Dec 14th, 12:00 AM

Fostering Human Agency: A Process for the Design of User-Centric XAI Systems

Maximilian Förster
University of Ulm, maximilian.foerster@uni-ulm.de

Mathias Klier
University of Ulm, mathias.klier@uni-ulm.de

Kilian Kluge
University of Ulm, kilian.kluge@uni-ulm.de

Irina Sigler
University of Ulm, irina.hardt@uni-ulm.de

Fostering Human Agency: A Process for the Design of User-Centric XAI Systems

Completed Research Paper

Maximilian Förster
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
maximilian.foerster@uni-ulm.de

Mathias Klier
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
mathias.klier@uni-ulm.de

Kilian Kluge
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
kilian.kluge@uni-ulm.de

Irina Sigler
University of Ulm
Helmholtzstr. 22
89081 Ulm, Germany
irina.sigler@uni-ulm.de

Abstract

The emerging research field of Explainable Artificial Intelligence (XAI) addresses the problem that users do not trust or blindly follow AI systems that act as black boxes. XAI research to date is often criticized for not putting the user at the center of attention. Against this background, we design a process to systematically guide the instantiation, calibration, and quality control of XAI systems such that they foster human agency and enable appropriate trust in AI systems. The process can be applied independent of the XAI method, application domain, and target user group. It incorporates the principles of user-centric design, insights into explanations from the social sciences, and established XAI evaluation scenarios. Following the Design Science methodology, we demonstrate the practical applicability of our artifact and evaluate its efficacy in a realistic setting. Our work contributes to the design of user-centric XAI systems and the quest for human agency in AI.

Keywords: Explainable Artificial Intelligence, User-Centric Design, Human-AI Interaction

Introduction

Artificial Intelligence (AI) is increasingly employed for a wide range of tasks, first and foremost, decision support (HLEG-AI 2019). However, "Without AI systems [..] being demonstrably worthy of trust, unwanted consequences may ensue and their uptake might be hindered" (HLEG-AI 2019, p. 4). Thus, AI systems need to guarantee human agency (HLEG-AI 2019), as otherwise, users might either blindly follow an AI system's recommendation or merely distrust and not use it (Herse et al. 2018; Rader and Gray 2015). The key impediment to human agency is the fact that many AI systems appear as "black boxes" that do not provide users with sufficient information to make an informed choice regarding their recommendations (Guidotti et al. 2019b; Wachter et al. 2018).

In light of this challenge, the research field of Explainable Artificial Intelligence (XAI) aims at AI systems that are both highly performant and empower their users to comprehend, appropriately trust, and scrutinize them (Abdul et al. 2018; DARPA 2017). In particular, XAI provides approaches to automatically generate explanations along with AI systems' outputs (Rai 2020). In this context, explanations are human-

understandable lines of reasoning for why an AI system maps a given input to a specific output (Abdul et al. 2018). As the primary motivation for providing explanations is to enable human agency (HLEG-AI 2019; Nunes and Jannach 2017), the user-centricity of explanations is a prerequisite (Ribera and Lapedriza 2019). User-centricity is the "extent to which a system, product, or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241-210 2019, sec. 3.13). However, while substantial progress has been made in developing and demonstrating XAI methods (Barredo Arrieta et al. 2020), research to date is criticized for not putting the users at the center of attention (Kirsch 2018; Mittelstadt et al. 2019). Recent research has begun to address this call by examining insights from social sciences (Miller 2019) and evaluating explanations from users' perspectives (Doshi-Velez and Kim 2018; Förster et al. 2020; Weerts et al. 2019). However, while these efforts yield first valuable insights into a better understanding of XAI users, findings remain fragmented, and users are still not systematically incorporated into the development of XAI methods. This creates a situation best described as "inmates running the asylum," with researchers constructing explanations they themselves appreciate rather than explanations that generate value for their users (Miller et al. 2017; Mittelstadt et al. 2019). Indeed, to the best of our knowledge, no approach exists that effectively and systematically guides researchers and practitioners in the user-centric design of XAI systems.

Against this background, we propose a novel IT artifact that guides researchers and practitioners in instantiating, calibrating, and controlling the quality of user-centric XAI systems. The design of our artifact is informed by prior work on user-centric design, insights into understanding XAI users, and methods to evaluate XAI systems. Our artifact takes the shape of a process inspired by well-established processes in the fields of data mining and data science. Following the Design Science methodology (Hevner et al. 2004; Sonnenberg and vom Brocke 2012), we demonstrate and rigorously evaluate our process by applying it to a use case transferable to other AI applications. Our contribution to research and practice is twofold. First, we conceptualize and evaluate a user-centric XAI process to guide researchers and practitioners in the design of XAI systems. Second, we demonstrate how to effectively incorporate processes from data mining and data science, principles of user-centric design, insights from the social sciences into characteristics, structures, and presentation modes of explanations, and evaluation frameworks in XAI into a unified process.

The remainder of this paper is structured as follows: In the next section, we discuss relevant literature in the fields of user-centricity, XAI, as well as data mining and data science that inform the design of our artifact. Subsequently, we propose a process for instantiating, calibrating, and controlling the quality of a user-centric XAI system. Then, we demonstrate the applicability of the artifact and evaluate its efficacy. Afterward, we discuss the implications of our research for theory and practice, reflect on limitations of our work, and conclude with directions for further research.

Theoretical Background

Explanations for AI decisions

A significant issue of many state-of-the-art AI systems is their opacity, or "black box" character, which means that their inner workings are so intricate that the reasons for their decisions appear impenetrable to the user (Guidotti et al. 2019b). A prominent example of opaque systems are deep neural networks (Doran et al. 2018), which are comprised of a stack of layers of artificial neurons, whose outputs depend (typically non-linearly) on the outputs of the neurons in the next-lower layer (Goodfellow et al. 2016). Opacity induces critical challenges regarding the adoption of AI systems and resulting consequences. First, opacity hinders AI's societal acceptance, as it contributes to users' distrust in the AI's decisions and consequently reduces their willingness to consider or accept recommendations (Herse et al. 2018). Second, opacity impedes human agency, as users lack the information and transparency needed to reflect critically on an AI system's decision before following or acting on it (Rader and Gray 2015). Explanations that accompany the AI system's decisions can provide the level of transparency needed to scrutinize AI decisions (HLEG-AI 2019; Nunes and Jannach 2017), enabling users to appropriately trust the system (DARPA 2017; HLEG-AI 2019). Accordingly, they are seen as a promising path in the quest for a trustworthy AI (HLEG-AI 2019).

In light of the challenges of both low AI adoption due to a lack of users' trust and the harmful consequences of AI systems that impede human agency, the research field of XAI provides algorithms for automatically generating explanations (Doran et al. 2018). The call for explanations for AI systems has attracted

considerable attention from researchers. For an overview, see the reviews by Barredo Arrieta et al. (2020) and Guidotti et al. (2019b). XAI systems, in their most basic form, consist of an algorithm generating explanations for an AI system and an explanation interface (DARPA 2017). Often, XAI algorithms are model-agnostic (Rai 2020, Guidotti et al. 2019b) and generate "post hoc interpretations" (Lipton 2018, p. 6). Thus, they can be used for any kind of AI system while not influencing its performance.

XAI explanations build on elements such as visualizations, feature-relevance, or counter-examples, with most approaches using a combination thereof. Visualizations convey the reasons for a decision through plots and graphics, e.g., partial dependence plots (Green and Kern 2010). Feature-relevance explanations measure the importance each feature has in generating the output, e.g., through estimation of Shapley values (Lundberg and Lee 2017). Algorithms generating counter-examples go a step further and explain a decision by contrasting it to another comparable decision (Wachter et al. 2018), inspired by how humans construct explanations themselves (Lipton 2000). XAI literature suggests two main lines of approaches to finding a suitable counter-example such that the explanation is meaningful: algorithms relying on locally approximating the AI system with a simpler model from which explanations are derived (Guidotti et al. 2019a) and algorithms computing explanations directly from the AI system, often framing the search for a counter-example as an optimization problem (Dhurandhar et al. 2019; Wachter et al. 2018). While the technical realization of XAI methods is an essential prerequisite, the call for explanations goes beyond providing post hoc interpretation for an AI system's output (Miller et al. 2017; Mittelstadt et al. 2019). It requires solutions that not only explain the recommendations of an AI system to users who do not understand its inner workings but further enable these users to contest and alter a recommendation (Doran et al. 2018; Wachter et al. 2018). In light of these challenges, to fully support human agency, the research field of XAI needs to place its users at the center of attention (Abdul et al. 2018; Kirsch 2018).

