Association for Information Systems

AIS Electronic Library (AISeL)

Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023

Data Analytics for Business and Societal Challenges

Dec 11th, 12:00 AM

The Power of Trust: Designing Trustworthy Machine Learning Systems in Healthcare

Mariska Fecho
Technische Universität Darmstadt, mariska.fecho@tu-darmstadt.de

Anne Zöll
TU Darmstadt, anne.zoell@tu-darmstadt.de

Follow this and additional works at: https://aisel.aisnet.org/icis2023

Recommended Citation

Fecho, Mariska and Zöll, Anne, "The Power of Trust: Designing Trustworthy Machine Learning Systems in Healthcare" (2023). *Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023*. 3.
https://aisel.aisnet.org/icis2023/dab_sc/dab_sc/3

This material is brought to you by the International Conference on Information Systems (ICIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in Rising like a Phoenix: Emerging from the Pandemic and Reshaping Human Endeavors with Digital Technologies ICIS 2023 by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

The Power of Trust: Designing Trustworthy Machine Learning Systems in Healthcare

Completed Research Paper

Mariska Fecho[1]
Technical University of Darmstadt
Hochschulstraße 1, 64289 Darmstadt
mariska.fecho@tu-darmstadt.de

Anne Zöll[1]
Technical University of Darmstadt
Hochschulstraße 1, 64289 Darmstadt
anne.zoell@tu-darmstadt.de

Abstract

Machine Learning (ML) systems have an enormous potential to improve medical care, but skepticism about their use persists. Their inscrutability is a major concern which can lead to negative attitudes reducing end users trust and resulting in rejection. Consequently, many ML systems in healthcare suffer from a lack of user-centricity. To overcome these challenges, we designed a user-centered, trustworthy ML system by applying design science research. The design includes meta-requirements and design principles instantiated by mockups. The design is grounded on our kernel theory, the Trustworthy Artificial Intelligence principles. In three design cycles, we refined the design through focus group discussions (N1=8), evaluation of existing applications, and an online survey (N2=40). Finally, an effectiveness test was conducted with end users (N3=80) to assess the perceived trustworthiness of our design. The results demonstrated that the end users did indeed perceive our design as more trustworthy.

Keywords: Trust, Machine Learning, Healthcare, Design Science Research, Trustworthy AI, Artificial Intelligence

Introduction

Over the recent years, and particularly during the COVID-19 pandemic, the healthcare sector has been subject to unprecedented strains and challenges, repeatedly testing its limits worldwide (Tong et al. 2022). One outstanding challenge is that physicians are overworked, and patients often have to wait months for appointments. Beyond that, an increasingly aging society demands additional medical care, underlining the need for scalable and accessible solutions that can alleviate the burden on the healthcare sector while improving patient outcomes. The growing prevalence of Information Systems (IS) in healthcare, specifically Machine Learning (ML) systems designed to aid in medical diagnoses, is anticipated to transform the provision of medical services, potentially serving as the primary point of contact for patient care (Wang and Siau 2018). ML systems are able to identify diseases like cancer and strokes from medical images or assist physicians during surgeries (Esteva et al. 2017; Taylor et al. 2016). The evolution of digitization and the proliferation of big data have caused a shift in the paradigm of decision-making from being solely reliant on human expertise and intuition to an approach that is predominantly data-driven (Berg 1997; Lebovitz et al. 2021). Especially for data-intensive and repetitive processes like image recognition in radiology or dermatology, ML systems can help to reduce physicians' workload and analysis costs (Buck et al. 2021). In addition, ML systems have shown great potential for facilitating self-examinations towards diseases for end users with various conditions without the need for physicians to be involved in the diagnosis process from

[1] Note: Both authors contributed equally to this paper.

the beginning (Takiddin et al. 2021). For example, ML systems enable end users to submit data such as skin images, health metrics, and descriptions of symptoms, which are then evaluated using ML algorithms for a health assessment (e.g., Baldauf et al. 2020). Such ML systems hold immense promise for end users, as they provide a convenient and accessible means of assessing health, improving the availability of medical care, and potentially reducing the burden on the healthcare sector.

However, ML systems supporting end users in diagnosing diseases are met with skepticism (Baldauf et al. 2020). Reasons include insufficient performance and privacy concerns. The non-use of ML systems by end users is exacerbated by algorithmic aversion, a phenomenon where individuals tend to prefer human support over ML algorithmic support, even if the latter performs better (Dietvorst et al. 2015). For instance, a study found that when physicians were unable to comprehend the reasoning behind a diagnostic algorithm's conclusion, they chose to rely on their own expertise and experience instead (Lebovitz et al. 2021). This suggests that in high-risk environments such as healthcare, end users are more likely to trust human expertise than ML systems - especially when the decision-making process is opaque. Further factors such as inscrutability, biases and discrimination, and prediction inaccuracy could also hinder end users from building trust in ML systems (Berente et al. 2021; Gillath et al. 2021). Glikson and Woolley (2020) highlight the critical role that the notion of trust plays in shaping end users' perceptions of accepting ML advice (Dietvorst et al. 2015). In this sense, trust is paramount, as it helps overcome end users' skepticism and contributes to better adoption of ML systems.

Previous research has focused on the technical implementation of ML systems (Liu et al. 2020; Takiddin et al. 2021), particular factors influencing end users' trust in ML systems (Glikson and Woolley 2020; Li and Hahn 2022; Yang and Wibowo 2022), exploring the influence of trust on the adoption of ML systems (Handrich 2021; Lohoff and Rühr 2021), and identifying characteristics determining trustworthy ML (Kaur et al. 2022; Thiebes et al. 2020). There is, however, still a lack of research on how to design user-centered ML systems that reinforce trust in these technologies (Li and Hahn 2022; Riedl 2022). Thus, recent studies have called for concrete design recommendations for trust in user-centered ML systems (e.g., Riedl 2022). In addition, research studies on ML systems in the healthcare context have mostly focused on physicians as end users, often neglecting the perspective of end users without particular medical expertise (e.g., Jussupow et al. 2022; Lebovitz et al. 2021; Pumplun et al. 2023). Therefore, in this study, we refer to ML systems that support end users without requiring domain expertise. Our study aims to investigate the research question:

What design principles should be adopted to create trustworthy ML systems in healthcare for end users?

In this study, we present a socio-technical artifact, in our case, design principles (DP), for developing trustworthy ML systems to support end users. We employed a design science research (DSR) approach consisting of five phases: Awareness of the problem, suggestion, development, evaluation, and conclusion (Kuechler and Vaishnavi 2008). These phases were iterated in three design cycles. In this vein, we derived 13 DPs for the design of trustworthy ML systems. This paper contributes to IS research by, first, responding to recent calls of research for the development of trustworthy ML systems (e.g., Riedl 2022). Second, we extend the existing trust literature by applying the trustworthy artificial intelligence (TAI) principles (Thiebes et al. 2020) to the context of ML systems and developing DPs that increase trust in these systems. Third, we present a unique approach to designing ML systems, involving end users in all three design cycles and creating a social-technical artifact that meets the needs of its intended audience.

Theoretical Background

Trust in Machine Learning Systems

Trust has been widely researched in the IS domain (Glikson and Woolley 2020; McKnight et al. 2011; Thiebes et al. 2020). It's defined as the "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al. 1995). This definition has been used previously in the interpersonal domain (McKnight et al., 2011). It emphasizes that trust presupposes the vulnerability of the trustor, and it implies that the trustor is dependent on the actions of the trustee and cannot force the trustee to fulfill his expectations. Researchers argue that the trust definition applies beyond interpersonal relationships to the technology domain (Glikson & Woolley, 2020; McKnight et al., 2011). Trust in technologies comprises three dimensions: Functionality, reliability, and helpfulness

(McKnight et al. 2011). Functionality refers to the belief that a technology can successfully perform its intended task (i.e., provide necessary features to complete a task). A technology that works well and fulfills its intended purpose is considered functional. This property can help build trust in its ability to perform. Reliability describes the belief that a technology consistently functions properly. A technology that operates as expected and performs predictably in different situations is deemed reliable and can contribute to developing trust in its performance. Helpfulness refers to the belief that a technology offers sufficient assistance to end users, meaning that help and support functions provide necessary guidance. A technology that provides benefits to end users and supports them in achieving their goals is considered helpful, which can help build trust in its overall value. The technology trust constructs are used to evaluate the trustworthiness of a technology and can help end users decide whether they are comfortable using it. In addition, research has introduced technical concepts related to autonomous systems alongside trust in technologies (Lee and See 2004; Thiebes et al. 2020). Lee and See (2004) refer to the following three trusting beliefs to conceptualize trust in autonomous systems: Performance, purpose, and process. Performance refers to the ability demonstrated by autonomous systems to achieve their intended goal. Thus, performance is closely related to functionality. Purpose describes to what extent the autonomous system is used in the developer's intent. This concept corresponds to helpfulness by reflecting that an autonomous system has a positive orientation towards end users. Process refers to how appropriate the autonomous system is for a given task and how well it can achieve the operator's goals. Consequently, process relates to the concepts of reliability (Thiebes et al. 2020).

