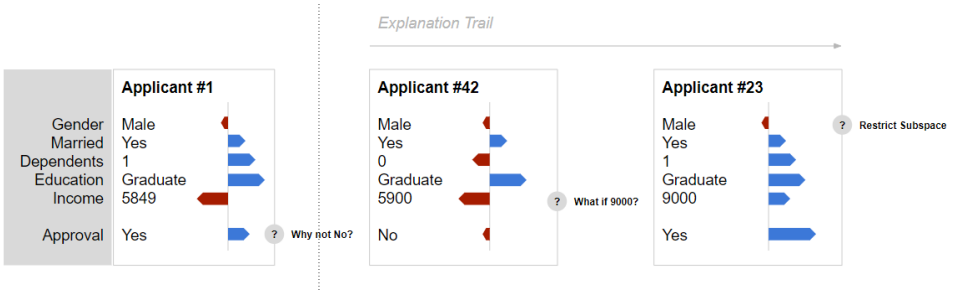


Figure 1. Schematic UI prototype of interactive explanation trail in reSHAPe: The outcome of each observation is explained through SHAP's feature attribution method (red=negative influence on outcome, blue=positive influence). Starting from an initial observation of interest, the user can select one follow-up question from a set of interaction options to validate their hypotheses. Each query returns an illustrative observation and adds it to the explanation trail.

We propose a web-based UI framework that enables developers to provide interactive explanations for end-users. We leverage existing model-agnostic post-hoc explanation-generating methods and integrate them into an interaction concept for navigating between the methods from a human-centered perspective. We build upon the methods provided by the SHAP framework (Lundberg and Lee, 2017). *SHAP (SHapley Additive exPlanations)* is a promising starting point as it unifies existing feature attribution methods (such as *LIME* and *DeepLIFT*) and connects them to additive Shapley values. Furthermore, it allows the generation of *local* and *global* explanations that are consistent with each other as they both use Shapley values as atomic units. This makes them suitable for guiding users through multi-stepped explanations following one line of thought.

However, prior research indicates that even experienced ML engineers have difficulties to use current visualizations of SHAP to effectively verify their hypotheses about an examined ML model (Kaur et al., 2020). Thus, with our framework we address the need for interactive exploration and verification of hypotheses. In a first step, we implement the follow-up operations proposed by Weld and Bansal (2019) for tabular data. From an initial triple of (*input*, *prediction*, *explanation)* provided by an XAI system the user can either:

- **Change the foil**: Contrast the triple with nearest-neighbour triples that resulted in a particularly different prediction to understand *"Why not prediction B?"*.

- **Restrict the subspace:** Request other triples that share the same value for one or more *input* features to understand "*How were similar inputs handled?*".
- **Sensitivity analysis**: Request the minimal changes required to one or more *input* features that result in a different *prediction* and *explanation* to understand *"How stable is the prediction?"*.
- **Explorative perturbation**: Change the values of one or more *input* features of an observation to explore the effects on the *prediction* and its *explanation* and to understand *"What if?"*.
- **Global roll-up**: Contrast the triple's *local explanation* with the *global explanation* of the entire model to understand *"How representative is the observation?"*.

An XAI system with interactive explanations may derive additional information about the user's mental model and preferences from the trail of follow-up interactions. This additional information may be used to establish common ground and potentially improve the overall human-AI system performance. With our framework we aim to support developers with the front-end development of XAI systems for domain experts. We consider domain experts as end-users with a high level of expertise in a particular domain but typically limited expertise in ML, such as lawyers or accountants. We focus on decision-making situations where the domain expert may have concrete or vague hypotheses about the decision problem that guides their explanation needs and interaction.

Future Work

Upcoming research will investigate the potentials of interactive explanations and their evaluation with users in an application context. We collaborate with German chancelleries, lawyers, and a leading software vendor in the sensitive legal domain. We follow a human-centered design process to derive requirements and user needs. Based on these, we iteratively explore design opportunities for usable interactive explanations using prototypes and user studies. We plan to integrate our insights and artifacts in a modular toolkit for creating interactive explanation interfaces for tabular and textual data.

References

Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. InIJCAI-17 workshop on explainable AI (XAI), Vol. 8. 1.

John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning.ACM Trans. Interact. Intell. Syst.8, 2, Article 8 (June 2018), 37 pages. https://doi.org/10.1145/3185517

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. DOI: https://doi.org/10.1145/3313831.3376219

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces. ACM, 126–137.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Advances in neural information processing systems, pp. 4765-4774.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence267 (2019), 1 – 38. https://doi.org/10.1016/j.artint.2018.07.007

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. Distill(2018). https://doi.org/10.23915/distill.00010 https://distill.pub/2018/building-blocks.

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

Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. Commun. ACM 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486

A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Martin Schuessler
Technische Universität Berlin
Berlin, Germany
schuessler@tu-berlin.de

ABSTRACT

The interdisciplinary field of explainable artificial intelligence (XAI) aims to foster human understanding of black-box machine learning models through explanation methods. However, there is no consensus among the involved disciplines regarding the evaluation of their effectiveness - especially concerning the involvement of human subjects. For our community, such involvement is a prerequisite for rigorous evaluation. To better understand how researchers across the disciplines approach human subject XAI evaluation, we propose developing a taxonomy that is iterated with a systematic literature review. Approaching them from an HCI perspective, we analyze which study designs scholar chose for different explanation goals. Based on our preliminary analysis, we present a taxonomy that provides guidance for researchers and practitioners on the design and execution of XAI evaluations. With this position paper, we put our survey approach and preliminary results up for discussion with our fellow researchers.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

explainable artificial intelligence; explanation; human evaluation; taxonomy.

ACM Reference Format:
Michael Chromik and Martin Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. In *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies (ExSS-ATEC'20) Cagliari, Italy*. 7 pages.

1 INTRODUCTION

We have witnessed the widespread adoption of intelligent systems into many contexts of our lives. Such systems are often built on advanced machine learning (ML) algorithms that enable powerful predictions – often at the expense of interpretability. As these systems are introduced into more sensitive contexts of society, there is a growing acceptance that they need to be capable of explaining their behavior in human-understandable terms. Hence, much research is conducted within the emerging domain of *explainable artificial intelligence* (XAI) and *interpretable machine learning* (IML) on developing models, methods, and interfaces that are interpretable to human users – often through some notion of explanation.

However, most works focus on computational problems while limited research effort is reported concerning their user evaluation. Previous surveys identified the need for more rigid empirical evaluation of explanations [2, 5, 17]. The AI and ML communities often strive for *functional* evaluation of their approaches with benchmark data to demonstrate generalizability. While this is suitable to demonstrate technical feasibility, it is also problematic since often *"there is no formal definition of a correct or best explanation"* [24]. Even if a formal foundation exists, it does not necessarily result in practical utility for humans as the utility of an explanation is highly dependent on the context and capabilities of human users. Without proper human behavior evaluations, it is difficult to assess an explanation method's utility for practical use cases [26]. We argue that functional and behavioral evaluation approaches have their legitimacy. Yet, since there is no consensus on evaluation methods, the comparison and validation of diverse explanation techniques is an open challenge [2, 4].

In this work, we take an HCI perspective and focus on evaluations with human subjects. We believe that the HCI community should be the driving force for establishing rigorous evaluation procedures that investigate how XAI can benefit users. Our work is guided by three research questions:

- **RQ-1**: Which evaluation approaches have been proposed and discussed across disciplines in the field of XAI?
- **RQ-2**: Which study design decisions have researchers made in previous evaluations with human subjects?
- **RQ-3**: How can the proposed approaches and study designs be integrated into a guiding taxonomy for human-centered XAI evaluation?