User-Centric XAI

In the context of human agency, individuals are seen as "contributors to their life circumstances, not just products of them" (Bandura 2006). In line with this definition, the Independent High-Level Expert Group on AI set up by the European Commission demands that AI systems guarantee human agency, as "users should be able to make informed autonomous decisions regarding AI systems. They should be given the knowledge and tools to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system." (HLEG-AI 2019, p. 16). Explanations are a crucial element of allowing for such informed decisions, as they aim to enable the subject to understand the reasons for a decision as well as put them into a position to contest or affect it (Wachter et al. 2018). In this line of thought, empowering users to control and appropriately trust an AI system is the primary motivation for providing explanations (HLEG-AI 2019; Nunes and Jannach 2017). Thus user-centricity of explanations is a prerequisite (Ribera and Lapedriza 2019). Still, whereas "researchers in the ML and AI communities are working on making their algorithms explainable, their focus is not on usable, practical and effective transparency that works for and benefits people" (Abdul et al. 2018, p. 10). Indeed, while XAI research provides a wide array of algorithms to produce a diverse range of explanations for AI recommendations, it remains unclear what the end-user needs to scrutinize and appropriately trust an AI system (Förster et al. 2020; Wang et al. 2019).

User-centric design might answer this call, as it provides a design approach to developing solutions that focus on the users' needs and wants (Norman and Draper 1986). The establishment of user-centric design in the 1980s (cf. Norman and Draper 1986) marks a milestone in product and service development (Still and Crane 2017). In general, user-centric design is an approach that aims at improving usability, namely the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241-210 2019, sec. 3.13). In the context of XAI, this can be transferred to an approach that puts XAI users, whether they are laypeople or domain experts, at the center of attention and enables them to achieve the goal of empowering users to control and appropriately trust an AI system (Ribera and Lapedriza 2019). The premise of user-centric design has inspired a broad range of methods and principles (cf. Still and Crane 2017). With their fundamental principles, Gould and Lewis (1985) proposed an early focus on users and tasks, empirical measurement, and iterative design as crucial elements of user-centric design. The guidelines put forward by IDEO, a leading design-agency, additionally identify empathy and the need to learn from failure as vital principles (IDEO 2015). Likewise, guidelines such as the "People + AI Guidebook" by Google (2019) that specifically focus on human-centric AI products emphasize the need to consider user-centricity throughout

the entire product development flow and provide guidance on, e.g., identifying user needs and the design of feedback mechanisms. The generic ISO guideline for "human-centered design" provides a well-established framework, frequently used in academia and practice. It proposes six principles (ISO 9241-210 2019): First, the need for understanding user, task, and environment. Second, user involvement in design and development. Third, user-centric evaluation. Fourth, the need for an iterative process that, fifth, addresses the entire user experience. Sixth, a multidisciplinary team with diverse skills and perspectives is required. Based on these principles, the ISO guideline identifies four key activities that a user-centric design process needs to entail, namely, understanding the context, specifying user requirements, producing the solution, and evaluating the solution. If necessary, several iterations of these activities are to be performed until a satisfactory solution can be instantiated. Together, the principles and activities serve as general guidelines to achieve a user-centric design that can be adapted for application in specific contexts. For instance, Farinango et al. (2015) integrated them into user-centric software development processes.

Research into explanations for AI systems that represent human-understandable lines of reasoning and enable human subjects to gain control when interacting with an AI system and develop an appropriate level of trust (Abdul et al. 2018) can build on strong foundations (cf. Wang et al. 2019). Social sciences find that an explanation's "loveliness" contributes to its "likeliness" (Lipton 2000) and point out specific characteristics, structures, and presentation modes of explanations beyond factual correctness that can contribute to user appreciation. First, social sciences literature identifies explanation characteristics, e.g., shortness (Thagard 1989), that are appreciated in human-human interaction and hence might inform the design of XAI systems (cf. Förster et al. 2020; Miller 2019). Second, regarding the basic structure of an explanation, research refers to how humans construct explanations themselves, suggesting XAI methods to produce contrastive explanations (e.g., Wachter et al. 2018). These explanations do not list all causes that lead to a specific event but focus on why an AI system yielded a particular output (the *fact*) instead of another, similarly perceivable one (the *foil*) (cf. Lipton 1990). The difference between the fact and the foil, the *contrast*, explains the output. In the case of the rejection of a new credit line, the fact refers to the customer's situation (e.g., income and savings) leading to the rejection. The foil refers to a counterfactual scenario that would bring about an approval (e.g., higher income). The contrast is the difference between the customer's situation and the counterfactual scenario (e.g., difference in income). Aside from characteristics and structure, specific modes of presentation can improve intelligibility. These include, among others, visual, textual, symbolic, audible, audio-visual, or tabular (Wang et al. 2019). For example, Huysmans et al. (2011) found that decision tables are especially comprehensible, while Ribera and Lapedriza (2019) demonstrated that visualization serves to support decision-making in healthcare.

Prior work on the evaluation of automatically generated explanations provides first insights on the incorporation of user-centricity into the design of XAI systems. In this context, Doshi-Velez and Kim (2018) propose three scenarios for the evaluation of explainable systems. The first, *functionally-grounded evaluation*, does not require human involvement. One potential approach is to test against proxy measures for explanations, e.g., the length of an explanation as a measure for its simplicity or complexity, respectively (Martens and Provost 2014; Wachter et al. 2018). While efficient in terms of time and resource requirements, it remains unclear whether such proxy measures truly reflect the users' perception of explanations. Thus, the second scenario, *human-grounded evaluation*, is conducted with human subjects undertaking a simplified task to assess the quality of explanations from users' perspective (Förster et al. 2020; Mohseni and Ragan 2018; Weerts et al. 2019). The third scenario, *application-grounded evaluation* with real users in a real application setting, can serve as the final step in evaluating usability and effectiveness (Abdul et al. 2018). Practitioners and researchers are confronted with trade-offs when choosing the most suitable evaluation scenario, most notably the trade-off between including users and required effort. On the one hand, conducting experiments with human subjects is crucial for lowering the risk of being misled by assumptions that do not reflect users' perception (Weerts et al. 2019). On the other hand, both expenditure of time and costs are generally substantially higher for the human-grounded compared to the functionally-grounded scenario (Doshi-Velez and Kim 2018).

To sum up, while first valuable insights into user-centric explanations that can inform the design of XAI methods have been reported, to the best of our knowledge, no process for the systematic application of user-centric principles to the design of XAI systems exists. This situation poses a challenge for researchers and practitioners looking to incorporate user-centricity in the design of XAI methods, as they face a highly fragmented state of knowledge.

Data Mining and Data Science Processes

Incorporating principles of user-centric design into the design of systems is a challenge that arises beyond the field of XAI. In the broadest sense, the design of information systems always entails the challenge of solving technical tasks while meeting pre-defined objectives. To address this challenge, research areas related to XAI, e.g., data mining and data science, rely on processes (Martinez-Plumed et al. 2019). More specifically, these processes guide projects by translating business goals into well-defined technical tasks (Marbán et al. 2009). This general approach can be transferred to the design of XAI systems, as – similar to data and machine-learning models – their design requires an explorative approach, the translation of pre-defined business objectives into technical metrics (e.g., accuracy), and statistical testing of the solution. Thus, we take processes from the fields of data mining and data science as a starting point to incorporate user-centric objectives into the design of XAI systems. Data mining literature defines a process as a series of steps that are executed in sequence. It can include loops and iterations, which are "triggered by a revision process" (Kurgan and Musilek 2006, p. 4). Data mining processes typically contain three stages: They begin with steps to understand the business goals and context, followed by data preparation and analysis, and conclude with the evaluation, interpretation, and application of the results (Kurgan and Musilek 2006; Martinez-Plumed et al. 2019). As an example, consider CRISP-DM, the de facto standard process for data mining in research and practice (Martinez-Plumed et al. 2019). This process comprises six steps, namely, business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al. 2000). Many researchers based their further developed processes on CRISP-DM (Martinez-Plumed et al. 2019). For instance, Gertosio and Dussauchoy (2004) expanded the process to include increased user involvement. More recently, faced with new challenges such as ever-larger data volumes and the rise of machine learning, both IBM and Microsoft released updated and more versatile variants (Microsoft 2020; Rollins 2015). Data mining and data science processes serve to minimize risks through different validation steps, to reveal and remedy faults, and to facilitate resource allocation (Marbán et al. 2009; Rollins 2015). Due to their flexibility and scalability, they can be applied independent of project size and domain (Marbán et al. 2009). Further, the processes provide a general and replicable framework that allows projects to be executed by staff with diverse backgrounds (Moyle and Jorge 2001). Finally, the clear goal-definition enforced by data science processes fosters alignment between team members (Microsoft 2020). To sum up, processes in the areas of data mining and data science can inform the incorporation of user-centricity into the design of XAI systems. In particular, their basic structure can serve as a blueprint when designing a novel process for the design of user-centric XAI systems.