Previous research on *ML and trust* mainly refers to the organizational context and present literature reviews (Kaur et al. 2022; Li and Hahn 2022), develop DPs to manage customer processes (Emamjome and Rosemann 2021), or frameworks to explore how ML systems impact trust (e.g., FEAS framework (Toreini et al. 2020), TAI Principles (Thiebes et al. 2020). Empirical studies explore how trust can be transferred from known technologies and providers to ML systems (Renner et al., 2021) or investigate the influence of trust on ML adoption (Handrich, 2021; Lohoff & Rühr, 2021). Trust-related studies on the individual level were conducted conceptually by developing frameworks that either distinguish user personality and trust in ML systems (Riedl, 2022) or identify factors that affect trust in ML systems (Glikson & Woolley, 2020; Yang & Wibowo, 2022). Kim et al. (2021) explored the relationship between explainable AI and user behavior mediated by trust. The study revealed that trust effectively influences the interaction between humans and ML systems. The uniqueness of the role of trust within the context of ML systems is multifaceted. First, ML algorithms embedded in IT systems lack a physical presence. This lack of embodiment poses challenges to the development of trust between humans and ML. Human trust relies on physical cues, absent in ML systems. This absence of a visible identity makes the establishment of trust more complex and nuanced (Glikson and Woolley 2020; Li 2015). Second, ML systems possess a higher level of autonomy, enabling them to perform complex actions without direct human intervention (Berente et al. 2021). However, end users might not always be aware of the actual extent of ML's technological sophistication. This variability in perceived autonomy contributes to uncertainties in trusting ML systems. End-users may not be able to accurately assess when the ML is fully capable or when it may reach its limits. (Glikson and Woolley 2020). Third, the non-deterministic nature of ML systems introduces perceived risks in human-ML relationships (Chao et al. 2016). Due to the algorithmic nature of ML systems, these risks arise from the potential for them to make unexpected or incorrect decisions. Finally, ML system's invisible nature, coupled with its potential for erroneous functions, contributes to a unique trajectory of trust, which means that trust in ML systems changes based on the feedback regarding its accuracy (Glikson and Woolley 2020). Initially, high levels of trust can be quickly eroded when users encounter errors in ML systems, and rebuilding trust takes considerable time.

Previous literature on *ML systems in healthcare* has focused mainly on technical implementation. In particular, the literature has dealt with ML performance indicators for the diagnostic processes of diseases (Tofangchi et al. 2017), automated classification of patient data such as skin lesions using a convolutional neural network, the implementation of health telematics infrastructure (Schweiger et al., 2007), the development of collaboration platforms aiming in reinforcing the clinician–biostatistician relationship (Raptis et al., 2012), or the general implementation or design of mobile healthcare applications using ML (Greve et al., 2020; Ngassam et al., 2021). For instance, Greve et al. (2020) undertook the challenge of delivering non-communicable disease care for developing countries, a task demanding specialized medical equipment and expertise. To address this issue, the study set out to develop a mobile application to support community health workers in their routine care and counseling on non-communicable diseases. In

addition, a critical task in cancer treatment strategy is to identify and establish links between key patient characteristics while streamlining redundant data and inefficiencies. This optimization enables cancer centers to deliver faster and more successful patient-centered treatment plans. Tofangchi et al. (2017) successfully used its ML system to identify a set of essential characteristics for treatment advice, such as the inflammatory response of the tissue surrounding a tumor. A few studies have also emphasized the implementation of user-centered mobile healthcare applications based on ML and investigated the end users' overall willingness-to-use (Baldauf et al., 2020). In addition, previous research explores how ML conversational agents and chatbots could be designed to interact with patients (Nguyen et al., 2021). However, most of the identified studies are related to the organizational level and clinical decision support systems (Braun et al., 2022; Pumplun et al., 2023). The ongoing research conducted by Braun et al. (2022) focuses on the development of design principles tailored to the development of ML systems specifically intended for use in clinical and healthcare settings. Thus, they were conducted in clinics where ML systems interact with or are assessed by physicians for diagnosis. (Lebovitz et al., 2021). For instance, scholars explore how radiologists utilize diagnostic ML systems in clinical practice (Jussupow et al., 2022).

	Dimensions / Authors	Emamjome and Rosemann 2021	Glikson and Woolley 2020	Handrich 2021	Kaur et al. 2023	Kim et al. 2021	Li and Hahn 2022	Renner et al. 2021	Riedl 2022	Thiebes et al. 2020	Toreini et al. 2020	Yang and Wibowo 2022	Baldauf et al. 2020	Braun et al. 2022	Esteva et al. 2017	Greve et al. 2020	Jussupow et al. 2022	Lebovitz et al. 2021	Lohoff and Rühr 2021	Ngassam et al. 2021	Nguyen et al. 2021	Pumplun et al. 2023	Raptis et al. 2012	Rudin and Ustun 2018	Schweiger et al. 2007	Tofangchi et al. 2017
	Dimension	**Trust and ML**											**Healthcare and ML**													
Dimensions	Trust	x	x	x	x	x	x	x	x	x	x	x	x					x	x					x		
	Healthcare												x	x	x	x	x	x	x	x	x	x	x	x	x	x
	User-centered												x		x						x					
	Conversational Agents			x																	x					
	Physicians													x	x	x	x	x	x	x	x	x	x	x	x	x
	Patients												x							x					x	x
	Organizational	x	x	x	x		x	x		x	x			x	x	x	x	x	x	x	x	x	x	x	x	x
	Individual					x			x			x	x													
Methodology	Framework				x				x	x	x	x													x	
	Literature review		x		x		x																		x	
	Quantitative			x		x		x					x													
	Qualitative (Interviews)																	x		x						
	DSR	x												x		x					x	x				
	Techn. implementation												x		x	x				x	x		x	x	x	x
	Experiment																		x							
	Case Study																x									

Table 1. Overview Literature Review

Prior research has shown that interpreting the output of ML systems is challenging due to inscrutability (Berente et al. 2021), often likened to ML algorithms operating as black boxes with unexplainable inner logic (Adadi and Berrada 2018; Lebovitz et al. 2021; Rudin and Ustun 2018). Therefore, scholars have investigated how explainable ML systems could be designed to address the physician's needs (Pumplun et al., 2023) (see Table 1). Their findings suggest that ML systems should provide model and global explanations in clinical decision support systems when required. All in all, we identified a research gap in the design of user-centered trustworthy ML systems in healthcare, particularly within the individualized context. While the notion of trust is a widely explored concept in research, its practical application still presents challenges that have yet to be fully addressed (Emamjome and Rosemann 2021). Thus, we found that IS research lacks an in-depth exploration of how to design ML systems to earn end users' trust (Riedl 2022). Addressing this gap is important because patient safety, improved healthcare efficiency, and stakeholder acceptance depend on user-centered DPs that ensure trust in ML systems.

Problem Awareness: Challenges in Fostering End Users' Trust in ML Systems

ML is not widely deployed in the healthcare sector for end users' (Baldauf et al. 2020), but there is a tremendous need since ML techniques assist in detecting early indicators for diseases and improve overall efficiency while lowering the cost of care (Buck et al. 2021). As ML represents a highly intricate technology, the literature inevitably engenders a host of issues, such as *inscrutability*, *bias and discrimination*, *prediction inaccuracy*, and *privacy issues* (Berente et al. 2021; Gillath et al. 2021; Rai 2020). Our efforts have centered on mitigating these issues, closely aligned with our research goals.