The contribution of this workshop paper is two-fold: First, we introduce our methodology for taxonomy development and literature review guided by RQ-1 and RQ-2. The review aims to provide an overview of how evaluations are currently conducted and help identify suitable best practices. As a second contribution, we present a preliminary taxonomy of human evaluation approaches in XAI and describe its dimensions. Taxonomies have been used in many disciplines to help researchers and practitioners to understand and analyze complex domains [23]. Our overarching goal is to synthesize a human subject evaluation guideline for researchers and practitioners of different disciplines in the field of XAI. With this work, we put our review methodology and preliminary taxonomy up for discussion with our fellow researchers.

ExSS-ATEC'20, March 2020, Cagliari, Italy

2 FOUNDATIONS AND RELATED WORK

2.1 Evaluating Explanations in Social Sciences

Miller defines explanation as either a process or a product [16]. On the one hand, an explanation describes the cognitive process of identifying the cause(s) of a particular event. At the same time, it is a social process between an *explainer* (sender of an explanation) and an *explainee* (receiver of an explanation) with the goal to transfer knowledge about the cognitive process. Lastly, an explanation can describe the product that results from the cognitive process and aims to answer a why-question. In our paper, we refer to explanations from the product perspective. Psychologists and social scientists investigated how humans evaluate explanations for decades. Within their disciplines, *explanation evaluation* refers to the process applied by an explainee for determining if an explanation is satisfactory [16]. Scholars conducted experiments where they presented participants with different types of explanations as treatments. These experiments indicate that choosing one explanation over another is often an arbitrary choice heavily influenced by cognitive biases and heuristics [12]. The primary criteria of explainees are whether the explanation helps them to understand the underlying cause [16]. For instance, humans are more likely to accept explanations that are consistent with their prior beliefs. Furthermore, they prefer explanations that are simpler (i.e., with fewer causes), and more generalizable (i.e., that apply to more events). Also, the effectiveness of an explanation depends on the current information needs of the explainee. A suitable explanation for one purpose may be irrelevant for another. Thus, for an explanation to be effective, it is essential to know the intended context of use.

2.2 Explainable Artificial Intelligence (XAI)

Interpretability in machine learning is not a monolithic concept [15]. Instead, it is used to indirectly evaluate whether important desiderata, such as fairness, reliability, causality, or trust, are met in a particular context [4]. Some definitions of interpretability are rather *system-centric*. Doshi-Velez and Kim [4] describe it as a model's *"ability to explain or to present in understandable terms to a human."* Miller [16] takes a more *human-centered* perspective calling it *"the degree to which an observer can understand the cause of a decision"*. Human understanding can be fostered either by offering means of introspection or through explanations [3]. A large variety of methods exist for both approaches [9]. The term *interpretable machine learning* (IML) often refers to research on models and algorithms that are considered as inherently interpretable while *explainable AI* (XAI) often refers to the generation of (post-hoc) explanations or means of introspection for black-box models [27, 33]. A model's black-box behavior may manifest itself in two ways: either from complex architectures, as with deep neural networks, or from proprietary models (that may otherwise be interpretable), as with the COMPAS recidivism model [27]. The lines between IML and XAI are often seamless and the terms are often used interchangeably. For instance, DARPA's XAI program subsumes both terms with the objective to *"enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"* [10].

2.3 Evaluating Explanations in XAI

Multiple surveys of the ever-growing field of XAI exist. They formalize and ground the concept of XAI [1, 2], relate it to adjacent concepts and disciplines [1, 16], categorize methods [9], or discuss future research directions [1, 2]. All these surveys report a lack of rigid evaluations. Adadi et al. found that only 5% of surveyed papers evaluate XAI methods and quantify their relevance [2]. Similarly, Nunes and Jannach found that 78% of the analyzed papers on explanations in decision support systems lacked structured evaluations that go beyond anecdotal "toy examples" [24].

Some works have addressed the design and conduction of explanation evaluations in XAI. Gilpin et al. survey explainable methods for deep neural networks and describe a categorization of evaluation approaches at different stages of the ML development process [8]. Yang et al. provide a framework consisting of multiple levels of explanation evaluation [33]. Their definition of *persuasibility* (measuring the degree of human comprehension) focuses on the human and resonates with our notion of human subject evaluation. Our work aims to elaborate on their generic strategy of *"employing users for human studies"*. Nunes and Jannach reviewed 217 publications spanning multiple decades and briefly report findings from applied evaluation approaches [24]. Based on their survey they derive a comprehensive taxonomy that guides the design of explanations. However, their taxonomy omits aspects of evaluation. Mueller identified 39 XAI papers that reported empirical evaluations and qualitatively described chosen evaluation approaches along 9 dimensions [20].

While these works offer valuable ideas, they are limited in their scope and, thus, offer little guidance for XAI user evaluations. Of course, *"there is no standard design for user studies that evaluate forms of explanations"* [24]. However, we believe that a unified taxonomy is needed that integrates the most common ideas related to human subject evaluation and extends them with best practice examples. Such an actionable format can provide great benefit for researchers and practitioners by guiding them through the design and reporting of structured XAI evaluations.

3 METHODOLOGY

In this section, we outline our method of taxonomy development as well as the planned literature review. Our goal is to develop a comprehensive taxonomy for human subject evaluations in XAI. We seek to validate and iterate it through a structured literature review (SLR). Figure 2 illustrates our proposed methodology and the interplay between taxonomy and SLR.

3.1 Taxonomy Development

There are two approaches to constructing a taxonomy. Following the *conceptual-to-empirical* approach, the researcher proposes a classification based on a theory or model (deductive). In contrast, the *empirical-to-conceptual* approach derives the taxonomy from empirical cases (inductive). We follow the iterative process for taxonomy development proposed by Nickerson et al. [23]. Their method unifies both approaches in an iterative process under a shared *meta-characteristic* and defined *ending conditions*.

In line with RQ-3, we defined our meta-characteristic as the development of *a taxonomy for human subject evaluation of black-box*

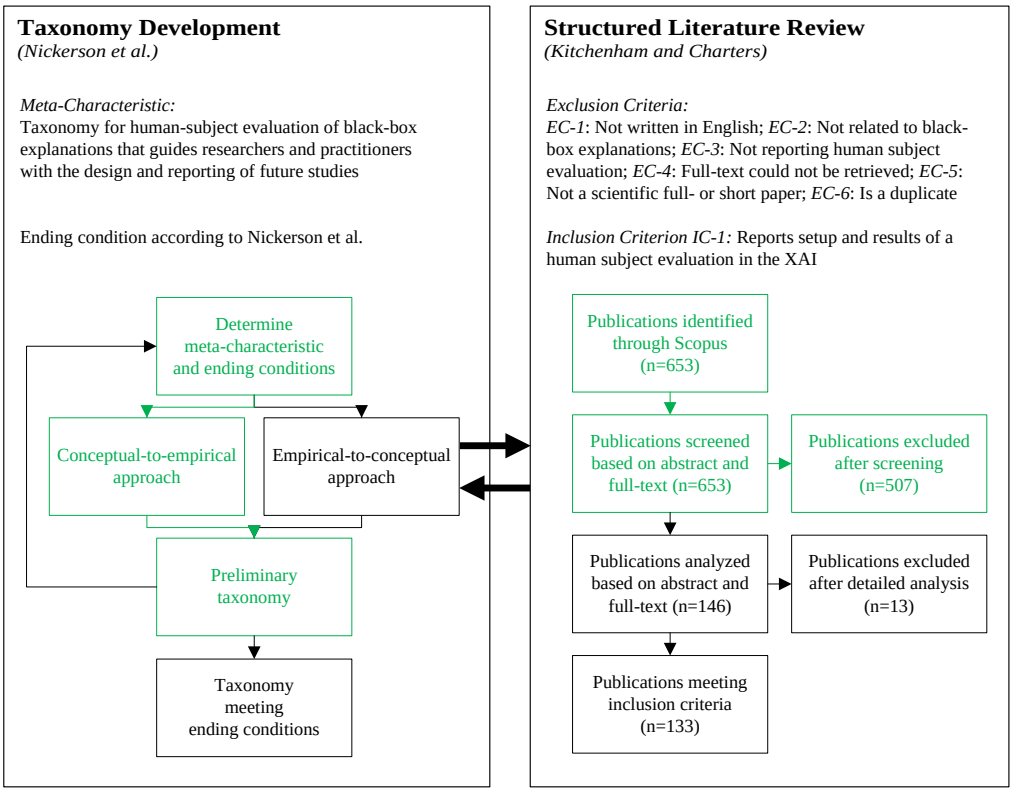


Figure 1: The proposed methodology for taxonomy development with an integrated structured literature review (SLR). Steps highlighted in green describe the preliminary results presented in this workshop paper.