Research Gap

In order to be beneficial to individuals and society, the proliferation of AI in everyday life requires that users are able "to comprehend and interact with AI systems to a satisfactory degree and, where possible, be enabled to reasonably self-assess or challenge the system" (HLEG-AI 2019, p. 4). In this regard, especially the opacity inherent to many AI systems is an impediment, as it leads to human subjects facing AI decisions without the means required to understand and contest them. Against this background, in the quest for trustworthy AI, the emerging research field of XAI aims to provide explanations that foster human agency (HLEG-AI 2019; Nunes and Jannach 2017). Over the past years, the automatic generation of explanations received tremendous research attention that resulted in the development of a wide range of algorithms (Barredo Arrieta et al. 2020). However, a majority of these have not yet been evaluated with human users, and the application of XAI systems in real-world contexts is still in its infancy (Abdul et al. 2018; Adadi and Berrada 2018). While first studies recently addressed the call for putting the user in the center of research attention (Kirsch 2018), e.g., examining insights from social sciences (Miller 2019) and evaluating explanations from users' perspectives (Förster et al. 2020), these findings remain fragmented. To the best of our knowledge, no process exists that systematically guides the design of user-centric XAI systems. Therefore, researchers and practitioners alike are at a high risk of designing XAI systems that do not provide value for their users (Miller et al. 2017; Mittelstadt et al. 2019). To close this gap, following the Design Science methodology (Hevner et al. 2004), we design and evaluate a novel process to instantiate, calibrate, and control the quality of user-centric XAI systems. Our process focuses on model-agnostic XAI methods and the end-users of XAI systems, such as domain experts and laypeople. The process can be applied to any application domain that entails an AI system augmenting human decision-making, excluding systems that fully automate it (Martin 2019). It places the users at the center of attention while striking a balance between costly and time-consuming user testing and calibration based on mathematical constructs.

A Novel Process to Design User-Centric XAI Systems

We design a novel process to instantiate, calibrate, and control the quality of an XAI system such that it is user-centric in that it enables and fosters human agency (DARPA 2017; HLEG-AI 2019). To this end, we design our "User-Centric XAI Process" (cf. Figure 1) based on well-established processes in the field of data mining and data science (cf. Martinez-Plumed et al. 2019) and the principles of user-centric design (ISO 9241-210 2019). The process integrates the evaluation framework for explainable systems by Doshi-Velez and Kim (2018) and incorporates research on explanations in the social sciences (cf. Miller 2019).

Basic Idea

We begin by briefly revisiting the problems researchers and practitioners face when designing XAI systems. First of all, XAI methods are novel algorithms that have yet to stand the test of practice and time (Wolf 2019). Further, contrary to the task-driven development of AI systems, in the context of XAI, users should be at the center of attention (Preece et al. 2018). Importantly, user-centricity in XAI reaches far beyond usability, as human agency is the primary goal, and further legal and ethical concerns demand consideration (HLEG-AI 2019). Due to a lack of experience and best practices to draw from, real-world XAI applications are at risk of failing their users and falling short of their stakeholders' high expectations (Weerts et al. 2019). Not because XAI methods are inherently incapable – quite the converse (Barredo Arrieta et al. 2020) – but because the designers of XAI systems lack the means to shift their focus from technical aspects to their lay or domain-expert end-users (Miller et al. 2017). Our artifact addresses this problem space based on the three core concepts sequential structure, user-centricity, and iterative calibration.

First, we design our artifact to provide a *sequential structure* that systematically guides the design of XAI systems. Inspired by well-established processes in the related fields of data mining and data science (Martinez-Plumed et al. 2019), we design a process comprising the phases Instantiation, Calibration, and Quality Control (cf. Kurgan and Musilek 2006). The Instantiation phase focuses on examining the application requirements as well as selecting and instantiating the XAI system. In the subsequent Calibration phase, the XAI system is adapted to produce explanations that fulfill the application-specific requirements in an iterative sequence of calibration and user testing. Finally, in the Quality Control phase, the deployed XAI system is continuously monitored and evaluated to assess its efficacy. Designing the artifact as a process exhibits three main advantages: First, a process provides a structured and replicable framework to systematically develop complex systems. Second, the process prescribes the definition of precise and unambiguous goals and ensures that they are not lost out of sight throughout the potentially lengthy and intertwined stages of system development. Third, the process places the technical development of the XAI system in the context of its users and the team designing it.

Second, our artifact emphasizes *user-centricity* by placing the end-user at the center of attention, as arguably, the users are the most critical stakeholders of XAI systems (Preece et al. 2018). On the one hand, usability is crucial for the successful application of an XAI system, as users have to accept and interact with it. On the other hand, fostering human agency is the primary goal in the design of XAI systems. However, users are rarely considered in XAI research to date (Kirsch 2018). By incorporating the principles of user-centric design (ISO 9241-210 2019) and building on XAI literature on user-centric explanations (cf. Wang et al. 2019), our "User-centric XAI process" ensures that the user is in focus at all times. For one, the Instantiation phase fosters a thorough understanding of the task, user, and environment, ensuring that the entire user experience is taken into account from the very beginning. Both the subsequent Calibration and Quality Control phase suggest an iterative approach guided by user feedback (Gould and Lewis 1985; IDEO 2015; ISO 9241-210 2019). Further, the design process is informed by insights from the social sciences regarding characteristics and presentation modes of explanations (cf. Miller 2019; Wang et al. 2019), such as the human preference for contrastive explanations or the need for explanations to be coherent.

Third, our artifact integrates the complementary XAI evaluation scenarios proposed by Doshi-Velez and Kim (2018) into a unified process of *iterative calibration* to enable efficient yet user-centric design of XAI systems. While user testing and involvement are indispensable to ensure user-centricity (ISO 9241-210 2019; Weerts et al. 2019), it is costly and time-consuming. Hence, it cannot be carried out continuously, but only on selected occasions. Against this background, we integrate functionally-grounded and human-grounded evaluation by interlinking them with proxy measures, i.e., mathematical constructs that reflect the user-centric requirements for the XAI system (Doshi-Velez and Kim 2018). Functionally-grounded

evaluation serves to calibrate the XAI system through optimizing its parametrization to specified target values of these proxy measures. This activity requires no user involvement and is thus economical in time and costs. To rigorously validate that the proxy measures truly reflect the users' perspective, we employ human-grounded evaluation, which tests explanations with human subjects, often on a simplified task that aims to capture the essential elements and features of the application setting (Doshi-Velez and Kim 2018).

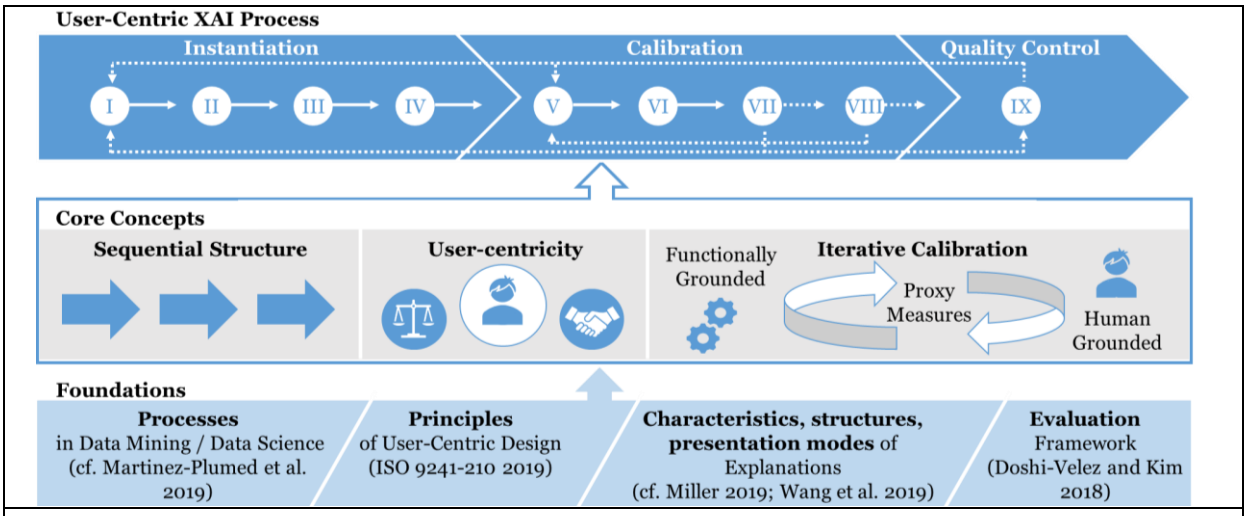

Figure 1. Basic idea and underlying core concepts of the User-Centric XAI Process

In the following, we describe the "User-Centric XAI Process" with its three phases and its nine corresponding steps (cf. Table 1) in detail. As a starting point, we assume a ready-to-use AI system that is accessible throughout the process. In line with the fifth principle of user-centric design (ISO 9241-210 2019), the process is conducted by a team with diverse backgrounds in, e.g., IS, AI, social sciences, or business-specific domains (cf. Mittelstadt et al. 2019).