P1-Inscrutability*:* Inscrutability refers to the difficulty of understanding and interpreting the ML systems' output (Berente et al. 2021). It can be attributed to the probabilistic nature of ML, which makes its output variable difficult for end users to interpret and understand (i.e., how the model generates its predictions) (Adadi and Berrada 2018; Berente et al. 2021). This inscrutability has multiple facets, including opacity, transparency, explainability, and interpretability (Berente et al. 2021). Recently, a senior scholar has suggested that "inscrutability can hamper end users' trust in the system, especially in contexts where the consequences are significant, and lead to the rejection of the systems." (Rai, 2020, p. 1). In healthcare ML system development, addressing inscrutability is paramount, as its implications can directly affect users' health and personal lives. Thus, the opacity inherent in ML systems may foster a sense of distrust, prompting end users to terminate their utilization of such systems. **P2-Bias and discrimination:** The issue of biases in ML algorithms is a major factor that contributes to the erosion of trust in these systems. Biases in ML refer to the existence of systematic errors or prejudices in the data, algorithms, or decision-making processes employed by ML systems (Berente et al., 2021; Rai, 2020). Bias and discrimination in ML systems occur when the training data used to build the model contains inherent biases, leading the model to replicate and even amplify those biases in its predictions. This problem arises when the training data is not representative of the real world diversity it is intended to reflect. Biased ML systems can lead to inaccurate or unreliable predictions or recommendations, which could potentially result in harm or negative impacts on end users. For instance, ML-based image recognition systems that have been trained on biased datasets may wrongly identify or exclude certain racial or ethnic groups, resulting in discriminatory surveillance practices. Inscrutability (P1) could exacerbate the biases, making it difficult to detect and correct any issues that may impact the system's performance (Lebovitz et al., 2021). When end users perceive ML systems as biased or discriminatory, they are likely to have less trust in the outputs. This can result in increased skepticism or even complete rejection (Lebovitz et al., 2021). This is particularly pertinent in healthcare ML systems, where biases can lead to discrimination against specific demographics, potentially resulting in unequal healthcare access and compromised health outcomes. **P3-Prediction inaccuracy***:* Prediction accuracy refers to an ML system's ability to produce precise outputs or forecasts, closely tied to achieving high performance (Lebovitz et al., 2021; Thiebes et al., 2020). Prediction inaccuracy occurs when ML systems fail to make accurate predictions on new or unseen data. This problem arises from various factors such as inadequate training data, insufficient feature representation, model overfitting, or inappropriate model selection (Rai, 2020). Prediction accuracy is one reliable factor of ML systems (Baskerville et al., 2015). If ML systems are not reliable or accurate, end users may question the effectiveness or usefulness of the ML system. Prediction accuracy in healthcare is critical since inaccuracies can profoundly impact lives. For instance, erroneous medical diagnoses by healthcare ML systems can lead to harm, fatalities, and a substantial erosion of trust in such systems (Davenport & Kalakota, 2019). **P4-Privacy***:* Privacy refers to the protection of personal and sensitive information from unauthorized access, use, or disclosure (Malhotra et al., 2004). ML systems in healthcare gather, process, and store vast amounts of sensitive user data (e.g., health conditions). Failure to protect end users' privacy can cause privacy concerns (Rai, 2020). Privacy concerns and lack of end users' control could result in mistrust, discontinuing use, and stopping the usage of the ML system in healthcare.

Overview of Design Science Research Process

Designing user-centered, trustworthy ML systems in healthcare requires a holistic approach that considers the social dimensions. It involves working with potential end users to ensure that ML systems meet their needs and expectations. DSR approach is most suitable since it involves the end user's perspective and allows for literately improving the socio-technical artifact. The DSR approach is also well-suited to addressing real-world challenges, such as those faced by the burdened healthcare sector (Gregor and Hevner 2013). Following the guidelines proposed by Kuechler and Vaishnavi (2008), we apply an iterative

DSR approach, which comprises five sub-phases: Awareness of the problems, suggestion, development, evaluation, and conclusion (see Figure 1). In sum, we have conducted three design cycles to develop 13 DPs.

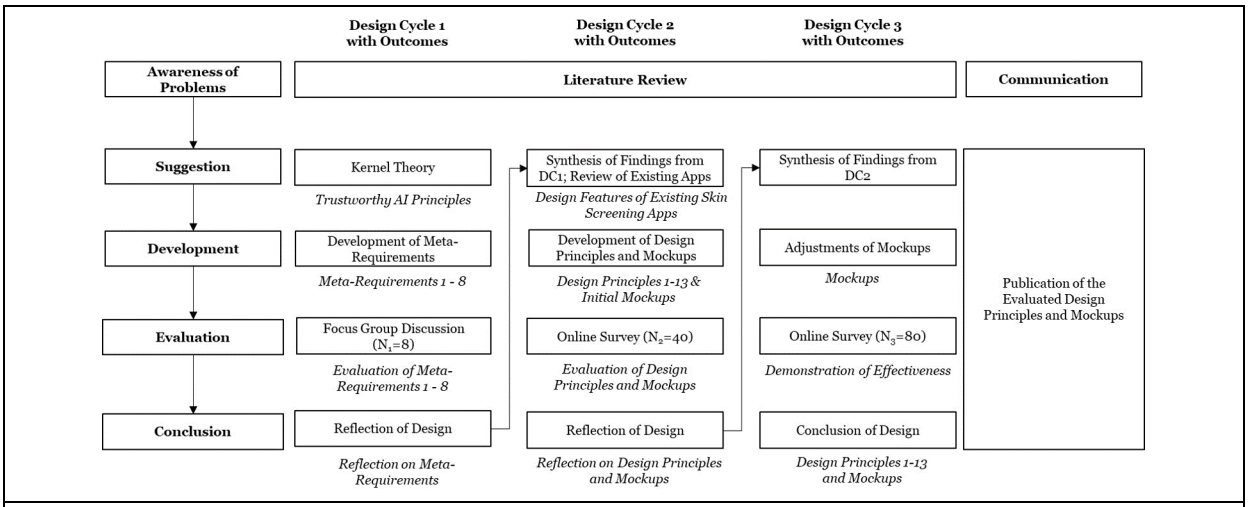

Figure 1. Design Science Research Process According to Kuechler and Vaishnavi (2008)

Based on a literature review, we have identified certain problems of ML systems that lead to lower end user trust and derive relevant meta-requirements (MRs) for designing trustworthy ML systems in healthcare. The literature review helps us to gather knowledge relevant to our problem, identify the research gap, and derive our kernel theory (Gregor and Hevner 2013). The concept of ML trustworthiness remains debated in research and practice. To address this, TAI frameworks and guidelines have emerged to advance ML technology (Thiebes et al. 2020). However, there is a significant gap in fully exploring the core TAI principles, namely beneficence, non-maleficence, autonomy, justice, and explicability (Liu et al. 2022; Thiebes et al. 2020). Applying these frameworks to end users is uncertain. Thus, our study is based on the TAI framework (Thiebes et al. 2020), which introduces TAI as an emerging goal. Thus, we decided to utilize the TAI principles as our kernel theory for several reasons: 1) It drew on a data-driven perspective, 2) it is based on the idea of building trust in automation technologies, and 3) the framework based on established trust theories (Mayer et al. 1995; McKnight et al. 2002; McKnight et al. 2011).

In the **first design cycle**, we developed the MRs for a trustworthy ML system based on the TAI principles. We conducted two video conferencing focus groups with a total of 8 potential app end users, lasting approximately 55 minutes each, to assess the MRs' suitability and relevance. The participants were 26 years old on average and were 50% male and female. As our objective was to assess MRs with end users, we selected participants who were already engaged with healthcare applications and had experience with ML systems. As there was limited existing knowledge regarding the design of such ML systems, focus group discussions were deemed appropriate at an early stage of the DSR study (Tremblay et al., 2010). These discussions allowed for direct interactions between end users, leading to the identification of specific MRs and needs. One author moderated the discussions with an initial introduction to the context of self-examination apps. We asked the participants about their expectations for the design of trustworthy ML systems and which features are important to them. Then, each MR was discussed. Following the open coding guidelines of Miles et al. (2019), two researchers separately coded the transcripts, categorized the emerging needs of the end users, and searched for supporting or contrasting arguments. All in all, the participants confirmed the MR1 - MR8 as important for the design of a trustworthy ML system. Generally, the focus group provided insights into what is important to them and how to promote trust, such as ensuring that they receive information about what the recommendation-for-action is. We took these comments into account when developing the DPs. In the **second design cycle,** we developed the socio-technical artifact, the DPs for ML systems, and an instantiation in the form of mockups. Based on the kernel theory and MRs, we formulate the DPs according to the recommendation of Gregor et al. (2020). We then identified a suitable use case in the healthcare domain. The objective was to enable the end user to independently use the ML system, enabling patients to regularly monitor their own health conditions. A second criterion for selecting the use case involved choosing an ML system that uses image recognition and employs a classification algorithm for disease identification. The third criterion was the identification of a disease

whose early detection is crucial for effective treatment. Consequently, we opted for the recognition of skin diseases. To design the mockups, we first analyzed existing ML systems for self-examined skin screening to develop a fundamental understanding of their functionality. We examined the skin screening applications and their design features listed in Table 2.