explanations that guides researchers and practitioners with the design and reporting of future studies. We start by applying the *conceptual-to-empirical* approach. To follow this approach, one needs to propose a classification based on a theory or model. We do this by consolidating proposed categories for XAI evaluation in prior work and connecting them with foundational literature on empirical studies. The resulting taxonomy describes an ideal type, which allows us to examine empirically how much current human subject evaluations deviate from an ideal type.

3.2 Structured Literature Review

As part of the *empirical-to-conceptual* iteration, we aim to validate and iterate the taxonomy using a structured literature review (SLR). In line with RQ-2, the review's objective is to capture how researchers currently evaluate XAI methods and systems with human subjects. Through this, we seek to find out how structured and precise we can describe the field using our taxonomy. During this process, we also aim to iterate the taxonomy. The planned SLR follows established approaches proposed by Kitchenham and Charters [13]. In the following, we outline the proposed search strategy.

Source Selection: An exploratory search for XAI on Google Scholar indicated that relevant work is dispersed across multiple publishers, conferences, and journals. Thus, we use the Scopus database as a source as it integrates publications from relevant publishers such as ACM, IEEE, and AAAI.

Search Query: Through our exploratory search, we obtained an initial understanding of relevant keywords, synonyms, and related concepts that helped us to construct a search query. We found that different terms are used between the disciplines to describe the field of XAI and human subject evaluation approaches. Early research does not explicitly state the expressions *XAI* nor *explainable artificial intelligence*. Thus, our search queries are composed of *groups* and *terms*. Groups refer to a specific aspect of the research question and limit the search scope. Terms have a similar semantic meaning within the group domain or are often used interchangeably. We are interested in the intersection of 3 groups that can be phrased using different terms. Table 1 shows our used groups and terms.

Study Selection Criteria: We filtered the search results by six exclusion criteria (EC) and one inclusion criterion (IC). We are interested in primary studies that report the setup and result of human subject

Table 1: Groups and terms used for search query

Group	Terms
1 - Explainable	explainability, explainable, explanation, explanatory, interpretability, interpretable, intelligibility, intelligible, scrutability, scrutable, justification
2 - AI	XAI, AI, artificial intelligence, machine learning, black-box, recommender system, intelligent system, expert system, intelligent agent, decision support system
3 - Human Subject Evaluation	user study, lab study, empirical study, online experiment, human experiment, human evaluation, user evaluation, participant, within-subject, between-subject, probe, crowdsourcing, Mechanical Turk

evaluations in the XAI context (IC-1). We limit the survey to publications addressing the *black-box explanation problem*, according to Guidotti et al. [9] (EC-2). Furthermore, we exclude publications that do not report *human-grounded* or *application-grounded* evaluations according to Doshi-Velez and Kim [4] (EC-3). We applied the exclusion criteria in cascading order, i.e., if we excluded publications due to one EC, we did not assess any following criteria.

Study Analysis: So far, we conducted the search procedure for Scopus in September 2019, which returned a total of 653 potentially relevant publications. Both authors filtered the returned publications by the inclusions and exclusion criteria to control for inter-rater effects. We discussed differing assessments until we reached consensus. We are currently in the process of analyzing the publications that met the inclusion criterion.

4 TAXONOMY OF HUMAN SUBJECT EVALUATION IN XAI

In the following section, we describe relevant dimensions of black-box explanation evaluation with human subjects. We group identified characteristics into *task-related*, *participant-related*, and *study design-related* dimensions. The outlined taxonomy is a preliminary result after the first iterations of the conceptual-to-empirical approach based on propositions in prior work. Furthermore, the taxonomy was validated and refined based on a small subset consisting of 34 publications from the structured literature review following the empirical-to-conceptual approach.

4.1 Task Dimensions

Mohseni and Ragan distinguish two **types of human involvement** in the evaluation of explanations [18]. In the *feedback* setting, participants provide feedback on actual explanations. Experimenters determine the quality of the explanations through this feedback. In contrast, in the *feed-forward* setting no explanations are provided. Instead, humans are generating examples of reasonable explanations serving as a benchmark for algorithmic explanations.

Doshi-Velez and Kim distinguish two types of human subject evaluations that differ in their **level of task abstraction** [4]: *Application-grounded* evaluations conduct experiments within a real application context. Typically, this requires a high level of participant expertise. The quality of the explanation is assessed in measures of the application context, typically with a test of performance. *Human-grounded* evaluations conduct simplified or abstracted experiments that aim to maintain the essence of the target application.

Multiple **types of user tasks** have been proposed to elicit the quality of explanations [4, 18, 33]. We suggest distinguishing them by the information provided to the participant and the information inquired in return. In *verification* tasks, participants are provided with input, explanation, and output and asked for their satisfaction with the explanation. *Forced choice* tasks extend this setting. Here, participants are asked to choose from multiple competing explanations. In the case of *forward simulation* tasks, participants are presented with inputs as well as explanations and need to predict the system's output. *Counterfactual simulation* tasks, present participants with an input, an explanation, an output, and an alternative output (the counterfactual). Based on these, they predict what input changes are necessary to obtain the alternative output. In *"Clever Hans" detection* tasks, participants need to identify and possibly debug flawed models, e.g., a naive or short-sighted predictor [14]. *System usage* tasks are characterized by participants using the system and its explanations for its primary purpose, e.g., a decision-making situation. The quality of the explanation is assessed in terms of decision quality. In *annotation* tasks, participants provide a suitable explanation given input and output of a model.

Explanations are provided to users with very different goals in mind. For their effective evaluation, researchers need to ensure that the **intended explanation goal(s)** are aligned with their intended evaluation goal(s), and vice versa. Also, calibration of the individual goals of participants with the intended explanation goal(s) might be necessary (e.g., through a briefing before the task) [31]. We distinguish 9 common explanation goals, which are derived from [24, 30, 32]: *transparency* aims to explain how the system works, *scrutability* aims to allow users to tell the system it is wrong, *trust* aims to increase the user's confidence in the system, *persuasiveness* aims to convince the user to perform an action, *satisfaction* aim to increase the ease of use or enjoyment, *effectiveness* aims to help users make good decisions, *efficiency* aims to make decisions faster, *education* aims to enable users to generalize and learn, *debugging* aims to enable users to identify defects in the system. In the case of multiple intended explanation goals, their dependencies may be complementary, contradictory, or even unknown (e.g., the impact of transparency on trust).

Hoffman et al. describe multiple **levels of task evaluation** to assess a participant's understanding of and XAI system. Furthermore, they discuss suitable metrics for each level [11]. *Tests of satisfaction* measure participants' self-reported satisfaction with an explanation and their perception of system understanding. On this level, researchers can rarely be sure whether participants understand the system to the degree that participants claim. *Tests of comprehension* assess the participants' mental models of the system and tests their understanding, for example, through prediction tests and generative exercises. *Tests of performance* measure the resulting human-XAI system performance.