Phases and Steps

The **Instantiation phase**, which consists of four steps, is devoted to understanding the XAI system's application domain and the target users as well as defining requirements for the XAI system. Step I aims to build an understanding of the specific XAI application context. First, in line with typical data mining processes (Chapman et al. 2000; Kurgan and Musilek 2006; Microsoft 2020), the team investigates the XAI system's intended application domain and the corresponding business background. More concretely, ethical, legal, and regulatory requirements, as well as the availability of resources such as data and computation infrastructure, are examined (cf. Chapman et al. 2000; Martinez-Plumed et al. 2019). Second, informed by the first principle of user-centric design, i.e., the need for understanding user, task, and environment (Gould and Lewis 1985; IDEO 2015; ISO 9241-210 2019), the level of domain expertise the user possesses, the purpose of the XAI system for the users, and the risks associated with a lack of user-centric explanations are examined (Beaudouin et al. 2020). Next to the primary goal of enabling human agency, further aims, e.g., legal requirements derived from the responsibility for consequences resulting from the AI system's use, are captured (Beaudouin et al. 2020; HLEG-AI 2019; Mittelstadt et al. 2019). Further, the team needs to establish an understanding of the role the AI system plays in the decision-making process (Google 2019). More specifically, this includes placing the intended application on the augmentation-automation continuum and understanding the respective roles of the human agent and the AI system (Martin 2019). Additionally, the team should identify the potential for decision-making bias in the intended application (cf. Baron 2014; Tversky and Kahneman 2015) and investigate whether literature already suggests potential mitigation measures in the application domain, e.g., as in the case of medical diagnosis (cf. Lighthall and Vazquez-Guillamet 2015; Wang et al. 2019). Based on the developed understanding of the XAI system's application context and purpose, in Step II, the team derives a list of user-centric and technical application requirements from the users' perspective (cf. IBM 2016; Moyle and Jorge 2001). Then, the team needs to identify whether these requirements can be translated into specific modes of presentation and desired characteristics of explanations, respectively. More concretely, technical

modes of presentation concern how the explanations are conveyed to the user and range from visual, textual, or symbolic presentation to audible, audio-visual, or interactive explanations (Wang et al. 2019). The choice of a mode of presentation is informed by the AI system's technical structure and its application context, with a focus on user experience (cf. ISO 9241-210 2019). Characteristics of explanations describe their perception by the user, e.g., concrete, coherent, or relevant (Förster et al. 2020), and reflect the intention of the explanations. At this early stage of the process, the precise characteristics required to fulfill the intentions are not yet known. Hence, the team identifies initial characteristics guided by the literature (Miller 2019) or user interviews (Hall et al. 2019). While prior XAI research indicates a variety of requirements in different user groups and settings (Hall et al. 2019; Kirsch 2018; Miller 2019; Ribera and Lapedriza 2019; Wang et al. 2019), systematically building an in-depth user understanding constitutes an innovative contribution to XAI. Equipped with user-centric and technical requirements, in Step III, the team selects an XAI method as the foundation of the XAI system. Similar to the selection of models in data mining processes (Kurgan and Musilek 2006) and data science (Goodfellow et al. 2016; Microsoft 2020), the selection is informed by the team's expertise and can require substantial literature research. In line with Hall et al. (2019), to be considered as the foundation of the XAI system, an XAI method has to be capable of generating explanations with the modes of presentation identified in Step II. Moreover, it needs to offer sufficient flexibility in its parametrization, such that its explanations can be tuned to exhibit the desired characteristics. The instantiation of the XAI system concludes Step III. At the end of the Instantiation phase, in Step IV, a set of preliminary proxy measures is derived. Proxy measures reflect the intended characteristics of explanations and are computed from the XAI system's output (Doshi-Velez and Kim 2018). The preliminary proxy measures can be based on examples from the literature (Dhurandhar et al. 2019; Guidotti et al. 2019b; Mothilal et al. 2020; Wachter et al. 2018) or, at this point in the process, be derived based on the team's intuition.

The four steps of the subsequent **Calibration phase** aim at calibrating the XAI system according to iteratively refined requirements. The alternation between functionally-grounded and human-grounded evaluation exploits the resource efficiency of the former. At the same time, the latter addresses the issue that most proxy measures available from the XAI literature have not been validated with users and are not equally applicable in all contexts (Adadi and Berrada 2018; Doshi-Velez and Kim 2018; Förster et al. 2020). The goal of Step V is to find a parametrization of the XAI system that leads to optimal explanations as indicated by the proxy measures. To systematically explore the XAI system's potentially vast parameter space, well-established methods for tuning machine-learning algorithms, such as grid search (Goodfellow et al. 2016), can be employed. In terms of the framework by Doshi-Velez and Kim (2018), each calculation of proxy measures for a possible parametrization of the XAI system constitutes a functionally-grounded evaluation. While research into XAI methods often ends with this step, generating explanations that satisfy the researchers' intuition (Miller et al. 2017; Mittelstadt et al. 2019), an explanation "is not a mathematical construct" but "the users are the leveling rule" (Kirsch 2018). Thus, to test if explanations exhibit the desired characteristics from the users' perspective, in Step VI, in accordance with the principles of user-centric design (ISO 9241-210 2019), we conduct a user test. More concretely, in a scenario of human-grounded evaluation (Doshi-Velez and Kim 2018), users evaluate explanations from the XAI system against competing and control explanations on a simplified task. Control explanations can, for instance, be ones written by experts (Förster et al. 2020) or instantiations of the XAI system from previous iterations. As far as possible, participants of the user test should be representative of the target audience. In the case of lay users, online platforms can provide access to a diverse demographic (Buhrmester et al. 2011). The user test yields data on how the explanations generated by the XAI system are perceived compared to control and competing explanations (Doshi-Velez and Kim 2018; Förster et al. 2020). Moreover, the test reveals which characteristics are most important from the users' perspective (Förster et al. 2020). In Step VII, the results of the user test are analyzed in two ways. First, the desired characteristics of explanations from the users' perspective are derived. Second, it is determined to what extent the XAI system produces explanations that exhibit these desired characteristics. This analysis constitutes a validation step, echoing the concept of validation in data science processes (Marbán et al. 2009; Rollins 2015): If the results are satisfying, the XAI system is fit for deployment, and the process can continue with the Quality Control phase. If a large fraction of explanations generated by the XAI system is found to be unsuitable, this might hint at oversights or incorrect assumptions in previous steps. In this case, the process should be discontinued and resumed with Step I. In all other cases, the results of the user test inform the adaption of the desired characteristics of explanations and the refinement of proxy measures in Step VIII. The desired characteristics are assessed and, if necessary, adapted based on both the conducted analysis and the investigation in Step I. Using the

data collected in the user test, it is analyzed whether each proxy measure indeed reflects its corresponding characteristic. If it does not, it is discarded. Subsequently, new proxy measures are introduced for characteristics that are not yet or not sufficiently captured. The refinement of the proxy measures concludes an iteration of the Calibration phase; the process proceeds with a new iteration, beginning with Step V.