Apps / Features	Instructions	Information: Purpose	States limitations	Data privacy; Limitation of data	Deletion of data	Verification by a physician	Appointment with physicians	Transfer results to physicians	Information about ML	Error communication	Information about accuracy	Explainability of results	Recommendation for action
SkinScreener	✓	✓	✓	-	-	-	-	-	-	-	-	✓	✓
SkinVision	✓	✓	-	-	-	-	-	-	✓	-	-	✓	✓
AIDermatologist	✓	✓	-	-	-	-	-	-	✓	-	-	-	✓
Scanoma	✓	-	-	✓	-	-	-	✓	-	-	✓	-	✓

Table 2. ML Systems for Self-Examined Skin Screening

For the design of the mockups, we utilized the interface design tool Figma to create a realistic representation of a skin screening application (Figma 2023). The final step in this second cycle was the evaluation of the instantiation of the DPs in the form of mockups. We used an anonymous online survey to evaluate the DPs with N_2=40 end users. We inquired about the participant's perception of the helpfulness of the DPs and mockups (McKnight et al. 2002) as well as their assessment of the ease of use of these components (Davis 1989). One approach involved requesting the participants to evaluate the variables using a 7-point Likert scale, while the other method involved soliciting their subjective opinions through a text input field. The participants of the survey were potential end users of ML systems, on average 32 years old and 50% male and female. All of the DPs were considered essential (mean values >=4.80 on a scale from 1="not at all" to 7="extremely"), and none of the participants expressed concerns about the development of a particular DP. Furthermore, we also adjusted the mockups based on the online survey feedback for the effectiveness test. In general, the initial DPs that were derived from the MRs based on the focus group discussions could be confirmed by the survey as our final set of 13 DPs. In the **third design cycle**, we tested the effectiveness of our DPs and of the instantiation in the form of mockups in an online survey with N_3=80 end users. Thereby, we referred to measures derived from the identified justificatory knowledge measure to what extent the trust changes in the designed ML system.

Results: Deriving Meta-Requirements and Design Principles

In the following, we present the MRs based on the related literature and kernel theory (TAI principles). Then, we present the derived DPs. Thiebes et al. (2020) introduced the concept of TAI by arguing that the full potential of AI will only be realized if trust can be established in its development, deployment, and use. We deem the TAI principles appropriate in our study since ML is a subcategory of AI (Berente et al. 2021). The TAI principles are characterized by: 1) Beneficence, 2) non-maleficence, 3) autonomy, 4) justice, and 5) explicability. These five principles are related to the trust in technology and automation beliefs mentioned in the theoretical background. The imperative for trustworthy ML systems becomes undeniable when they are applied in the context of human health. Based on the related literature and the TAI principles, we derived eight MRs for trustworthy ML systems in healthcare that frame our design theory (Gregor & Hevner, 2013). In addition, by formulating the MRs into concrete design recommendations, we derived 13 DPs (see Table 3) that ensure the development of a trustworthy ML system for end users.

Beneficence refers to the development, deployment, and use of ML that is beneficial to humanity by acting in the end user's best interest, trying to help or achieve certain benefits (McKnight et al. 2002; Thiebes et al. 2020). This principle refers to the two trusting beliefs, helpfulness, and purpose (Thiebes et al. 2020). Extant research shows that these trusting beliefs are essential indicators for measuring trust in technologies (McKnight et al. 2011; Renner et al. 2021). The content provided by ML systems influences end users' perceived information quality. In particular, end user's trust in ML systems relies on the systems' ability to provide precise, up-to-date, comprehensive, and relevant information that aids the end user's objectives and supports its task (Kim et al. 2021; Yen and Chiang 2021). Consequently, ML systems in healthcare should provide adequate and responsive help to end users (**MR1**). In summary, to meet MR1, we derived

our first DP (**DP1**) (see Table 4, summarizing the final set of derived DPs), which refers to leveraging trust in the ML system in healthcare by providing guidance and appropriate help to end users. This principle may be instantiated by the provision of short explanations in the form of instructions and advice to the end user. In line with the trusting belief purpose, trust in ML systems can be increased by providing the user information about its purpose, i.e., the goals it was designed to achieve (Lee and See 2004). Therefore, it is important to clearly communicate the purpose of ML systems in healthcare (**MR2**) (Amershi et al. 2019). According to Yen and Chiang (2021), end users interacting with ML systems expect informative conversations while minimizing the occurrence of irrelevant information. Conclusively, we derived the **DP2a** and **DP2b**. These two principles aim to increase trust in the ML system by providing information about the functionalities and limitations of the system. DP2a and DP2b may be instantiated by giving examples of how to use the ML system and how not to use it. **Non-maleficence** refers to the development, deployment, and use of ML in a way that avoids bringing harm to people by particularly protecting people's privacy (Thiebes et al. 2020). It relates to the trusting beliefs reliability and process. Advances in digitization have shifted the emphasis from the intuition-based expertise of a human expert to a more data-driven approach to decision-making (Berg 1997; Lebovitz et al. 2021). A large amount of data is essential for ML systems to derive patterns and make predictions about a certain problem (Duan et al. 2019). An ML system should aim to transmit data confidentially, integrally, and authentically to reduce concerns and comply with privacy protection regulations. If the ML system has mechanisms in place to protect personal information such as identity, location, and device data and ensure only authorized end users have access, it is more likely to be trusted (Robinson 2020). Thus, ML developers should enable end users to have control over their data in ML systems (Sheridan 2019). In summary, our third MR for ML systems in healthcare is to address end users' privacy concerns by implementing suitable measures for safeguarding their personal data (**MR3**). This results in DP3a and DP3b, which aim to protect the privacy of users. According to **DP3a**, the collection of data from the end user is kept to a minimum by collecting only the data that is actually needed for the health analysis. In addition, technical measures ensure that no information is disclosed to unauthorized parties. **DP3b** refers to leveraging trust in the ML system in healthcare by giving users control over their data (e.g., allowing them to permanently delete their data). **Autonomy** advocates for the promotion of human autonomy, agency, and control, which may include limiting the autonomy of ML systems when necessary (Thiebes et al. 2020). Due to ethical and legal aspects, ML systems are currently developed to support end users rather than to replace human experts (Roshanov et al. 2013; Takiddin et al. 2021). In contexts where tasks are primarily performed by human experts, end users usually expect humans to be in the decision loop (Yang and Wibowo 2022). This expectation stems from the direct impact these tasks may have on high-risk domains. Because ML systems lack physical appearance and have a high level of autonomy without human intervention, end user trust may be diminished. The inclusion of a human expert provides a critical layer of oversight, which helps to ensure that the decisions supported by ML systems are accurate and reliable (Faraj et al. 2018). Thus, ML systems in high-risk domains that include proper oversight mechanisms, such as involving a human expert in the final decision-making process (i.e., keeping "human-in-the-loop"), are generally considered more reliable and trustworthy than those that do not. In addition, the involvement of a human expert can help to address questions that end users may have regarding. Therefore, it should be possible for a human expert to intervene in the decision-making process of ML systems in healthcare, if necessary (**MR4**). To satisfy the fourth MR, the DP4a, DP4b, and DP4c should be followed. According to these DPs, a human expert (i.e., a physician) should be involved in the health analysis. In particular, **DP4a** aims to enhance trust by incorporating a human expert to review and, if necessary, rectify the results of the ML system. Furthermore, **DP4b** enables the user to directly schedule a consultation with a human expert. **DP4c** aims to enable the ML system's results to be promptly conveyed to a human expert. **Justice** describes the utilization of ML to amend past inequities, the creation of shareable and subsequent distribution of benefits through ML, and thwarting the creation of new harms and inequities by ML. It relates to the trusting beliefs reliability and process (Thiebes et al. 2020). Biases in the data used for training ML systems can cause algorithms to have disparate impacts on the results for disadvantaged groups (Teodorescu et al. 2021). Thus, it is essential that trustworthy ML systems in healthcare avoid providing biased or discriminating information by enhancing the diversity of the data sets and including multiple groups and conditions in algorithmic development (**MR5**). Therefore, we have derived our fifth DP. **DP5** mandates the provision of information regarding the operation of the ML system's algorithm to end users. System reliability, accessibility, and timeliness of functional features are crucial factors for assessing the quality of an ML system (Bedué and Fritzsche 2022; Yang and Wibowo 2022). Thus, an ML system is perceived as trustworthy when it is easily accessible and free from errors such