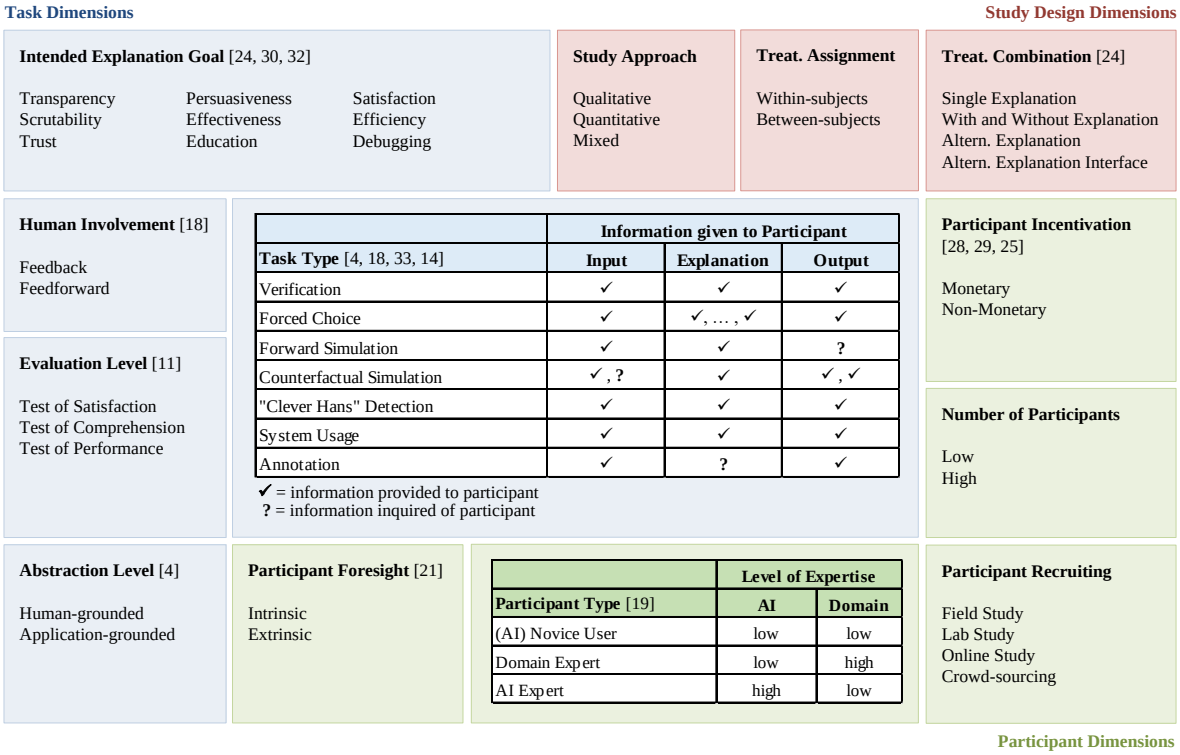


Figure 2: Preliminary taxonomy of human subject evaluation in XAI based on the conceptual-to-empirical approach.

4.2 Participant Dimensions

Mohseni et al. distinguish between several **participant types**: *AI novices* who are usually end-users, data experts (including *domain experts*), and *AI experts* [19]. This distinction is important as user expertise strongly influences other participant-related dimensions. For example, Doshi-Velez and Kim [4], referencing the work of Neath and Surprenant [22], point out that user expertise determines what kind of cognitive chunks participants apply to a situation. The expertise of participants may determine the **recruiting method** and **number of participants**. Recruiting difficulty is likely to increase with the required level of participants' expertise [4]. One can recruit novices in large numbers via *crowd-sourcing*. In contrast, domain or AI experts are usually harder to identify and recruit. They are often invited to a targeted *online study*, a *lab study*, or a *field study*. According to Narayana et al., the user study task may have dependencies with the **level of participant foresight** [21]. In an *intrinsic* setting, the participant's understanding of the context is solely based on the provided information. Thus, all participants are assumed to have equal knowledge about the context. Such types of experiments are usually suitable for novices. In an *extrinsic* setting, participants can additionally draw upon external facts, such as prior experience, that may be relevant for assessing the quality of an explanation, e.g., for spotting model flaws. Such a setting may be more suitable for data experts. However, it also makes controlling for participants' knowledge more difficult.

Incentivization of participants is another relevant dimension. According to Sova and Nielsen, it should be chosen considering study length, task demand, and participant expertise [28]. Stadtmüller and Porst advise us to use a *monetary incentive* for participants [29]. However, several *non-monetary incentives* are known to be effective as well (e.g., gifts for already paid employees) [25, 28]. Prost and Briel found that participants may take part in a study because of study-related incentives (e.g., curiosity, sympathy, or entertainment), personal-incentive (e.g., professional interest or a promise made), or altruistic reasons (e.g., to benefit science, society, or others) [25]. Esser argues that researchers should consider incentives in their combination such that the benefits of participating out-weigh the perceived cost [6].

4.3 Study Design Dimensions

The study design of evaluations may follow a *qualitative*, *quantitative*, or *mixed* **study approach**. In experimental studies, experimenters assign treatments to groups of participants. Applied to the context of explanation evaluations, we can distinguish four common types of **treatments combinations** in line with Nunes and Jannach [24]: *single treatment* (i.e., no alternative treatment), *with*

and without explanation (i.e., no explanation is alternative treatment), *alternative explanation* (i.e., varying information provided in explanations between treatments with other aspects of user interface fixed), *alternative explanation interface* (i.e., varying user interfaces between treatments). Furthermore, we can distinguish study designs by the **treatment assignment**: *Between-subjects designs* study the differences in understanding between groups of participants, each usually assigned to one treatment. In contrast, *within-subject designs* study differences within individual participants who are assigned to multiple treatments.

5 LIMITATION AND FUTURE WORK

Our preliminary taxonomy has limitations. The taxonomy is neither collectively exhaustive nor mutually exclusive. Thus, it does not meet the ending conditions of taxonomy development [23]. We aim to refine and iterate the taxonomy with the results from the proposed structured literature review.

Furthermore, human subject evaluations in XAI are typically embedded in a broader context, which may create dependencies and limit applicable evaluation approaches. Dependencies may arise from the explanation design context, such as the form of an explanation, its contents, or its underlying generation method. Multiple taxonomies have been developed for guiding the design of explanations [7, 24]. Nunes and Jannach proposed an elaborate explanation design taxonomy [24]. However, their taxonomy omits aspects of evaluation. For now, we have abstained from relating our preliminary human subject evaluation taxonomy with this prior work, but plan to integrate them in later iterations.

6 CONCLUSION

In this work, we gave a brief overview of recent efforts on explanation evaluation with human subjects in the growing field of XAI. We proposed a methodology for developing a comprehensive taxonomy for human subject evaluation that integrates the knowledge from multiple disciplines involved in XAI. Based on ideas from prior work, we presented a preliminary taxonomy following the conceptual-to-empirical approach. Despite its limitations, we believe our work is a starting point for rigorously evaluating the utility of explanations for human understanding of XAI systems. Researchers and practitioners developing XAI explanation facilities and systems have been asked to *"respect the time and effort involved to do such evaluations"* [4]. We aim to spark a discussion at the workshop on how to support them along the way.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

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

[3] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, Vol. 8. 1.

[4] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretability. *CoRR* abs/1702.08608 (2017). arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[5] F. K. Dosilovic, M. Brcic, and N. Hlupic. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

[6] Hartmut Esser. 1986. Über die Teilnahme an Befragungen. *ZUMA Nachrichten* 10, 18 (1986), 38–47.