The final **Quality Control phase** is devoted to continuous evaluation of the deployed XAI system under real-world conditions. First, the output of the XAI system is monitored, similar to the monitoring and maintenance of applications developed in data mining and data science processes (Microsoft 2020; Moyle and Jorge 2001). In a scenario of continuous functionally-grounded evaluation (Doshi-Velez and Kim 2018), it is observed if generated explanations fall in the identified target range of the final set of proxy measures, e.g., through a dashboard (Microsoft 2020). Second, corresponding to the application-grounded evaluation scenario described by Doshi-Velenz and Kim (2018), the effects of the XAI system on its users are observed. One potential area of investigation might be the presence of cognitive bias that impairs decision-making (cf. Baron 2014; Tversky and Kahneman 2015) and the extent to which the XAI system serves to mitigate or amplify it. An early and thorough assessment of the XAI system's impact and efficacy is especially important in cases where the participants of the user test conducted in Step VI and the target users differ significantly. Thus, e.g., in the case of a consumer entertainment application, which could be tested with a diverse group of lay users during calibration, a small-sized survey might be sufficient. On the other end of the spectrum, a system that supports decision making in healthcare may require several rounds of testing (Wang et al. 2019). If an application-grounded evaluation reveals that the XAI system does not yet or no longer meet its requirements, the process can be resumed with Step I or V at the team's discretion.

Table 1. Steps and Tasks of the User-Centric XAI Process

Instantiation	I Context and user specification	• Investigate application domain and business background • Identify the purpose of the XAI system • Create an understanding of the target user group
	II Application requirements	• Derive user-centric and technical requirements • Identify modes of presentation and desired characteristics of explanations based on requirements
	III Instantiation of XAI method	• Select an XAI method as the foundation of the XAI system • Instantiate the XAI system
	IV Preliminary proxy measures	• Select a preliminary set of proxy measures
Calibration	V Calibration with proxy measures	• Find parameters of the XAI system that lead to optimal explanations according to the proxy measures
	VI Evaluation with users	• Conduct a user test to evaluate generated explanations against competing and control explanations
	VII Analysis of users' perception	• Derive desired explanation characteristics from users' perspectives • Evaluate if explanations generated by the XAI system meet the desired explanation characteristics • *If results are satisfying, go to Step IX, if results are unsuitable, go to Step I, otherwise go to Step VIII*
	VIII Refinement of proxy measures	• Validate and refine the desired characteristics of explanations • Validate and adapt the set of proxy measures • *Go to Step V*
Quality Control	IX Evaluation and monitoring	• Continuously monitor the XAI system's output • Evaluate the XAI system's efficacy under real-world conditions • *If the XAI system does not fulfill its requirements, go to Step I or V*

Table 1. Steps and Tasks of the User-Centric XAI Process

Demonstration and Evaluation

As an essential part of the Design Science research process (Hevner et al. 2004), we demonstrate and evaluate the applicability and efficacy of our artifact in a realistic setting. We provide quantitative evidence that the artifact fulfills its objective and rigorously assess the efficacy of its core concepts.

Setting

Evaluation of the applicability and effectiveness of a design artifact requires demonstration of an instantiation in a setting that closely resembles the three "realities" system, task, and users (Sonnenberg and vom Brocke 2012). Against this background, we select an AI-based smartphone app for plant species detection as our use case. With the app, lay users can take pictures of leaves, whose species is detected by an AI system and displayed as text (cf. Förster et al. 2020). The task is to add an XAI system to the app that generates accompanying textual explanations that help the user understand the AI system's reasoning. We use a simulated prototype of the app closely modeled after real-world examples, such as the smartphone app Plantix or the AI system for identifying plant diseases presented by Ramcharan et al. (2019). The AI system classifies leaves using a neural network trained on a publicly available real-world dataset of shape and texture attributes extracted from 340 images of leaf specimen from 30 different plant species (Silva et al. 2013, 2014).

Application of the User-centric XAI Process

At the beginning of the Instantiation phase, in Step I, we identified that the app targets a lay audience with an interest in nature. In turn, explanations had to be comprehensible without expert knowledge. The focus of explanations was to entertain and educate users, helping them to gradually improve their botany knowledge while casually interacting with the app. Building on these insights, in Step II, we chose contrastive explanations, as the primary purpose of the explanations was to convey information about the leaf classification (cf. Miller 2019). Motivated by both comprehensibility and constrained smartphone screen space, we opted for short textual natural-language explanations displayed alongside the picture of the classified leaf as the mode of presentation. Turning to characteristics, in addition to the comprehensibility requirement, we assessed that explanations should be faithful to the AI system. While explanations should be as general as possible to facilitate learning, they should nevertheless fully explain the specific classification result. Research in social sciences suggests shortness, generality, and coherence (Lombrozo 2012; Thagard 1989), as well as relevance (Hilton and Erb 1996; McClure 2002), as characteristics of explanations that humans generally value. As the starting point to select an XAI method, in Step III, we considered that the AI system used for plant species detection utilized a feature extraction algorithm to extract the leaf's features from a picture (Silva et al. 2013, 2014). Thus, the input to the neural network at the heart of the AI system is a vector x_{fact} with numerical, scalar features, which could serve as the input to the XAI system as well. We selected the popular, frequently used algorithm proposed by Wachter et al. (2018), which is compatible with the input data, ensures faithfulness by directly operating on the AI system, and can be computed efficiently for neural networks (Dhurandhar et al. 2019). In a nutshell, this algorithm computes contrastive explanations by framing the search for a suitable counterfactual as an optimization problem. The approach builds on minimizing an objective function with two terms: First, the squared and weighted Euclidean distance between the AI system's output $f(x)$ and the foil y_{foil}. Second, the Manhattan distance $|x_{fact} - x|$ weighted by the mean absolute deviation MAD of each feature in a representative dataset, to ensure that $x_{fact} - x$ is sparse:

$$o(x) = \lambda \left\| f(x) - y_{foil} \right\|^2 + \sum_i \frac{\left| x_{\text{fact},i} - x_i \right|}{MAD_i} \qquad (1)$$

The XAI method by Wachter et al. (2018) has several parameters that influence the properties of generated explanations: First, the parameter λ in its objective function balances the foil's faithfulness with the sparsity of the contrast. Second, the optimization can either be conducted for a specified amount of steps or stopped once the AI system's classification of the foil surpasses a confidence threshold. Third, the resulting contrast vector can be pruned of small values. To this end, setting a threshold (Wachter et al. 2018), pruning greedily to arrive at the minimal contrast that sustains the foil's classification (Mothilal et al. 2020), or pruning features in order of ascending feature importance (Förster et al. 2020; Lundberg and Lee 2017) are all

established options. We transferred the contrast vector $\Delta x = x_{foil} - x_{fact}$ to natural language text via a custom basic text generation engine (Förster et al. 2020). The resulting explanations follow the pattern "The leaf was classified as y_{fact} and not y_{foil}. In order to be classified as y_{foil}, the leaf would need to be <comparative> <adjective> … and <comparative> <adjective>." including one comparative/adjective pair for each non-zero entry of the contrast Δx. In Step IV, we defined a set of preliminary proxy measures. Initially, we did not have insight into which characteristics were valued by the users of the plant species detection app. Hence, we selected simple measures for faithfulness, comprehensibility, and generality based on literature. To measure faithfulness, we determined whether the foil was indeed classified as the foil class (Martens and Provost 2014). As a preliminary proxy measure for comprehensibility, we determined the length of an explanation by counting the number of non-zero entries in the contrast (Wachter et al. 2018). Finally, we used the distance to the closest point in the AI system's training dataset as a preliminary proxy measure for the generality of explanations (Guidotti et al. 2019a).

We conducted three iterations of the Calibration phase. In each iteration, in Step V, we undertook a grid search of the XAI system's parameter space to find a configuration that performed well regarding the proxy measures (cf. Goodfellow et al. 2016). To this end, we selected a set of values for each parameter of the XAI system, instantiated it for each possible combination of parameter values, and generated explanations for 100 facts randomly sampled from the dataset. We calculated each proxy measure's value for each of the explanations and selected the parameter combination that best fitted the desired value range and balance. Then, in Step VI, we conducted a binary choice experiment with users in the shape of an online study presented via a web interface built with the oTree framework (Chen et al. 2016). In the experiment, users interacted with the simulated prototype of the app. First, they were presented with a leaf picture (the fact) and asked to match it to one of four possible plant species. Had the user matched correctly, the second-most likely plant-species, according to the AI system, subsequently served as the foil. Were they mistaken, the plant species they had selected was used as the foil. Second, the user saw two alternative explanations, either generated by the XAI system or through a control method, e.g., an explanation written by a researcher or generated by picking the closest data point labeled as the foil class from the dataset. Users selected the explanation they preferred or indicated when they found both to be unsuitable. Third, users selected the characteristics that influenced their decision from a pre-defined list and had the opportunity to give additional free-text justification. All users completed multiple cycles, each time judging a new pair of explanations. The study setup is described in more detail in Förster et al. (2020).

In the following, we detail each of the three iterations of the Calibration phase, with a special focus on the analysis of the user test (Step VII) and subsequent refinement of the proxy measures (Step VIII).