as miscalculations, inaccuracies, misinterpretations, over- or underestimations (Kim and Peterson 2017). Consequently, ML systems in healthcare should be able to detect and correct system errors and inaccuracies (**MR6**). This results in the sixth DP. **DP6** aims to increase trust by automatically notifying end users of system errors. DP6 also relates to DP1. Thus, an ML system's trustworthiness is determined by its reliability (i.e., the ML systems exhibit the same and expected behavior over time) (Hoff and Bashir 2015) and accuracy. Thus, **MR7** aims to maximize the reliability of ML systems in healthcare by achieving a high level of accuracy in performing specific tasks or functions. To meet our seventh MR, we derived the DP7. According to **DP7**, system errors will be sent directly to the support service to ensure and improve functionality. **Explicability** refers to the development, deployment, and use of explainable ML by producing interpretable ML models whilst maintaining high levels of performance and accuracy (Thiebes et al. 2020). This principle relates to the trusting beliefs functionality and performance. With the increasing complexity and non-deterministic nature of ML models and the potential impact of their decision process, the necessity for transparent and explainable models has grown increasingly important. Transparency in ML algorithms and the capacity to offer clear explanations for ML-generated results are pivotal factors affecting end user trust in ML predictions (Glikson and Woolley 2020). In particular, increased transparency and explainability can positively influence end user's trust in adhering to the advice provided by the ML system (Ebrahimi and Hassanein 2019; Glikson and Woolley 2020; Strich et al. 2021) because it enables end users to reliably judge process characteristics of the ML system (Lee et al. 2019). Finally, our eighth MR refers to maximize the transparency and explainability of ML systems in healthcare (**MR8**). Thus, we have derived our DP8a and DP8b. By providing users with information that helps them understand the results of the ML system, **DP8a** aims to increase the transparency of the results and trust in the ML system. According to **DP8b**, users should receive appropriate recommendations for action.

TAI		**Description of DP**
To increase trust in ML systems in healthcare, developers need to implement measures to ensure that…		
Beneficence	MR1	…the ML system provides end users with brief explanations that can be easily understood without prior domain and technical knowledge in the form of instructions and advice on how to use the ML system correctly. (DP1)
	MR2	…the purpose of use is clearly stated, and end users are informed about how and for what the results of the ML system can be used. (DP2a)
		…the limitations of the ML system are presented to the end users. (DP2b)
Non-maleficence	MR3	…only the necessary end user data is collected in compliance with relevant data protection regulations, and such data is safeguarded by robust technical measures. (DP3a)
		…end users are given control over their data, for example, by allowing them to permanently delete their data. (DP3b)
Autonomy	MR4	…the result of the ML system is verified by a human expert and corrected if necessary. (DP4a)
		…for critical results, it is possible to arrange a prompt appointment with a human expert directly via the system. (DP4b)
		…the ML system's result can (optionally) be sent to a human expert for documentation and potential follow-up actions. (DP4c)
Justice	MR5	…the proposed system aims to provide end users with information about how the ML algorithms work and the data on which the algorithm is based. (DP5)
	MR6	… if there is a system error that causes improper functioning of the ML system, an automatic notification is sent to the end user, and the corresponding error code will be automatically transmitted to the support service. (DP6)
	MR7	…if there is a system error that causes the ML system not to work properly, provide the end user with a notification, and the corresponding error code is relayed to the support service. (DP7)
Explicability	MR8	…the result of the ML system and its interpretation are provided in an appropriate information density and quality so that they are comprehensible for end users without prior domain and technical knowledge. (DP8a)
		…end users receive appropriate and understandable recommendations for action in the case of both negative and positive results, considering factual communication of the results. (DP8b)

Table 4. Description of DP

Effectiveness of the Trustworthy Machine Learning System

We conducted a scenario-based online survey to evaluate our socio-technical artifact, comprising the DPs for ML systems, by utilizing instantiated mockups. The objective of this survey was to investigate the efficacy of our DPs in fostering trust in ML systems. We followed established guidelines in IS to design our online survey (Lowry et al. 2016).

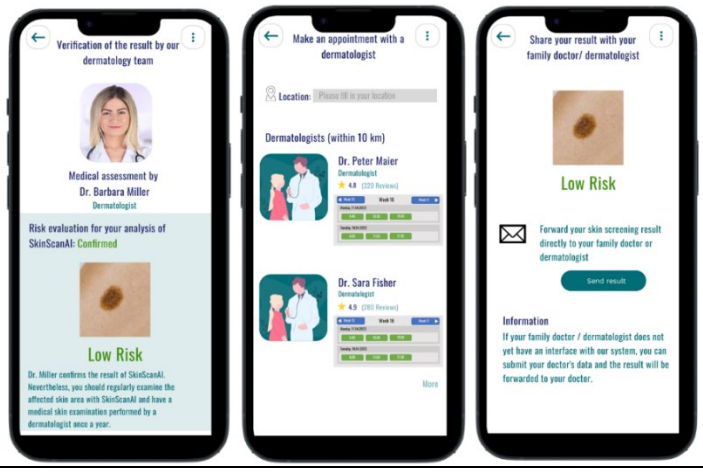

Figure 2. Examples of Mockups (DP4a-4c)

Design of Effectiveness Test: We tested the same scenario in two different groups, the control and the treatment group. We adhered to the procedures outlined by Mettler et al. (2014) for carrying out controlled experiments aimed at assessing designs. We then juxtaposed SkinScanAI, which was constructed based on DPs aligned with the TAI principles (see Table 4), with a variant of the mockups devoid of any trust elements. The stimulus (i.e., the new and improved design) is presented exclusively to the treatment group, while the control group remains unexposed. Therefore, we developed mockups for an ML system, applying our context of the skin screening process identified in the second design cycle. We present a basic skin screening process representing the control group. For the treatment group, we designed mockups based on the DPs. To prevent the priming of participants, we chose a between-subject study design in which participants were randomly assigned to one of the two groups (i.e., control or treatment group). Specifically, the goal is to scan a skin lesion with a mobile phone camera using the fictional ML system, namely SkinScanAI. In Table 5, we present the descriptions of the mockups for the control group and the treatment group. In addition, we present in Figure 2 examples of the mockups representing DP4a-DP4c.

Control Group	Treatment Group
MU1: The end user must first register to create their own profile. To do so, he/she must enter his/her name, gender, and email address.	DP3a, DP3b: Specifies that SkinScanAI collects the end user's age and gender. Optionally, the end user can also add information about their family doctor. In addition, the app provides information about its privacy mechanisms, which are compliant with current data privacy regulations. To give end users control, SkinScanAI also includes a feature to permanently delete their data.
MU2: Presents the necessary requirements for the use of SkinScanAI.	DP2a, DP2b: Presents the purpose, necessary requirements, and limitations for use.
MU3: Demonstrates the skin scan process by taking a photo of the lesion. Provides the end user with information on how to focus on the lesion.	DP1: Demonstrates the photo-taking process with clear instructions on how to take a photo of the lesion. SkinScanAI automatically provides feedback to the end user on whether the lesion is in focus, recognizable, and whether the lesion has been successfully detected.

MU4: The end user receives an exemplary result that includes an image of the lesion and the risk assessment (i.e., "Our algorithm did not detect a problematic skin lesion").	DP7, DP8a, DP8b: The end user receives a result that illustrates three risk assessment scenarios (i.e., low, medium, and high risk) and includes an image with the framed lesion. For each risk assessment, the end user is provided with an easy-to-understand explanation of how to interpret the result and the recommended course of action.
-	DP4a, DP4b, DP4c: Three mockups provide (1) additional verification of the risk assessment by physicians, (2) the option to share the result directly with the family doctor or dermatologist, and (3) the feature to schedule an appointment with a local dermatologist (i.e., online booking).
-	DP5: Provides information about how the SkinScanAI algorithm works, its average accuracy, and the database used to train the algorithm.
MU5: Provides information that if a system error occurs while using SkinScanAI, the end user can email the information to the support team.	DP6: Displays a message informing the end user that in the event of a system error, a message containing information about the error will be automatically sent to the SkinScanAI provider and that they should restart the skin screening process.