[7] Gerhard Friedrich and Markus Zanker. 2011. A Taxonomy for Generating Explanations in Recommender Systems. *AI Magazine* 32, 3 (Jun. 2011), 90–98. https://doi.org/10.1609/aimag.v32i3.2365

[8] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89. https://doi.org/10.1109/DSAA.2018.00018

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (aug 2018). https://doi.org/10.1145/3236009

[10] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* 40, 2 (Jun. 2019), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

[11] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *CoRR* abs/1812.04608 (2018). arXiv:1812.04608 http://arxiv.org/abs/1812.04608

[12] Frank C. Keil. 2006. Explanation and Understanding. *Annual Review of Psychology* 57, 1 (2006), 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100 arXiv:https://doi.org/10.1146/annurev.psych.57.102904.190100 PMID: 16318595.

[13] B. Kitchenham and S Charters. 2007. Guidelines for performing Systematic Literature Reviews in Software Engineering. *Keele University and University of Durham, Technical Report EBSE-2007-01* (2007).

[14] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10, 1 (2019), 1096.

[15] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages. https://doi.org/10.1145/3236386.3241340

[16] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1 – 38. https://doi.org/10.1016/j.artint.2018.07.007

[17] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. In *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*. http://people.eng.unimelb.edu.au/tmiller/pubs/explanation-inmates.pdf

[18] Sina Mohseni and Eric D. Ragan. 2018. A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning. *CoRR* abs/1801.05075 (2018). arXiv:1801.05075 http://arxiv.org/abs/1801.05075

[19] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *CoRR* abs/1811.11839 (2018). arXiv:1811.11839 http://arxiv.org/abs/1811.11839

[20] Shane T. Mueller, Robert R. Hoffman, William J. Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *CoRR* abs/1902.01876 (2019). arXiv:1902.01876 http://arxiv.org/abs/1902.01876

[21] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *CoRR* abs/1802.00682 (2018). arXiv:1802.00682 http://arxiv.org/abs/1802.00682

[22] Ian Neath and Aimee Surprenant. 2002. *Human Memory* (2 edition ed.). Thomson/Wadsworth, Australia ; Belmont, CA.

[23] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3 (2013), 336–359. https://doi.org/10.1057/ejis.2012.26 arXiv:https://doi.org/10.1057/ejis.2012.26

[24] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Dec. 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[25] Rolf Porst and Christa von Briel. 1995. *Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen? Oder: Gründe für die Teilnahme an Panelbefragungen*. Vol. 1995/04.

[26] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and Measuring Model Interpretability. *CoRR* abs/1802.07810 (2018). arXiv:1802.07810 http://arxiv.org/abs/1802.07810

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

[28] Deborah Hinderer Sova and Jacob Nielsen. 2003. How to Recruit Participants for Usability Studies. https://www.nngroup.com/reports/how-to-recruit-participants-usability-studies/, accessed December 20th, 2019.

[29] Sven Stadtmüller and Rolf Porst. 2005. *Zum Einsatz von Incentives bei postalischen Befragungen*. Vol. 14.

[30] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop (ICDEW '07)*. IEEE Computer Society, Washington, DC, USA, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[31] Nadya Vasilyeva, Daniel A Wilkenfeld, and Tania Lombrozo. 2015. Goals Affect the Perceived Quality of Explanations.. In *CogSci*.

[32] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 601.

[33] Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating Explanation Without Ground Truth in Interpretable Machine Learning. *CoRR* abs/1907.06831 (2019). arXiv:1907.06831 http://arxiv.org/abs/1907.06831

Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems

Michael Chromik
LMU Munich
Munich, Germany
michael.chromik@ifi.lmu.de

Malin Eiband
LMU Munich
Munich, Germany
malin.eiband@ifi.lmu.de

Sarah Theres Völkel
LMU Munich
Munich, Germany
sarah.voelkel@ifi.lmu.de

Daniel Buschek
LMU Munich
Munich, Germany
daniel.buschek@ifi.lmu.de

ABSTRACT

The rise of interactive intelligent systems has surfaced the need to make system reasoning and decision-making understandable to users through means such as *explanation facilities*. Apart from bringing significant technical challenges, the call to make such systems explainable, transparent and controllable may conflict with stakeholders' interests. For example, intelligent algorithms are often an inherent part of business models so that companies might be reluctant to disclose details on their inner workings. In this paper, we argue that as a consequence, this conflict might result in means for explanation, transparency and control that do not necessarily benefit users. Indeed, we even see a risk that the actual virtues of such means might be turned into *dark patterns*: user interfaces that purposefully deceive users for the benefit of *other* parties. We present and discuss such possible dark patterns of explainability, transparency and control building on dark UX design patterns by Grey et al. The resulting dark patterns serve as a thought-provoking addition to the greater discussion in this field.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Explainability; Explanations; Transparency; Dark Patterns; Interpretability; Intelligibility; User Control; Intelligent Systems.

ACM Reference Format:
Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019*. 6 pages.

1 INTRODUCTION

Intelligent systems that are empowered by advanced machine learning models have successfully been applied in closed contexts to well-structured tasks (e.g., object recognition, translations, board games) and often outperform humans in those. These advancements fostered the introduction of intelligent systems into more sensitive contexts of human life, like courts, personal finance or recruiting, with the promise to augment human decision-making in those.

However, the effectiveness of intelligent systems in sensitive contexts cannot always be measured in objective terms. Often they need to take soft factors, like safety, ethics and non-discrimination, into account. Their acceptance will greatly depend on their ability to make decisions and actions interpretable to its users and those affected by them. Introducing interpretability through *explanation facilities* [15] is widely discussed as an effective measure to support users in understanding intelligent systems [9, 24]. Yet, these measures are located at the intersection of potentially conflicting interests between decision-subjects, users, developers and company stakeholders [36].

First, companies may not see the benefit to invest in potentially costly processes to include explanations and control options for users *unless* they improve their expected revenues in some way. Second, creating suitable explanations of algorithmic reasoning presents a major technical challenges in itself that often requires abstraction from the algorithmic complexity [28, 29]. Furthermore, those systems are often integrated with critical business processes. Companies might be reluctant to disclose explanations that honestly describe their reasoning to the public as it might have an impact on their reputation or competitive advantage. Forcing companies to do so by law, like the *right to explanation* as part of the European Union *General Data Protection Regulation (GDPR)* [32], will most likely not result in meaningful explanations for users.

Therefore, we see a danger that means for algorithmic explanation, transparency and control might not always be designed by practitioners to *benefit* users. We even see a risk that users might consciously be deceived for the benefit of *other* parties. Such carefully crafted deceptive design solutions have gained notoriety in the UI design community as *dark patterns* [3].

In this paper, we extend the notion of prominent dark UX patterns [13] to algorithmic explanation, transparency and control. We discuss situations of opposing interests between the creator and receiver of algorithmic explanation, transparency and control means that could be potentially argued as questionable or unethical and contribute to the discussion about the role of design practitioners in this process.

IUI Workshops'19, March 20, 2019, Los Angeles, USA

2 BACKGROUND

2.1 Explanations in Intelligent Systems

Haynes et al. define intelligent systems as *"software programs designed to act autonomously and adaptively to achieve goals defined by their human developer or user"* [15]. Intelligent systems typically utilize a large knowledge data base and decision-making algorithms. Following Singh [31], a system is intelligent if users need to *"attribute cognitive concepts such as intentions and beliefs to it in order to characterize, understand, analyze, or predict its behavior"*.