In the first iteration, our main focus was on tuning the XAI system such that it produced both faithful and short explanations (median length 1 with mean absolute deviation from the median (MAD) 1.0, generality 98%, faithfulness 100%). We subsequently conducted a user test with a small number of users (N=38) recruited among university students. While the audience was not representative of the intended target demographic, it allowed for a cost-efficient and rapid first validation of the application requirements. Through analysis of the collected data (Step VII), we uncovered the full set of decisive characteristics in the application context. As expected, this set comprised shortness, coherence, generality, and relevance. Additionally, it included length, concreteness, and consistency, which we found by analyzing users' free-text justifications. When assessing which characteristics were valued most, we found that short explanations consisting of just a single feature were often considered inferior to longer explanations by the users, contrary to XAI literature (Martens and Provost 2014; Wachter et al. 2018). Accordingly, in Step VIII, we relaxed the goal of creating explanations that were as short as possible. However, as the evaluation had revealed additional characteristics and we, therefore, had not gathered data on the perception of explanations regarding the complete set of characteristics, we decided to undergo another iteration and keep the set of proxy measures unaltered.

For the second iteration, we calibrated the XAI system to yield longer explanations (median length 2 with MAD 0.93, generality 98%, faithfulness 100%). We conducted a user test with significantly more users (N=144) recruited on the online platform Clickworker. This population was diverse in age, educational background, and gender and closely resembled the target audience in this regard. In Step VII, we analyzed which characteristics users named most frequently when selecting an explanation. This analysis revealed concreteness (named for 34.7% of judged pairs), coherence (34.3%), and relevance (32.9%) as the decisive characteristics of explanations, which users chose significantly more frequently than the expected average

($p<0.001$, one-sided binomial test). A subsequent analysis of the relationship between perceived characteristics uncovered a co-occurrence of concreteness and length (selected together in 33.5% of cases, Wilson score 95%-confidence interval 28.9%-38.4%) as well as a co-occurrence between relevance and shortness (34.4%, 29.1%-40.2%). In Step VIII, we observed a strong correlation between the proxy measure for length and users' perception of length (Pearson correlation 0.73, $p<0.05$) and shortness (-0.87, $p<0.01$). However, it remained unclear whether the number of features itself or linguistic properties such as the length of the sentence or the presence of comparatives were the decisive factor. Regarding the proxy measures for faithfulness and generalizability, the evaluation results did not reveal a clear link with any of the desired characteristics. We concluded that in the given scenario, users were not necessarily looking for a complete or faithful explanation, but were satisfied with a concrete explanation coherent with their expectations. While we abandoned the generalizability measure, we continued to require faithfulness, since the delivery of correct explanations was an important requirement. In summary, we found that users generally perceived the explanations generated by the XAI system as lacking in concreteness, relevance, and coherence, but sometimes appreciated their shortness. Motivated by the clear link between length and perceived characteristics, we more closely analyzed the collected data in this regard. We observed that users perceived explanations of length three and four most consistently as concrete, relevant, and coherent. Accordingly, we constructed a new proxy measure CRC for these characteristics by determining whether an explanation fell in that range. At the end of the second iteration, we assessed that the quality of explanations was not yet satisfactory. Still, given the new CRC measure, we were hopeful that another iteration of the Calibration phase would yield a significant improvement.

During the third iteration, we calibrated the XAI system to the new CRC measure while maintaining faithfulness (median length 3 with MAD 0.26, faithfulness 98%). To this end, we adapted the contrast pruning to leave a minimum of three features in any explanation. The generated explanations fell into the target range in 94% of cases, whereas for the previous parametrization, only 25% did. Further, to better convey the magnitude of the contrast, we added more nuanced comparatives based on the difference between fact and foil relative to the features' standard deviation in the dataset. In Step VI, we again conducted a human-grounded evaluation (N=100) through the Clickworker platform. Users judged newly generated explanations against the explanations generated with the previous parametrization of the XAI system as well as human-made explanations as a benchmark. Our analysis in Step VII confirmed the previous finding that concreteness, relevance, and coherence were decisive characteristics for selecting an explanation. In more than two out of three cases (69.1%, $p<0.001$), the explanations generated with the new parametrization were preferred to that of the previous iteration. Users perceived them as more concrete (40.5%, $p<0.001$), more coherent (34.5%, $p<0.05$), and longer (54.3%, $p<0.001$). Overall, we found that the XAI system's explanations outperformed both that of the previous parametrizations as well as the human-made explanations. Thus, we deemed the XAI system fit for deployment.

In the case of a real-world application, in Step IX, the XAI system would be monitored and evaluated throughout its lifecycle. On a technical level, we would continuously monitor the explanations produced by the XAI via a dashboard (Microsoft 2020). If the explanations fell below specified thresholds (e.g., faithfulness below 95%), we would resume the process with Step V. To verify that the XAI system's explanations indeed entertain and educate the app's users as intended, we could occasionally present short surveys to or conduct interviews with randomly selected users of the app.

Evaluation

We evaluate our artifact, the "User-centric XAI process," with respect to its objective and its efficacy (Hevner et al. 2004). As detailed in the previous section, the process succeeded in guiding the instantiation and calibration of a user-centric XAI system that produced explanations that exhibit the identified decisive characteristics. Specifically, the explanations were perceived as concrete, relevant, and coherent by users while being faithful to the explained AI system. We identify the parametrization of the second iteration of the Calibration phase as state of the art (SotA). At this stage, the XAI system's parametrization was informed by recent literature (cf. Instantiation phase) as well re-calibrated (median length 2 with MAD 0.93, generality 98%, faithfulness 100%) based on the results of the user test conducted in the first iteration of the Calibration phase. We argue that this choice of a benchmark is justified due to the absence of previous reports on XAI systems in the application context and the fact that the predominant practice in XAI research to date is to instantiate XAI methods without the involvement of users (Abdul et al. 2018; Wachter et al. 2018). Indeed, the parametrization of the second iteration of the Calibration phase exceeds the current de-

facto standard in XAI research (Abdul et al. 2018; Kirsch 2018; Miller et al. 2017; Mittelstadt et al. 2019). Hence, the second iteration's parametrization of the XAI system represents, if anything, an upper bound on the SotA. Utilizing the data collected in Step VI of the third iteration, we compare the results for our final XAI system with that of the second iteration (SotA) and the human benchmark. More specifically, we analyze the participants' preferences for one out of two different contrastive explanations in a binary choice experiment (Doshi-Velez and Kim 2018). We additionally identify the reasons for participants preferring an explanation, which they select from a pre-defined list of characteristics (cf. Förster et al. 2020). The analysis (cf. Table 2) reveals that the XAI system calibrated with our novel process significantly outperforms the SotA directly as well as in comparison to the human benchmark. As described in detail above, this can be conclusively attributed to the perception of the XAI system's explanations as more concrete (40.5%, $p < 0.001$), more coherent (34.5%, $p < 0.005$), and longer (54.3%, $p < 0.001$).

Table 2. Comparison of State of the Art and Artifact's final XAI system.

	Users prefer explanations over that of competing approach	Users prefer explanations over human benchmark
State of the Art	30.95% ($p < 0.001$)	40.6% ($p < 0.01$)
Artifact's final XAI system	**69.05%** ($p < 0.001$)	**65.3%** ($p < 0.001$)

Results from the user test of iteration 3. p-values given are for a one-sided binomial test (H_0=50%).