Table 5. Descriptions of Mockups

Questionnaire: The online survey includes a scenario, demographic questions, and a representative scale for the target variable, trust in technology, that was slightly adapted to fit the context. Furthermore, an attention check was implemented in the survey to detect inattentive participants. Before presenting the mockups and questions to the participants, we asked them to imagine that they would like to examine a specific skin lesion (e.g., moles) using the ML system SkinScanAI. Then, the participants were guided through each step of the skin screening process by showing them mockups and particular functions. Afterwards, the participants were asked to assess trust in SkinScanAI consisting of the three dimensions: functionality, reliability, and helpfulness. To measure trust, we refer to our kernel theory from the literature (Iivari 2020) and lean on the established scale by McKnight et al. (2011) (see Table 6).

Constructs	**Items** (7-point Likert scale, 1=strongly disagree, 7=strongly agree)
Functionality	SkinScanAI has the functionality I need; SkinScanAI has the features required for my tasks; SkinScanAI has the ability to do what I want it to do.
Reliability	SkinScanAI is a very reliable piece of software; SkinScanAI does not fail me; SkinScanAI is extremely dependable; SkinScanAI does not malfunction for me.
Helpfulness	SkinScanAI supplies my need for help through a help function; SkinScanAI provides competent guidance through a help function; SkinScanAI provides whatever help I need; SkinScanAI provides very sensible and effective advice if needed.

Table 6. Constructs and Items

Data Analysis: Overall, we collected N_3=80 participants by using the established online platform Prolific (Palan and Schitter 2018), of which 40 participants were assigned to each of the two groups. The participants were evenly distributed between males (50%) and females (50%). On average, the participants were 31.5 years old, while most of them (68.8%) were between 18-33 years old. The gathered data was analyzed by using the statistical software SPSS 27. As our data for trust, including functionality, reliability, and helpfulness, was not distributed normally (Shaphiro and Wilk 1965), we applied the two-step approach for transforming the trust variables to normal distribution. First, we calculated the fractional rank of the variables, resulting in uniformly distributed probabilities, and applied an inverse-normal transformation to form a variable of normally distributed z-scores (Templeton 2011). After applying these two-steps approach, the Shapiro-Wilk test revealed that the trust variables, including functionality, reliability, and helpfulness, are normally distributed since the significances were higher than 0.05 ($p > 0.05$). In addition, the Levene's test indicated that the variances for the constructs were statistically equal, which confirms variance homogeneity. Then, we applied a t-test and could indeed confirm that the mean values in the treatment group were significantly different at the 1% significance level ($p<0.001$) (see Table 7). Thus, end users perceive the trustworthy DPs higher compared to the control group. In addition, the effect sizes were

measured by using Cohen's d, and the results confirmed large effect sizes (Cohen 1988). In conclusion, the results of our evaluation showed that the ML system based on the derived DPs received higher trust than the one based on a basic ML skin screening process.

Trust	**Control Group**		**Treatment Group**		**Results t-test**		
Dimensions	Mean	Stddev	Mean	Stddev	t-statistics	p-value	Cohens's d
Functionality	4.613	1.458	5.775	0.824	-4.560	0.000	-1.020
Reliability	4.144	1.329	5.112	0.749	-4.322	0.000	-0.973
Helpfulness	4.525	1.435	5.775	0.711	-4.818	0.000	-1.067

Table 7. Results of T-Test

Discussion

We developed 13 DPs with the goal of ensuring a user-centered and trustworthy design for ML systems through three design cycles. Prior to the initiation of the design process, we conducted a literature review to identify problems of trusting ML systems. By using the TAI principles as the kernel theory, we derived MRs for trustworthy ML systems. The DPs were refined through focus group discussions and an online survey. The final effectiveness test confirmed that established DPs indeed increase trust in ML systems.

DP1: Delivering precise and succinct instructions is fundamental to guarantee that end users acquire a thorough understanding of the appropriate utilization of ML systems. In healthcare, the incorrect use of these systems can lead to false diagnoses or other adverse outcomes for end users that may impact individuals' lives. Hence, it is crucial to ensure that end users understand how to use these systems correctly. **DP2a and DP2b** acknowledge the limitations and purpose of ML systems, aiding end users in avoiding excessive reliance on the system's outcomes, which can lead to incorrect decisions and misguided conclusions. In this way, end users can avoid making incorrect usage decisions and drawing false conclusions. These DPs are in line with previous literature, which stated that it should be clear what the system can do (Amershi et al. 2019). **DP3a**, protecting end users' data and ensuring regulatory compliance is essential in healthcare due to the sensitive nature of patient data. Establishing trust by preserving the privacy of end users is pivotal for the widespread adoption and continuous use of ML systems. In addition, **DP3b** is about giving end users control over their data by allowing them to choose whether or not to share their personal information with the healthcare provider. More importantly, end users request a technical feature to delete their data. Thus, these DPs increase trust and are designed to mitigate problem P4. In addition, previous research has shown that the implementation of privacy-preserving mechanisms can increase end user trust in technology (Bansal et al. 2015). **DP4a, DP4b, and DP4c** involve human experts in the decision-making process due to the complexity and non-deterministic nature of ML systems, as this is important to end users due to the barriers to ML adoption, namely inscrutability (P1) and inaccurate predictions (P2). Especially in high-risk domains such as healthcare, providing an additional source of trust for ML systems is crucial, as inaccurate results could have negative consequences for end users. Therefore, ML systems in healthcare are currently developed primarily to support medical diagnosis and cannot replace human experts (Takiddin et al., 2021). Our findings align with earlier studies, which have demonstrated that collaborative work between humans and machines can yield superior outcomes (e.g., Sturm et al., 2021). Thus, these DPs could mitigate the problems P1 and P2. **DP5,** raising awareness among end users about the functioning of ML algorithms is crucial to increase transparency, as it is key to building trust in ML systems and thus mitigating problem P1 (Adadi & Berrada, 2018). End users can make more informed decisions regarding the suitability of an ML system for their needs, as well as whether to depend on its outputs when they possess a thorough understanding of the system's inner workings and the underlying data. For example, if an ML system is developed using training data from middle-aged people, it may not be suitable for analyzing health conditions of senior citizens. **DP6** informs end users of system errors and provides support services. This information helps to build trust and ensure that end users can rely on the ML system. Especially in the healthcare context, providing direct support to end users when needed is important to avoid negative attitudes about system performance (e.g., Emamjome & Rosemann, 2021). **DP7** informs end users about the interpretation of the accuracy because the trajectory of trust in ML systems has mostly focused on the way trust in ML changes based on the feedback regarding its accuracy (Glikson and Woolley 2020). In addition, it is more important than informing end users about isolated metrics (Lebovitz et al., 2021; Pumplun et al., 2023). In healthcare, providing information on how to

interpret the accuracy helps end users understand and interpret the output of the ML system, thereby mitigating P3 and increasing trust. **DP8a and DP8b** ensure that end users can understand and respond to the output of the ML system without the need for domain knowledge (i.e., medical expertise). This holds significance due to the elevated autonomy of ML systems. Incorporating these DPs can effectively mitigate uncertainties, thereby fostering enhanced end user confidence in adhering to the advice provided by ML systems (e.g., scheduling a medical appointment) (Ebrahimi & Hassanein, 2019; Strich et al., 2021). These DPs aim to mitigate P1. Ensuring that end users understand the output accordingly can help prevent incorrect conclusions that may have harmful effects on end users. Due to this issue, Pumplun et al. (2023) designed an ML clinical decision support system with additional explanation features for physicians. In summary, adherence to all these principles can ensure a user-centered development of ML systems guided by TAI principles, ultimately leading to increased end users' trust in ML systems.

Theoretical Contributions

Our theoretical contribution is threefold. **First,** we respond to the recent calls for research (e.g., Riedl (2022)) to examine and develop trustworthy ML systems. Many studies focus on the technical implementation of ML systems, for instance, developing performance measures (e.g., accuracy, robustness) (e.g., Tofangchi et al. 2017). However, factors increasing trust in ML systems go beyond these algorithmic model characteristics because trust in the ML context is of great importance (see chapter, "Trust in Machine Learning Systems"). In addition, previous research on trust in ML systems mainly employed empirical methods to describe end user behavior (Renner et al. 2021). Thus, current studies still lack a deep understanding of trust in ML systems, particularly on how to design these systems to increase trust (Li and Hahn 2022; Riedl 2022). By deriving and evaluating concrete DPs for trustworthy ML systems, we contribute to the theory of *how* end users' trust in ML systems can be achieved by design. This is particularly important to expand current trust research on trust antecedences (e.g., Glikson and Woolley 2020). The DPs provide appropriate rationales for the underlying mechanisms, helping researchers to understand the development of end users' trust in ML systems. Thus, our results deepen and expand the understanding of trust in ML systems by particularly providing guidelines on how to derive trust in ML systems. **Second**, our research expands the trust literature stream by developing DPs for ML systems. By developing these DPs, we were able to specify the overarching TAI principles, thus guiding future research. It is worth highlighting that the end user placed significant emphasis on acknowledging the decisive role of human experts during the design process due to the uniqueness of trust in ML systems. Consequently, the design incorporates a human-in-the-loop approach to ensure effective decision-making and optimize outcomes. This is an important finding for IS research to understand whether or under what conditions ML systems will fully automate or augment human work processes. **Third,** previous research has mostly focused on ML systems in an organizational context (e.g., Lebovitz et al. 2021; Pumplun et al. 2023) rather than from the individual perspective of non-specialist end users (i.e., users without medical expertise). Our research is unique due to the design of a user-centered ML system. In doing so, we involved potential end users in all three design cycles, which means that we considered the perspective and needs of end users. This is important because end users are the target group for these systems and will be interacting with the ML system. Thus, we create a social-technical artifact in the form of DPs and an instantiation in the form of mockups that meet the needs of its intended audience.