Many of the intelligent systems developed today are based on increasingly complex and non-transparent machine learning models, which are difficult to understand for humans. However, sensitive contexts with potentially significant consequences often require some kind of human oversight and intervention. Yet, even intelligent systems in everyday contexts often confuse users [11]. For example, social network users are not aware that the news feed is algorithmically curated [6]. These insights result in ongoing research activities to improve the interpretability of those systems. *Interpretability* is the degree to which a human can understand the cause of a decision [26]. Interpretability can be achieved either by *transparency* of the model's inner workings and data, or *post-hoc explanations* that convey information about a (potentially) approximated cause – just like a human would explain [24].

Different stakeholders (e.g., creator, owner, operator, decision-subjects, examiner) of an intelligent system may require different means of interpretability [35]. Creators may demand *transparency* about the system's algorithms, while operators might be more interested how well the system's conceptual model fits their mental model (*global explanation*). Decision-subjects, on the other hand, may be interested in the factors influencing their individual decision (*local explanation*). This paper focuses on the interplay between owners of intelligent systems and decision-subjects using it.

Explanation facilities [15] are an important feature of usable intelligent systems. They may produce explanations in forms of textual representations, visualizations or references to similar cases [24]. The explanations provided may enable users to better understand why the system showed a certain behaviour and allow them to refine their mental models of the system. Following Tomsett [35] we define *explainability* as the level to which a system can provide clarification for the cause of its decision to its users.

Previous research work suggests that explanation facilities increase users' trust towards a system [23, 28] and user understanding [10, 18, 20]. However, how to present effective and usable explanations in intelligent systems is still a challenge that lacks best practices [22]. Due to the complexity of intelligent systems, explanations can easily overwhelm users or clutter the interface [18]. Studies by Bunt et al. [7] indicate that the costs of reading explanations may outweigh the perceived benefits of users. Moreover, some researchers warn that it may also be possible to gain users' trust with the provision of meaningless or misleading explanations [36]. This might leave users prone to manipulation and give rise to the emergence of dark patterns.

2.2 Dark Patterns

In general, a *design pattern* is defined as a proven and generalizing solution to a recurring design problem. It captures design insights in a formal and structured way and is intended to be reused by other practitioners [12]. Design patterns originate from architecture [1], but have been adopted in other fields such as software engineering [12], proxemic interaction [14], interface design [33], game design [37], and user experience design [13]. In contrast, an *anti pattern* refers to a solution that is commonly used although being considered ineffective and although another reusable and proven solution exists [17].

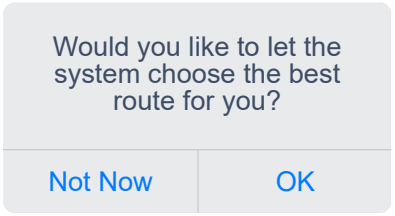


Figure 1: Exemplary interface for the *Restricted Dialogue* dark pattern. Users are not given a "No" option.

In 2010, Harry Brignull coined the term *dark pattern* [3] to describe *"a user interface that has been carefully crafted to trick users into doing things [...] with a solid understanding of human psychology, and they do not have the user's interests in mind"* [5]. He contrasts dark patterns to "honest" interfaces in terms of trading-off business revenue and user benefit [4]: while the latter put users first, the former deliberately deceive users to increase profit within the limits of law. Brignull [3] identified twelve different types of dark patterns and collects examples in his "hall of shame". Gray et al. [13] further clustered these dark patterns into five categories: *Nagging, Obstruction, Sneaking, Interface Interference* and *Forced Action*.

3 DARK PATTERNS OF EXPLAINABILITY, TRANSPARENCY AND CONTROL

What makes a pattern *dark* in the context of explainability, transparency and control? We see two general ways: the *phrasing* (of an explanation), and the way it is integrated and depicted in the *interface* (of explanation facilities). We build on the five categories of dark UX design patterns by Gray et al. [13] and apply them to the context of explainability, transparency, and user control, along with concrete examples (Table 1).

3.1 Nagging

Nagging is defined as a *"redirection of expected functionality that may persist over one or more interactions"* [13]. Transferred to the context of this paper, Nagging interweaves explanation and control with other, possibly hidden, functionality and thus forces users to do things they did not intend to do or interrupts them during their "actual" interaction.

3.1.1 Example 1: Restricted Dialogue. One example that Gray et al. present in their paper are pop-up dialogues that do not allow permanent dismissal. This could be easily transferred to our context: for example, an intelligent routing system could take control away from users with the tempting offer "Would you like to let the system

Dark Pattern by Gray et. al. [13]	**Transfer** to Explainability and Control	**Example Phrasings** of Explanation	**Example Interfaces** of Explanation Facilities
Nagging: *"redirection of expected functionality that may persist over one or more interactions"*	Interrupt users' desire for explanation and control	Restricted Dialogue	Hidden Interaction
Obstruction: *"making a process more difficult than it needs to be, with the intent of dissuading certain action(s)"*	Make users shun the effort to find and understand an explanation while interacting with explanation or control facilities	Information Overload, Nebulous Prioritization	Hidden Access, Nested Details, Hampered Selection
Sneaking: *"attempting to hide, disguise, or delay the divulging of information that is relevant to the user"*	Gain from user's interaction with explanation/control facilities through hidden functions	Explanation Marketing	Explanation Surveys
Interface Interference: *"manipulation of the user interface that privileges certain actions over others."*	Encourage explainability or control settings that are preferred by the system provider	Unfavorable Default	Competing Elements, Limited View
Forced Action *"Requiring the user to perform a certain action to access [...] certain functionality"*	Force users to perform an action before providing them with useful explanations or control options	Forced Data Exposure, Tit for Tat	Forced Dismissal

Table 1: Examples of dark patterns in the *phrasing* of explanations and the *interface* of explanation facilities. The examples are built upon the categorization by Gray et al. [13].

choose the best route for you?", where users can only select "Not now" or "OK", but have no "No" option (see Figure 1).

3.1.2 Example 2: Hidden Interaction. Nagging might include linking on-demand explanations with hidden advertisements: A click on "Why was this recommended to me?" on an ad could indeed open the explanation, but also the ad link (e.g., in two browser tabs).

3.2 Obstruction

Gray et al. define *Obstruction* in UX design as *"making a process more difficult than it needs to be, with the intent of dissuading certain action(s)"*. In the context of this paper, Obstruction makes it hard to get (useful) explanations about the system's decision-making and to control the algorithmic settings. Users thus might shun from the additional effort this takes and rather accept the system as is.

3.2.1 Example 1: Information Overload. Moreover, the use of very technical language to explain system behaviour and decision-making, or very lengthy explanations would most probably discourage users from reading the given information at all (see Figure 3. This might be comparable to what we currently see in end user licence agreements: the use of very technical language *and* a very lengthy presentation format results in users skipping the system prompt [2].

3.2.2 Example 2: Nebulous Prioritization. When explaining a decision or recommendation with a large number of influencing factors, the system might limit those factors by some notion of "importance" to not overwhelm the user. However, limiting factors requires a (potentially arbitrary) prioritization, which might be used to obfuscate sensitive factors, like family or relationship statuses. The explanation could be framed vaguely (e.g., "This recommendation is based on factors such as..." – i.e. not claiming to present all factors).

3.2.3 Example 3: Hidden Access. One way to obstruct the path to information could be to avoid "in-situ" links to explanations (e.g., offer no direct explanation button near a system recommendation). Instead, the option for explanation and control could be deeply hidden in the user profile and thus difficult to access.

3.2.4 Example 4: Nested Details. Similarly, the information *detail* could be distributed, for example nested in many links: When users want to have more than a superficial "This was shown in your feed, because you seem to be interested in fashion", they would have to take many steps to reach the level of detail that satisfies their information need.

3.2.5 Example 5: Hampered Selection. The system could also make activating explanations tedious for users by forcing them to do this for, say, every single category of product recommendation without giving a "select all" option. This could resemble the difficult cookie management practices seen today on many ad-financed websites. In another example setting, the information in an intelligent routing system could be spread along different sections of the recommended route and thus would have to be activated for each section separately.