Table 2. Comparison of State of the Art and Artifact's final XAI system

To assess the efficacy of the "User-centric XAI process," in the following, we examine each of its three core components (cf. Basic Idea). First, the process proved well-suited to reach the objective. More precisely, we found all phases and steps to be indispensable and placed in a sensible order. In the beginning, the Instantiation phase guided the team from understanding the application context towards the instantiation of a suitable XAI system. The translation of the application requirements into the mode of presentation and desired characteristics provided the essential foundation for the Calibration phase. Entering this phase with a functionally-grounded evaluation ensured that the first parametrization of the XAI system presented to users was already well-tested, in turn enabling the collection of reliable data. Further, the team had the opportunity to familiarize themselves with the potentials and constraints of the XAI system, which added to the first round of analyses. As is evident from the mixed results obtained in the user tests of the first two iterations of the Calibration phase, requiring multiple iterations of calibration and rigorous user testing was invaluable. It ensured that the XAI system verifiably fulfilled all application requirements before it was deemed fit for deployment. Second, incorporating the principles of user-centric design ensured that the user was at the center of attention throughout the process. On a technical level, the XAI system built on a well-known algorithm was capable of generating explanations right after its instantiation. However, as the first user test unambiguously revealed, these explanations failed both to satisfy the users and to meet the requirements. Importantly, the demonstration highlighted that, in line with the quest for fostering human agency, the process incorporates user-centricity beyond user satisfaction. On the one hand, the team emphasized the characteristics users appreciated most when calibrating the XAI system. On the other hand, however, based on the identified needs and expectations of the users, faithfulness was kept as a requirement even when the analyses of user tests revealed that in the test, it was not a decisive consideration for participants. Third, the iterative integration of functionally-grounded and human-grounded evaluation was invaluable. The functionally-grounded evaluation proved indispensable to find a suitable parametrization of the XAI system. We conservatively estimate that across the three iterations of the Calibration phase conducted for the demonstration, we generated and assessed well above 250,000 explanations. On the one hand, it would have been impossible to evaluate even a fraction of these with human users. On the other hand, without validated proxy measures, functionally-grounded evaluation would have been futile. Here, the first iteration of the Calibration phase revealed that despite the thorough assessment of the application context and extensive research in the literature, the first parametrization of the XAI system failed to generate explanations that met the objective. On the contrary, the human-grounded evaluation revealed that the common assumption that users prefer concise explanations did not hold. The second iteration uncovered the relative importance of characteristics and the set of decisive characteristics, enabling us to find an empirically validated proxy measure for concreteness, relevance, and coherence. This proxy measure enabled us to systematically find a parametrization that outperformed the current state of the art both directly and with respect to a common human benchmark (cf. Table 2).

Conclusion, Limitations, and Directions for Further Research

Opacity renders the recommendations of an AI system unintelligible to the user (Guidotti et al. 2019b), which impedes human agency, hinders societal acceptance, and thus poses a critical impediment to exploit AI's potential to benefit individuals and society (HLEG-AI 2019). Indeed, opacity fosters distrust, both reducing users' willingness to accept AI decisions (Herse et al. 2018) as well as inhibiting users' critical reflection before following an AI recommendation (Rader and Gray 2015). In light of this challenge, XAI aims to provide automatically generated explanations for AI systems (HLEG-AI 2019) that enable users to understand, contest, and alter an AI system's decisions (Doran et al. 2018). While a plethora of approaches has been demonstrated in the quest to provide explanations (Adadi and Berrada 2018), XAI research is criticized for not putting the user at the center of attention (Kirsch 2018). While first studies examine insights from the social sciences (Miller 2019) and evaluate explanations from users' perspectives (Förster et al. 2020), a process is needed that effectively guides researchers and practitioners in the design of user-centric XAI systems.

Against this background, we designed the "User-centric XAI process" to systematically guide the instantiation, calibration, and quality control of XAI systems such that they foster human agency and enable appropriate trust in AI systems. Our artifact's sequential structure is based on well-established processes from the fields of data mining and data science (cf. Martinez-Plumed et al. 2019). It incorporates the complementary scenarios for the evaluation of explainable systems proposed by Doshi-Velez and Kim (2018) and the principles of user-centric design (ISO 9241-210 2019), as well as insights from the social sciences into characteristics and presentation modes of explanations appreciated by users (cf. Miller 2019). We demonstrated the practical applicability of our artifact and rigorously evaluated its efficacy in the realistic setting of a smartphone app. Our contribution to research and practice is twofold. First, following the Design Science methodology (Hevner et al. 2004), we conceptualized and evaluated a process that effectively and systematically guides researchers and practitioners in the design of user-centric XAI systems. We contribute to the successful development and application of XAI systems by providing a structure for their design that keeps the user at the center of attention. At the same time, the iterative calibration ensures an appropriate balance between costly user testing and efficient optimization towards well-founded proxy measures. While focusing on post hoc interpretability and model-agnostic XAI methods, our process can be applied independently of the underlying AI system and application domain. Second, we demonstrate how to effectively incorporate processes from data mining and data science, principles of user-centric design, insights from the social sciences, and evaluation frameworks for XAI systems into a unified process. This unification puts research from different disciplines into the context of XAI and the quest for human agency, enabling researchers to identify the broader implications and links between previously fragmented findings.

Although our research provides a substantial step towards the design of user-centric XAI systems that foster human agency, it is subject to several limitations. First, notwithstanding the strength of our experiment, we evaluated our process only for one single use case and did not observe long-term effects. Nevertheless, the artifact is well-founded and does not rely on particular properties of the AI system. We encourage researchers and practitioners to apply and evaluate our process in different domains and especially with different target groups to investigate how the process varies with and can be adapted to suit different levels of the end-users' expertise. In addition, as our use case is built on an application focusing on AI as augmentation, we invite future research to explore the applicability of our process in the context of automated decision-making and interactive AI systems that allow users to contribute their expertise. Overall, studies that observe long-term effects as well as whether the XAI systems designed with our process indeed empower users in real-world applications are of particular interest. Second, translating user requirements into modes of presentation and characteristics and subsequently deriving proxy measures constitute major elements of our process. While a large variety of proxy measures has been reported in the literature, it remains an open question whether proxy measures that truly reflect the users' perception can be identified for all characteristics of explanations. Although our process is well-suited to uncover novel, domain-specific proxy measures, it cannot provide guarantees. Hence, we encourage research both into new proxy measures as well as into the fundamental question whether, in principle, proxy measures can be constructed for any characteristic. Third, designing an XAI system following the "User-centric XAI process" demands significant expenditure of time and costs, even though our artifact is designed mindful of resources, most importantly by aiming for an optimal balance of functionally-grounded and human-

grounded evaluation. As the involvement of users is indispensable for the design of user-centric systems (ISO 9241-210 2019), we invite further research to shed light on alternative means of including users and the associated trade-offs. Finally, given the challenge of human agency and societal acceptance of AI, providing explanations for AI systems and defining processes to tailor them to user requirements constitute an important element but cannot account for all aspects within and beyond the research field of XAI. In particular, the quest for human agency raises challenges for XAI deployments in organizations, which we invite future research to investigate. With our work, we hope to encourage XAI researchers to put the user in the focus of their attention, thereby pushing this fascinating research field forward.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. 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.

Adadi, A., and Berrada, M. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access* (6), pp. 52138–52160.

Bandura, A. 2006. "Toward a psychology of human agency," *Perspectives on Psychological Science* (1:2), pp. 164–180.

Baron, J. 2014. "Heuristics and biases," *The Oxford Handbook of Behavioral Economics and the Law,* pp. 3–27.

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., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115.

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., D'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., and Parekh, J. 2020. "Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach," *SSRN Electronic Journal*.

Buhrmester, M., Kwang, T., and Gosling, S. D. 2011. "Amazon's Mechanical Turk," *Perspectives on Psychological Science* (6:1), pp. 3–5.

Chapman, P., Clinton, J., Kerber, R., Kabaza, T., Reinartz, T., Shearer, C., and Wirth, R. 2000. "CRISP-DM 1.0: Step-by-Step Data Mining Guide," SPSS.

Chen, D. L., Schonger, M., and Wickens, C. 2016. "OTree—An Open-Source Platform for Laboratory, Online, and Field Experiments," *Journal of Behavioral and Experimental Finance* (9), pp. 88–97.

DARPA 2017. "Explainable Artificial Intelligence (XAI)." DARPA. (https://www.darpa.mil/program/explainable-artificial-intelligence, accessed September 4, 2020).

Dhurandhar, A., Pedapati, T., Balakrishnan, A., Chen, P.-Y., Shanmugam, K., and Puri, R. 2019. "Model Agnostic Contrastive Explanations for Structured Data," *ArXiv* (1906.00117).

Doran, D., Schulz, S., and Besold, T. R. 2018. "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives," in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017*, Bari.

Doshi-Velez, F., and Kim, B. 2018. "Considerations for Evaluation and Generalization in Interpretable Machine Learning," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Cham: Springer, pp. 3–17.

Farinango, C. D., Benavides, J. S., and Lopez, D. M. 2015. "OpenUP/MMU-ISO 9241-210. Process for the Human Centered Development of Software Solutions," *IEEE Latin America Transactions* (13:11), IEEE Computer Society, pp. 3668–3675.

Förster, M., Klier, M., Kluge, K., and Sigler, I. 2020. "Evaluating Explainable Artificial Intelligence – What Users Really Appreciate," in *Proceedings of the European Conference on Information Systems 2020*, Marrakesh.

Gertosio, C., and Dussauchoy, A. 2004. "Knowledge Discovery from Industrial Databases," *Journal of Intelligent Manufacturing* (15:1), pp. 29–37.

Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*, Cambridge, MA: MIT Press.

Google 2019. "People + AI Guidebook." Google. (https://pair.withgoogle.com/guidebook, accessed September 4, 2020).

Gould, J. D., and Lewis, C. 1985. "Designing for usability: key principles and what designers think," *Communications of the ACM* (28:3), pp. 300–311.

Green, D. P., and Kern, H. L. 2010. "Modeling Heterogeneous Treatment Effects in Large-Scale Experiments Using Bayesian Additive Regression Trees," in *Proc. Annu. Summer Meeting Soc. Political Methodol.*, pp. 1–40.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. 2019a. "Factual and Counterfactual Explanations for Black Box Decision Making," *IEEE Intelligent Systems* (34:6), pp. 14–23.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2019b. "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* (51:5), pp. 1–42.

Hall, M., Harborne, D., Tomsett, R., Galetic, V., Quintana-Amate, S., Nottle, A., and Preece, A. 2019. "A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems," in *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, Macau.

Herse, S., Vitale, J., Tonkin, M., Ebrahimian, D., Ojha, S., Johnston, B., Judge, W., and Williams, M. 2018. "Do You Trust Me, Blindly? Factors Influencing Trust Towards a Robot Recommender System," in *27th IEEE International Symposium on Robot and Human Interactive Communication*, Nanjing, pp. 7–14.

Hevner, March, Park, and Ram. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.

Hilton, D. J., and Erb, H.-P. 1996. "Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance," *Thinking & Reasoning* (2:4), pp. 273–308.

HLEG-AI. 2019. "Ethics Guidelines for Trustworthy Artificial Intelligence," Brussels: Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., and Baesens, B. 2011. "An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models," *Decision Support Systems* (51:1), pp. 141–154.

IBM. 2016. "Analytics Solutions Unified Method," *Analytics Services Datasheet*, Armonk, NY: IBM Corporation.

IDEO 2015. "The Field Guide to Human-Centred Design." IDEO. (https://www.designkit.org//resources/1, accessed September 4, 2020).

ISO 9241-210. 2019. "Ergonomics of Human-System Interaction — Part 210: Human-Centred Design for Interactive Systems," Geneva: International Organization for Standardization.

Kirsch, A. 2018. "Explain to Whom? Putting the User in the Center of Explainable AI," in *Proceedings of the 1st International Workshop on Comprehensibility and Explanation in AI and ML 2017*, Bari.

Kurgan, L. A., and Musilek, P. 2006. "A Survey of Knowledge Discovery and Data Mining Process Models," *The Knowledge Engineering Review* (21:1), pp. 1–24.

Lighthall, G. K., & Vazquez-Guillamet, C., 2015. "Understanding decision making in critical care," *Clinical Medicine & Research* (13:3-4), pp. 156–168.

Lipton, P. 1990. "Contrastive Explanation," *Royal Institute of Philosophy Supplement* (27), pp. 247–266.

Lipton, P. 2000. "Inference to the Best Explanation," in *A Companion to the Philosophy of Science*, W. H. Newton-Smith (ed.), Maiden, MA: Blackwell, pp. 184–193.

Lipton, Z. C. 2018. "The Mythos of Model Interpretability," *Queue* (16:3), pp. 1–27.

Lombrozo, T. 2012. "Explanation and Abductive Inference," in *The Oxford Handbook of Thinking and Reasoning*, K. J. Holyoak and R. G. Morrisson (eds.), Oxford: Oxford University Press.

Lundberg, S., and Lee, S.-I. 2017. "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, pp. 4765–4774.

Marbán, O., Segovia, J., Menasalvas, E., and Fernández-Baizán, C. 2009. "Toward Data Mining Engineering: A Software Engineering Approach," *Information Systems* (34), pp. 87–107.

Martens, D., and Provost, F. 2014. "Explaining Data-Driven Document Classifications," *MIS Quarterly* (38:1), pp. 73–99.

Martin, K., 2019. "Designing Ethical Algorithms," *MIS Quarterly Executive* (18:2), pp. 129–142.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez Orallo, J., Kull, M., Lachiche, N., Ramirez Quintana, M. J., and Flach, P. A. 2019. "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Transactions on Knowledge and Data Engineering*.

McClure, J. 2002. "Goal-Based Explanations of Actions and Outcomes," *European Review of Social Psychology* (12:1), pp. 201–235.

Microsoft. 2020. "Team Data Science Process Documentation," Microsoft. (https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/, accessed September 4, 2020).

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences," *Artificial*

Intelligence (267), pp. 1–38.
Miller, T., Howe, P., and Sonenberg, L. 2017. "Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences," in *IJCAI-17 Workshop on Explainable AI (XAI) Proceedings*, Melbourne, pp. 36–42.
Mittelstadt, B., Russell, C., and Wachter, S. 2019. "Explaining Explanations in AI," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, pp. 279–288.
Mohseni, S., and Ragan, E. D. 2018. "A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning," *ArXiv* (1801.05075).
Mothilal, R. K., Sharma, A., and Tan, C. 2020. "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, pp. 607–617.
Moyle, S., and Jorge, A. 2001. "RAMSYS - A Methodology for Supporting Rapid Remote Collaborative Data Mining Projects," in *Proceedings of the ECML/PKDD'01 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*, Freiburg, pp. 20–31.
Norman, D. A., and Draper, S. W. (eds.). 1986. "User Centered System Design: New Perspectives on Human-Computer Interaction," *User Centered System Design* (1st ed.), Boca Raton, FL: CRC Press.
Nunes, I., and Jannach, D. 2017. "A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems," *User Modeling and User-Adapted Interaction* (27), pp. 393–444.
Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. 2018. "Stakeholders in Explainable AI," in *Artificial Intelligence in Government and Public Sector Proceedings*, Arlington, VA.
Rader, E., and Gray, R. 2015. "Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, pp. 173–182.
Ramcharan, A., McCloskey, P., Baranowski, K., Mbilinyi, N., Mrisho, L., Ndalahwa, M., Legg, J., and Hughes, D. P. 2019. "A Mobile-Based Deep Learning Model for Cassava Disease Diagnosis," *Frontiers in Plant Science* (10), pp. 1-8.
Rai, A. 2020. "Explainable AI: From black box to glass box," *Journal of the Academy of Marketing Science* 48(1), pp. 137-141.
Ribera, M., and Lapedriza, A. 2019. "Can We Do Better Explanations? A Proposal of User-Centered Explainable AI," in *Joint Proceedings of the ACM IUI 2019 Workshops*, Los Angeles, CA.
Rollins, J. B. 2015. "Foundational Methodology for Data Science," *IBM Analytics Whitepaper*, Somers, NY: IBM Corporation.
Silva, P. F. B., Marcal, A. R. S., and da Silva, R. A. 2014. "Leaf Dataset," *UCI Machine Learning Repository*. (https://archive.ics.uci.edu/ml/datasets/Leaf, accessed April 30, 2020).
Silva, P. F. B., Marçal, A. R. S., and da Silva, R. M. A. 2013. "Evaluation of Features for Leaf Discrimination," in *International Conference Image Analysis and Recognition*, Póvoa do Varzim, pp. 197–204.
Sonnenberg, C., and vom Brocke, J. 2012. "Evaluations in the Science of the Artificial – Reconsidering the Build-Evaluate Pattern in Design Science Research," in *Design Science Research in Information Systems. Advances in Theory and Practice*, Las Vegas, NV, pp. 381–397.
Still, B., and Crane, K. 2017. *Fundamentals of User-Centered Design: A Practical Approach*, Boca Raton, FL: CRC Press.
Thagard, P. 1989. "Explanatory Coherence," *Behavioral and Brain Sciences* (12), pp. 435–467.
Tversky, A., and Kahneman, D. 2015. "Causal schemas in judgments under uncertainty," in *Progress in Social Psychology: Volume 1*, New York, NY: Psychology Press, pp. 49–72.
Wachter, S., Mittelstadt, B., and Russell, C. 2018. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology* (31:2), pp. 841–887.
Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. 2019. "Designing Theory-Driven User-Centric Explainable AI," in *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, pp. 1–15.
Weerts, H. J. P., van Ipenburg, W., and Pechenizkiy, M. 2019. "A Human-Grounded Evaluation of SHAP for Alert Processing," in *Proceedings of the KDD Workshop on Explainable AI*, Anchorage, AK.
Wolf, C. T. 2019. "Explainability Scenarios: Towards Scenario-Based XAI Design," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, Marina del Rey, CA, pp. 252–257.