Practical Contributions

We contribute to practice by, **first**, providing a social-technical artifact in the form of DPs, which can serve as practical guidance for developers and researchers to develop trustworthy ML systems. We instantiate the DPs by mockups not only to illustrate the crucial DPs for end users but also to demonstrate how these principles can be put into practice. By doing so, we are addressing key challenges in the development of ML systems, such as the lack of transparency, the involvement of humans in the decision-making process, and the infrequent use of ML systems by end users. As a result, the potential of ML in the healthcare domain can be better leveraged. This contribution is highly relevant to society and could help to increase ML system adoption. **Second**, the increasing pressure on the healthcare sector can be relieved by involving end users in the diagnostic process and supporting them with self-examination tools as the first point of medical

contact. Even if these systems cannot replace human physicians due to legal, ethical, and validation reasons (Roshanov et al. 2013), physicians' workload can be reduced by providing end users with the availability of self-examination tools at home. By involving end users in the diagnostic process, ML systems in healthcare have the potential to reduce the number of physicians' appointments, which in turn can help to ease the burden on the healthcare sector. Moreover, ML systems can also empower end users to take control of their health and well-being, as they can monitor their symptoms and keep track of their health data in a convenient and accessible manner. **Finally**, we can assist healthcare providers by demonstrating DPs to increase trust in ML systems that can promote continuous use. End users are more likely to use ML systems continuously if they perceive them as trustworthy and effective (Glikson and Woolley 2020). Through the promotion of user-centered design, healthcare providers can improve the end user's experience and create a sense of trust in the ML systems. In addition, the continued use of ML systems can lead to the collection of more accurate and comprehensive health data, consequently resulting in improved diagnosis and treatment outcomes. Thus, the results of our study can ultimately benefit both end users and healthcare providers by improving health outcomes and promoting more efficient and trustworthy ML systems.

Limitations, Future Research, and Conclusion

Overall, our research aims to address the problem of an overburdened healthcare sector by developing an ML system that prioritizes the needs and preferences and engages the end user in the diagnosis process of diseases. We recognized that end users may not trust these systems due to issues such as inscrutability, biases and discrimination, prediction inaccuracy, and privacy concerns. To address these challenges, we focused on designing a trustworthy ML system that increases trust. We applied a DSR approach and used the TAI principles as our kernel theory to develop a socio-technical artifact consisting of DPs and instantiation mockups in three design cycles. Our final evaluation test demonstrated the effectiveness of our DPs in increasing end user trust in ML systems. Our research provides important insights into the design of trustworthy ML systems and contributes to the growing body of knowledge on the development of systems in high-risk domains such as healthcare.

Our study also has limitations. **First**, the proposed DPs have not been technically implemented in a prototype. While the mockups provide a visualization of the DPs, the implementation feasibility remains unclear. Future research could focus on developing a prototype based on the proposed DPs, the instantiation in the form of mockups, and evaluating the technical feasibility. This would involve the ability to provide transparent explanations of their decision-making processes. It would also require addressing any technical challenges that arise during the implementation process, such as issues related to data privacy, system integration, and end user experience. By conducting such research, we could gain a better understanding of the implications of implementing ML systems and other high-risk domains. Future research could concentrate on empirical research of end user acceptance of the DPs using a prototype of the ML system. Thus, it could lead to valuable insight into the differences between DPs, for instance, related to autonomy or transparency. In addition, further research could explore the impact of individual differences and integrate personality traits (e.g., Big Five) and user characteristics, as these may impact the end user acceptance of ML systems (Riedl 2022). **Second**, the DPs are based on a broader range of literature and are not limited to healthcare. They encompass general concepts for building trustworthy technologies, which can be applied to the design of ML systems in other high-risk domains. For example, the principles may be useful in other diagnostic or treatment settings in healthcare, as well as in other high-risk environments like finance. Nonetheless, it's important to replicate this research in other contexts to provide empirical evidence of the principles' transferability to other high-risk domains or to identify any specific needs for each domain. **Third**, the effectiveness of the final set of DPs was evaluated in the specific context of a skin screening process. Future research could assess these DPs in varied scenarios for broader result generalizability. While our study aimed to evaluate DPs' impact on enhancing end user trust, we must acknowledge a limitation in our approach. We opted to use McKnight et al.'s (2011) established trust measurement scale due to its wide applicability. However, this choice led to the challenge of not being able to individually evaluate the impact of each DP on trust. To address this limitation and offer a more nuanced understanding, future research endeavors should focus on conducting separate evaluations for each DP. This could entail a variety of methodologies, including online surveys to gauge user perceptions, as well as interactive sessions involving end users, such as focus group discussions or interviews.

Acknowledgments

Funded by the German Research Foundation – 251805230/GRK 2050.

References

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

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. 2019. "Guidelines for Human-AI Interaction," in *CHI*, Glasgow, Scotland.

Baldauf, M., Fröehlich, P., and Endl, R. 2020. "Trust Me, I'm a Doctor – User Perceptions of AI-Driven Apps for Mobile Health Diagnosis," in *International Conference on Mobile and Ubiquitous Multimedia,* New York, USA.

Bansal, G., Zahedi, F., and Gefen, D. 2015. "The role of privacy assurance mechanisms in building trust and the moderating role of privacy concern," *European Journal of Information Systems* (24:6), pp. 624-644.

Bedué, P., and Fritzsche, A. 2022. "Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption," *Journal of Enterprise Information Management* (35:2), pp. 530-549.

Berente, N., Gu, B., Recker, J., and Santhanam, R. 2021. "Managing artificial intelligence," *MIS quarterly* (45:3), pp. 1433-1450.

Berg, M. 1997. *Rationalizing Medical Work: Decision-support Techniques and Medical Practices*, New Bakersville: MIT Press.

Braun, M., Harnischmacher, C., Lechte, H., and Riquel, J. 2022. "Let's Get Physic(AI)L – Transforming AI-Requirements of Healthcare into Design Principles," in *European Conference on Information Systems,* Timişoara.

Buck, C., Hennrich, J., and Kauffmann, A. 2021. "Artificial Intelligence in Radiology – A Qualitative Study on Imaging Specialists' Perspectives," in *International Conference on Information Systems,* Austin.

Chao, C.-Y., Chang, T.-C., Wu, H.-C., Lin, Y.-S., and Chen, P.-C. 2016. "The interrelationship between intelligent agents' characteristics and users' intention in a search engine by making beliefs and perceived risks mediators," *Computers in Human Behavior* (64), pp. 117-125.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Davis, F. 1989. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS quarterly* (13:3), pp. 319-340.

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2015. "Algorithm aversion: people erroneously avoid algorithms after seeing them err," *Journal of experimental psychology. General* (144:1), pp. 114-126.

Duan, Y., Edwards, J. S., and Dwivedi, Y. K. 2019. "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *International Journal of Information Management* (48), pp. 63-71.

Ebrahimi, S., and Hassanein, K. 2019. "Empowering Users to Detect Data Analytics Discriminatory Recommendations," in *International Conference on Information Systems,* Munich, Munich.

Emamjome, F., and Rosemann, M. 2021. "Managing trust- A design theory and design principles," in *International Conference on Information Systems,* Austin.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. 2017. "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* (542:7639), pp. 115-118.

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28:1), pp. 62-70.

Figma. 2023. "The modern interface design tool," available at https://www.figma.com/de/, accessed on Apr 20 2023.

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., and Spaulding, R. 2021. "Attachment and trust in artificial intelligence," *Computers in Human Behavior* (115).