3.3 Sneaking

The dark pattern of *Sneaking* is defined as *"attempting to hide, disguise, or delay the divulging of information that is relevant to the user"* [13]. Following this dark pattern, systems could use UI

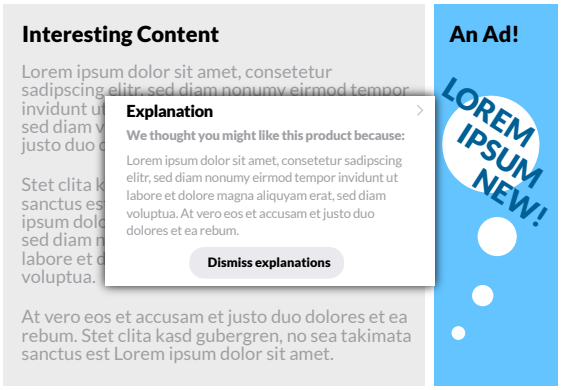


Figure 2: Exemplary interface for the *Limited View* dark pattern. Users are encouraged to dismiss explanations since they are layouted in a way that annoyingly covers the main content of the website.

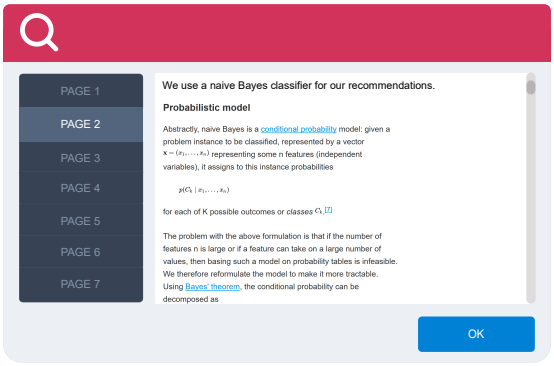


Figure 3: Exemplary interface for the *Information Overload* dark pattern. The given explanation is lengthy and uses technical language not suitable for non-experts (example article copied from Wikipedia).

elements for explainability and control, to sneak in information motivated by different intentions than interpretability.

3.3.1 Example 1: Explanation Marketing. For example, a web advertisement service could explain a particular ad by showing previously seen ads which the user had seemed to be interested in. Thus, the user's interest in an explanation is utilized to present multiple (potentially paid) advertisements. In a similar fashion, an online shop could use the opportunity of explaining product recommendations to promote further products. Also, ads might be directly integrated into the phrasing of explanations. For instance, an intelligent maps application might explain its routing decisions along the lines of "This route is recommended because it passes by the following stores you've visited in the past...".

3.3.2 Example 2: Explanation Surveys. Another approach might present an explanation and ask users for feedback in order to improve future explanations. This way, a company might enrich its user data and utilize it apart from explanation.

3.4 Interface Interference

Gray et al. [13] define this dark pattern as *"manipulation of the user interface that privileges certain actions over others."* In our context, this dark pattern privileges UI settings and user states that do *not* contribute to – or actively suppress – explainability, transparency, and user control.

3.4.1 Example 1: Unfavorable Default. For example, a dark pattern in this category could preselect a "hide explanations" option during the user onboarding in a financial robo-advisor system. This could be motivated to the user as "uncluttering" the dashboard or UI layout in general.

3.4.2 Example 2: Limited View. Explanations and control elements could also be layouted in a way that significantly reduces the space for the actual content or interferes with viewing it. This could encourage users to dismiss explanations to increase usability. Even simpler, links to an explanation might be presented in a barely visible manner. Figure 2 shows an example.

3.4.3 Example 3: Competing Elements. Further integration of explanations with the system's business model might involve, for instance, starting a count down timer upon opening an explanation for a booking recommendation to compete for the user's attention. This timer could indicate a guaranteed price or availability, thus putting pressure on the user to abandon the explanation view in order to continue with the booking process.

3.5 Forced Action

This dark pattern is defined as *"requiring the user to perform a certain action to access (or continue to access) certain functionality"* [13]. In our context, the user could be forced to perform an action that (also) dismisses functionality or information related to explainability, transparency and control.

3.5.1 Example 1: Forced Data Exposure. This dark pattern could be used to collect valuable user data under the pretext of explanation. The user might be forced to provide further personal information (e.g., social connections) before receiving personalized explanations. Otherwise, the user would be left off with a generic high-level explanation.

3.5.2 Example 2: Forced Dismissal. A user could be forced to dismiss an explanation pop-up in order to see the results of a request displayed underneath (e.g., during the investment process of a robo-advisor system). This dismissal might be interpreted as a permanent decision to no longer display any explanations.

3.5.3 Example 3: Tit for Tat. Regarding transparency, an e-commerce recommender system might force the user to first confirm an action (e.g., place an order) before it displays the factors that influenced

the recommendation. For instance, the system might proclaim that so far not enough data is available to explain its recommendation.

4 SUMMARY AND DISCUSSION

In this paper, we presented possible dark patterns of explanation, transparency and control of intelligent systems based on the categorization of dark UX design patters by Gray et al. [13]. We see the possibility that simple legal obligations for explanation might result in dark patterns rather than user benefits (e.g., similar to cookies settings on many ad-financed websites). Instead, with our work we intend to promote the on-going research on explainability as well as the discussion on explanation standards and their effects on users.

4.1 What Are the Consequences of Dark Patterns?

We see several possibly negative consequences of dark patterns in this context: Users might be annoyed and irritated by explanations, developing a negative attitude towards them. Examples include explanations presented in the *Nagging* patterns, which automatically open an advertisement along with the explanation; *Forced Action* patterns, which hinder the user to access desired results; or *Sneaking* patterns, which disguise advertisements as explanations. Similarly, users might lose interest in explanations when *Interface Interference* or *Obstruction* patterns are applied, which e.g., show long and tedious to read explanations. As a consequence, users might dismiss or disable explanations entirely.

On the other hand, users might not recognize explanations when they are hidden in profile settings. When users know that intelligent systems *must* provide explanations by law, the absence of explanations might mistakenly make users believe that the system does not use algorithmic decision-making. Hence, users might develop an incorrect understanding of algorithmic decision-making in general.

Furthermore, *Obstruction* patterns might lead to explanations which promote socially acceptable factors for algorithmic decision-making and withhold more critical or unethical ones. As a result, this might hinder the formation of correct mental models of the system's inner workings. Hence, users might not be able to critically reflect on the system's correctness and potential biases. As previous work in psychology suggests, users might accept *placebo* explanations without conscious attention as long as no additional effort is required from them [21]. When explanations use very technical language and are difficult to understand, users might simply skip them. This lack of knowledge and uncertainty about the underlying factors influencing the algorithm might lead to *algorithmic anxiety* [16].

4.2 Which Further Dark Patterns May Appear in this Context?

In this paper, we transferred the dark pattern categories by Gray et al. [13] to explainability and control of intelligent systems. However, there might be further patterns in this context. For example, we propose a pattern based on *Social Pressure* that uses information about other people – who are relevant to the user – in a way that is likely to be unknown or not endorsed by those people. For example, when Bob is shown an advertisement for diet products, explained by "Ask Alice about this", he might be annoyed with Alice without her knowledge. Similarly, Alice's boss might be recommended a lingerie shop that also "Alice might be interested in".

4.3 How Do Dark Patterns Affect Complex Ecosystems?

In this paper, we examined dark patterns which deceive decision-subjects who have means of directly interacting with the intelligent system. However, the ecosystem model of an intelligent system might be more complex and involve multiple stakeholders [35]. For example, in a financial decision-support context the system could ascertain the creditworthiness of a person (decision-subject), but only present an incontestable subset of reasons to the bank employee (operator) to not impact the reputation of the company (owner).

4.4 Can All Aspects of Dark Patterns Be Avoided?

Intelligent systems often use machine learning algorithms, which have hundreds of input variables. If all of these variables are explained, the explanation consists of a long list of text, which we identified as a dark pattern above. On the other hand, if they only show a subset of input variables for an explanation, this might bias the user's mental model, which is another dark pattern. Some explanations might be easier to understand for users than others. Hence, future studies have to evaluate which explanations are most helpful for users to understand the system.

4.5 How Can Dark Patterns Inform Research and Design?

In general, reflecting on dark patterns can be useful for HCI researchers and practitioners to learn how to do things properly by considering how not to do them. As a concrete use case, dark patterns can serve as a baseline for empirical studies to evaluate new design approaches: For example, a new explanation design could be compared against a placebo explanation – and not (only) against a version of the system with no explanation at all. Finally, dark patterns raise awareness that having *any* explanations is not sufficient. Instead, they motivate the HCI community to work on specific guidelines and standards for explanations to make sure that these actually support users in gaining awareness and understanding of algorithmic decision-making.

5 CONCLUSION

The prevalence of intelligent systems poses several challenges for HCI researchers and practitioners to support users to successfully interact with these systems. Explanations of how an intelligent system works can offer positive benefits for user satisfaction and control [19, 34], awareness of algorithmic decision making [27], as well as trust in the system [8, 25, 30]. Since 2018, companies are legally obliged to offer users a *right to explanation*, enshrined in the *General Data Protection Regulation* [32].

However, providers of intelligent systems might be reluctant to integrate explanations that disclose system reasoning to the public

in fear of a negative impact on their reputation or competitive advantage. Hence, legal obligations alone might not result in useful facilities for explanation and control for the end user.

In this paper, we have drawn on the notion of dark UX patterns [3] to outline questionable designs for explanation and control. These arise from explanation facilities that are not primarily designed with the users' benefits in mind, but purposely deceive users for the benefit of other parties.

In conclusion, we argue that while a legal right to explanation might be an acknowledgement of the necessity to support users in interacting with intelligent system, it is not sufficient for users nor our research community. By pointing to potential negative design outcomes in this paper, we hope to encourage researchers and practitioners in HCI and IUI communities to work towards specific guidelines and standards for "good" facilities for explanation, transparency and user control.

REFERENCES

[1] Christopher Alexander, Sara Ishikawa, Murray Silverstein, Max Jacobson, Ingrid Fiksdahl-King, and Shlomo Angel. 1977. *A Pattern Language: towns, buildings, construction.* Oxford University Press, Oxford, UK.

[2] Omri Ben-Shahar. 2009. The Myth of the "Opportunity to Read" in Contract Law. *European Review of Contract Law* 5, 1 (2009), 1–28.

[3] Harry Brignull. 2010. Dark Patterns. darkpatterns.org.

[4] Harry Brignull. 2011. Dark Patterns: Deception vs. Honesty in UI Design. https://alistapart.com/article/dark-patterns-deception-vs.-honesty-in-ui-design, accessed November 28, 2018.

[5] Harry Brignull. 2014. Dark Patterns: User Interfaces Designed to Trick People. http://talks.ui-patterns.com/videos/dark-patterns-user-interfaces-designed-to-trick-people, accessed November 28, 2018.

[6] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (jan 2017), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

[7] Andrea Bunt, Matthew Lount, and Catherine Lauzon. 2012. Are explanations always important?: A study of deployed, low-cost intelligent interactive systems. In *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI '12)*. ACM, New York, NY, USA, 169–178. https://doi.org/10.1145/2166966.2166996

[8] Henriette Cramer, Bob Wielinga, Satyan Ramlal, Vanessa Evers, Lloyd Rutledge, and Natalia Stash. 2009. The effects of transparency on perceived and actual competence of a content-based recommender. *CEUR Workshop Proceedings* 543 (2009), 1–10. https://doi.org/10.1007/s11257-008-9051-3

[9] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv e-prints* (Feb. 2017). https://arxiv.org/abs/1702.08608

[10] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (jun 2018), 1–37. https://doi.org/10.1145/3185517

[11] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When People and Algorithms Meet:Assessing User-reported Problems to Inform Support inIntelligent Everyday Applications. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. ACM, New York, NY, USA.

[12] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. 1994. *Design Patterns: Elements of Reusable Object-Oriented Software.* Addison Wesley, Boston, MA, USA.

[13] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. 2018. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 534, 14 pages. https://doi.org/10.1145/3173574.3174108

[14] Saul Greenberg, Sebastian Boring, Jo Vermeulen, and Jakub Dostal. 2014. Dark Patterns in Proxemic Interactions: A Critical Perspective. In *Proceedings of the 2014 Conference on Designing Interactive Systems (DIS '14)*. ACM, New York, NY, USA, 523–532. https://doi.org/10.1145/2598510.2598541

[15] Steven R Haynes, Mark A Cohen, and Frank E Ritter. 2009. Designs for explaining intelligent agents. *International Journal of Human-Computer Studies* 67, 1 (2009), 90–110. https://doi.org/10.1016/j.ijhcs.2008.09.008

[16] Shagun Jhaver, Yoni Karpfen, and Judd Antin. 2018. Algorithmic Anxiety and Coping Strategies of Airbnb Hosts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 421, 12 pages. https://doi.org/10.1145/3173574.3173995

[17] Andrew Koenig. 1998. Patterns and Antipatterns. In *The Patterns Handbooks*, Linda Rising (Ed.). Cambridge University Press, New York, NY, USA, 383–389.

[18] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[19] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678

[20] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, New York, NY, USA, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[21] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology* 36, 6 (1978), 635–642. http://dx.doi.org/10.1037/0022-3514.36.6.635

[22] Brian Y. Lim and Anind K. Dey. 2011. Design of an intelligible mobile context-aware application. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '11)*. ACM, New York, NY, USA, 157–166. https://doi.org/10.1145/2037373.2037399

[23] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why* and *why not* explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[24] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[25] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. 2017. Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. In *Advances in Human Factors in Robots and Unmanned Systems*, Pamela Savage-Knepshield and Jessie Chen (Eds.). Springer International Publishing, Cham, 127–136. https://doi.org/10.1007/978-3-319-41959-6_11

[26] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *CoRR* abs/1706.07269 (2017). arXiv:1706.07269 http://arxiv.org/abs/1706.07269

[27] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173677

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.

[30] James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O'Donovan. 2015. Getting the Message?: A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 345–356. https://doi.org/10.1145/2678025.2701406

[31] Munindar P. Singh. 1994. *Multiagent systems.* Springer, Berlin, Heidelberg, Germany, 1–14. https://doi.org/10.1007/BFb0030532

[32] The European Parliament and Council. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* (2016).

[33] Jenifer Tidwell. 2012. *Designing interfaces: Patterns for Effective Interaction Design.* O'Reilly Media, Inc., Sebastopol, Canada. arXiv:arXiv:gr-qc/9809069v1

[34] N. Tintarev and J. Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*. IEEE, New York, NY, UDA, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[35] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv e-prints* (June 2018). http://arxiv.org/abs/1806.07552

[36] Adrian Weller. 2017. Challenges for Transparency. *CoRR* (2017). http://arxiv.org/abs/1708.01870

[37] José P Zagal, Staffan Björk, and Chris Lewis. 2013. Dark patterns in the design of games. In *Foundations of Digital Games 2013*.

Eidesstattliche Versicherung

(Siehe Promotionsordnung vom 12.07.11, § 8, Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eidesstatt, dass die Dissertation von mir selbstständig und ohne unerlaubte Beihilfe angefertigt wurde.

München, den 09.04.2021 Michael Chromik