Glikson, E., and Woolley, A. W. 2020. "Human Trust in Artificial Intelligence: Review of Empirical Research," *Academy of Management Annals* (14:2), pp. 627-660.

Gregor, S., and Hevner, A. R. 2013. "Positioning and presenting design science research for maximum," *MIS quarterly* (37:2), pp. 337-355.

Gregor, S., Kruse, L., and Seidel, S. 2020. "Research Perspectives: The Anatomy of a Design Principle," *Journal of the association for information systems* (21:6), pp. 1622-1652.

Greve, M., Lichtenberg, S., Diederich, S., and Brendel, A. B. 2020. "Supporting Non-Communicable Disease Prevention Through A MHealth Application in Decentralized Healthcare Systems: Action Design Research in Eswatini," in *European Conference on Information Systems,* A Virtual AIS Conference.

Handrich, M. 2021. "Alexa, you freak me out – Identifying drivers of innovation resistance and adoption of Intelligent Personal Assistants," in *International Conference on Information Systems,* Austin.

Hoff, K. A., and Bashir, M. 2015. "Trust in automation: integrating empirical evidence on factors that influence trust," *Human Factors* (57:3), pp. 407-434.

Iivari, J. 2020. "Editorial: A critical look at theories in design science research," *Journal of the association for information systems* (21:3), pp. 502-519.

Jussupow, E., Spohrer, K., and Heinzl, A. 2022. "Radiologists' Usage of Diagnostic AI Systems," *Business & Information Systems Engineering* (64:3), pp. 293-309.

Kaur, D., Uslu, S., Rittichier, K. J., and Durresi, A. 2022. "Trustworthy Artificial Intelligence: A Review," *ACM Computing Surveys* (55:2), pp. 1-38.

Kim, J., Giroux, M., and Lee, J. C. 2021. "When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations," *Psychology & Marketing* (38:7), pp. 1140-1155.

Kim, Y., and Peterson, R. 2017. "A Meta-analysis of Online Trust Relationships in E-Commerce," *Journal of interactive marketing* (38:1).

Kuechler, B., and Vaishnavi, V. 2008. "On theory development in design science research: anatomy of a research project," *European Journal of Information Systems* (17:5), pp. 489-504.

Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. 2021. "Is AI Ground Truth Really 'True'? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *MIS quarterly* (45:3), pp. 1501-1525.

Lee, J. D., and See, K. A. 2004. "Trust in Automation: Designing for Appropriate Reliance," *Human Factors* (46:4), pp. 50-80.

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. 2019. "Procedural Justice in Algorithmic Fairness," *Proceedings of the ACM on Human-Computer Interaction* (3:CSCW), pp. 1-26.

Li, J. 2015. "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies* (77), pp. 23-37.

Li, Y., and Hahn, J. 2022. "Review of Research on Human Trust in Artificial Intelligence," in *International Conference on Information Systems*, Copenhagen.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., and Tang, J. 2022. "Trustworthy AI: A Computational Perspective," *ACM Transactions on Intelligent Systems and Technology* (14:1), pp. 1-59.

Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G. S., Peng, L. H., Webster, D. R., Ai, D., Huang, S. J., Dunn, R. C., and Coz, D. 2020. "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine* (26:6), pp. 900-908.

Lohoff, L., and Rühr, A. 2021. "Introducing (Machine) Learning Ability as Antecedent of Trust in Intelligent Systems," in *European Conference on Information Systems*.

Lowry, P. B., D'Arcy, J., Hammer, B., and Moody, G. D. 2016. ""Cargo Cult" science in traditional organization and information systems survey research: A case for using nontraditional methods of data collection, including Mechanical Turk and online panels," *Journal of Strategic Information Systems* (25:3), pp. 232-240.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. 1995. "An Integrative Model Of Organizational Trust," *Academy of Management Review* (20:3), pp. 709-734.

McKnight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. 2011. "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on management information systems (TMIS)* (2:2), pp. 1-25.

McKnight, D. H., Choudhury, V., and Kacmar, C. 2002. "Developing and validating trust measures for e-commerce: An integrative typology," *Information systems research* (13:3), pp. 334-359.

Mettler, T., Eurich, M., and Winter, R. 2014. "On the use of experiments in design science research: a proposition of an evaluation framework," (34:1).

Miles, M. B., Huberman, A. M., and Saldaña, J. 2019. *Qualitative data analysis*: *A methods sourcebook*, Los Angeles: SAGE.

Palan, S., and Schitter, C. 2018. "Prolific.ac—A subject pool for online experiments," *Journal of Behavioral and Experimental Finance* (17), pp. 22-27.

Pumplun, L., Peters, F., Gawlitza, J. F., and Buxmann, P. 2023. "Bringing Machine Learning Systems into Clinical Practice: A Design Science Approach to Explainable Machine Learning-Based Clinical Decision Support Systems," *Journal of the association for information systems* (24:4).

Rai, A. 2020. "Explainable AI: from black box to glass box," *Journal of the Academy of Marketing Science* (48:1), pp. 137-141.

Renner, M., Sebastian Lins, Soellner, M., Scott Thiebes, and Ali Sunyaev. 2021. "Achieving Trustworthy Artificial Intelligence: Multi-Source Trust Transfer in Artificial Intelligence-capable Technology," in *International Conference on Information Systems,* Austin.

Riedl, R. 2022. "Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions," *Electronic Markets* (32), pp. 2021-2051.

Robinson, S. C. 2020. "Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI)," *Technology in Society* (63).

Roshanov, P. S., Fernandes, N., Wilczynski, J. M., Hemens, B. J., You, J. J., Handler, S. M., Nieuwlaat, R., Souza, N. M., Beyene, J., van Spall, H. G. C., Garg, A. X., and Haynes, R. B. 2013. "Features of effective computerised clinical decision support systems: meta-regression of 162 randomised trials," *BMJ* (346).

Rudin, C., and Ustun, B. 2018. "Optimized Scoring Systems: Toward Trust in Machine Learning for Healthcare and Criminal Justice," *Interfaces* (48:5), pp. 449-466.

Shaphiro, S., and Wilk, M. B. 1965. "An analysis of variance test for normality," *Biometrika* (52:3), pp. 591-611.

Sheridan, T. B. 2019. "Individual Differences in Attributes of Trust in Automation: Measurement and Application to System Design," *Frontiers in psychology* (10), p. 1117.

Strich, F., Mayer, A.-S., and Fiedler, M. 2021. "What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity," *Journal of the association for information systems* (22:2).

Takiddin, A., Schneider, J., Yang, Y., Abd-Alrazaq, A., and Househ, M. 2021. "Artificial intelligence for skin cancer detection: scoping review," *Journal of Medical Internet Research* (23:11).

Taylor, R. H., Menciassi, A., Fichtinger, G., Fiorini, P., and Dario, P. 2016. "Medical Robotics and Computer-Integrated Surgery," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib (eds.), Springer, pp. 1657-1684.

Templeton, G. F. 2011. "A two-step approach for transforming continuous variables to normal: implications and recommendations for IS research," *Communications of the Association for Information Systems* (28:1), pp. 41-58.

Teodorescu, M. H. M., Morse, L., Awwad, Y., and Kane, G. C. 2021. "Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation," *MIS quarterly* (45:3), pp. 1483-1499.

Thiebes, S., Lins, S., and Sunyaev, A. 2020. "Trustworthy artificial intelligence," *Electronic Markets* (31), pp. 447-464.

Tofangchi, S., Hanelt, A., and Bährnsen, F. 2017. "Distributed Cognitive Expert Systems in Cancer Data Analytics: A Decision Support System for Oral and Maxillofacial Surgery," in *International Conference on Information Systems,* South Korea.

Tong, Y., Tan, C. H., Sia, C. L., Shi, Y., and Teo, H. H. 2022. "Rural-Urban Healthcare Access Inequality Challenge: Transformative Roles of Information Technology," *MIS Quarterly* (46:4), pp. 1937-1985.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., and van Moorsel, A. 2020. "The relationship between trust in AI and trustworthy machine learning technologies," in *Conference on Fairness, Accountability, and Transparency,* New York.

Wang, W., and Siau, K. 2018. "Trust in health chatbots," in *International Conference on Information Systems,* San Francisco.

Yang, R., and Wibowo, S. 2022. "User trust in artificial intelligence: A comprehensive conceptual framework," *Electronic Markets* (32), pp. 2053-2077.

Yen, C., and Chiang, M.-C. 2021. "Trust me, if you can: a study on the factors that influence consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported assessments," *Behaviour & Information Technology* (40:11), pp. 1177-1194.