Electronic Markets (2025) 35:66
https://doi.org/10.1007/s12525-025-00799-3

RESEARCH PAPER

Explainable AI for online disinformation detection: Insights from a design science research project

Isabel Bezzaoui[1] · Carolin Stein[1] · Christof Weinhardt[1] · Jonas Fegert[1]

Received: 30 August 2024 / Accepted: 16 May 2025

Abstract
The pervasive threat of online disinformation challenges the integrity of the digital public sphere and the resilience of liberal democracies. This study conceptualizes and evaluates an explainable artificial intelligence (XAI) artifact specifically designed for disinformation detection, integrating confidence scores, visual explanations, and detailed insights into potentially misleading content. Based on a systematic empirical literature review, we establish theoretically informed design principles to guide responsible XAI development. Using a mixed-method approach, including qualitative user testing and a large-scale online study ($n = 344$), we reveal nuanced findings: while explainability features did not inherently enhance trust or usability, they sometimes introduced uncertainty and reduced classification agreement. Demographic insights highlight the pivotal role of age and trust propensity, with older users facing greater challenges in comprehension and usability. Users expressed a preference for simplified and visually intuitive features. These insights underscore the critical importance of iterative, user-centered design in aligning XAI systems with diverse user needs and ethical imperatives. By offering actionable guidelines and advancing the theoretical understanding of explainability, this study contributes to the development of transparent, adaptive, and effective solutions for disinformation detection in digital ecosystems.

Keywords Disinformation · Explainable artificial intelligence · Design science research · Digital platforms · User experience

Introduction

The manipulation of information through online disinformation represents a profound threat to the integrity of the digital public sphere and the functioning of liberal democracies (Del Vicario et al., 2016). This challenge has been increasingly acknowledged in Information Systems (IS) research (Weinhardt et al., 2024), especially as the rapid proliferation of manipulated content—exacerbated by the capabilities of generative artificial intelligence (AI) (Hanley & Durumeric, 2023)—has escalated beyond electoral contexts, becoming a pervasive societal issue (Truong et al., 2024; Williams et al., 2024). With digital platforms now central to public discourse, ensuring the accuracy and trustworthiness of information is more critical than ever. Disinformation has far-reaching consequences not only for society but also for digital platform governance, business credibility, and user engagement, areas of growing interest in digital market research (Siering et al., 2021; Schlagwein & Hu, 2017). In response to this threat, advancements in AI offer promising approaches for moderating disinformation (Ansar & Goswami, 2021; Shu et al., 2020; Wei et al., 2019). However, deploying AI in such a sensitive domain presents new challenges, particularly regarding the transparency, reliability, and user acceptance of algorithmic decisions. In 2018, the European Commission enacted the General Data Protection Regulation (GDPR), which mandates a right for explanations to end-users directly impacted by an algorithmic decision (Voigt & Von dem Bussche, 2017). This legal framework highlights the importance of designing AI systems that can provide clear and understandable reasoning for their decisions, particularly in contexts where these systems operate autonomously (Mohseni et al., 2019).

Explainable AI (XAI), while not universally defined (Thiebes et al., 2021), encompasses diverse efforts to

Responsible Editor: Xiao-Liang Shen.

✉ Isabel Bezzaoui
bezzaoui@fzi.de; isabel.bezzaoui@kit.edu

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Published online: 16 July 2025

enhance the transparency and trustworthiness of AI by making its decision-making processes more understandable to users (Adadi & Berrada, 2018). The XAI research domain is expansive and interdisciplinary (Brasse et al., 2023), encompassing the fields of IS, human–computer interaction (HCI), and social sciences, involving collaboration among researchers and practitioners across diverse disciplines (Miller, 2019). The application of XAI holds particular relevance in high-stakes situations or use cases where a model output directly impacts human decision-making (Blackman & Ammanath, 2022; Confalonieri et al., 2021). In platform-mediated environments, this includes decisions affecting content visibility, user trust, and perceptions of platform fairness—key factors in maintaining engagement and commercial viability (Lehrer et al., 2018; Siering et al., 2021).

Disinformation—i.e., the intentional dissemination of false or misleading information to deceive the public (European Commission, 2018)—can greatly impact individuals and society. It has become a means of hybrid warfare attacking liberal societies from within (Shu et al., 2017) and was, therefore, rated as the most severe threat anticipated over the next two years (World Economic Forum, 2024). These dynamics can have significant political repercussions, influencing elections and spreading disinformation during crises such as the COVID-19 pandemic and conflicts in regions like the Levant (Bessi & Ferrara, 2016; Murphy, 2023; Pennycook et al., 2020). The intentional nature of disinformation requires detection systems that go beyond technical accuracy. Effective detection tools must not only identify harmful content but also provide interpretable, evidence-based explanations for their decisions to establish trust and credibility (Stitini et al., 2022). This need is particularly critical for disinformation because its contentious nature often provokes skepticism regarding interventions, raising concerns about political bias, censorship, and fairness. Unlike misinformation, where user misunderstandings can often be remedied with factual corrections (Vraga & Bode, 2020), disinformation demands systems that can justify decisions to diverse user groups, including platform users, moderators, and regulators. Therefore, XAI represents a strategic tool not only for enhancing algorithmic transparency but also for safeguarding platform governance and business sustainability in an increasingly complex information environment (Lehrer et al., 2018; Maedche et al., 2019).

The dissemination of disinformation through Online Social Networks (OSN) underscores the urgent need for automated detection systems that respond swiftly and effectively. However, in online discussions, interventions such as moderation are often perceived as controversial, raising concerns about transparency and potential censorship (Mathew et al., 2020). Introducing AI-based moderation software for disinformation detection could exacerbate these concerns, as algorithms are frequently viewed as unreliable and opaque (Gorwa et al., 2020; Suzor et al., 2019). Integrating XAI-based models could help break the black box effect by providing necessary context, allowing end-users to evaluate the veracity of news content independently and reliably. Despite growing interest in XAI, the intersection of explainability and disinformation remains underexplored (Guo et al., 2022; Rjoob et al., 2021). Current research primarily focuses on technical accuracy and detection efficacy, with limited attention to the user-centric design principles necessary for building transparency in AI-based disinformation detection systems (Wells & Bednarz, 2021). By focusing on disinformation rather than misinformation, this study emphasizes the heightened technical and social complexities of disinformation detection, where transparency, user trust, and contextual explanations are paramount. Specifically, the objective is to create a user-centric foundation for developing an XAI model applicable to digital platforms and social media channels. Guided by the principles of Design Science Research (DSR) (Hevner et al., 2004; Thuan et al., 2019), the study is driven by the following research question (RQ):

> RQ: How should an (X)AI-based tool for detecting online disinformation be designed to foster user trust, comprehension, and usability by leveraging explainability and transparency?

This research advances theoretical understanding by integrating user-centric principles into designing XAI systems for disinformation detection, focusing on how user feedback and contextual explanations can enhance trust, comprehensibility, and usability. Specifically, we extend prior work in IS and HCI by identifying design principles that balance transparency with user perception, challenging the assumption that greater transparency always improves user experience (Gunning & Aha, 2019; Haque et al., 2023). Using a DSR approach, this paper details two iterative design cycles aimed at developing an XAI-based disinformation detection tool. These cycles synthesize insights from a structured literature review, empirical user feedback, and theoretical perspectives on responsible AI design. The key contribution of this study lies in its development of actionable guidelines for creating XAI systems that are not only technically robust but also aligned with user expectations in sensitive and high-stakes domains. Our findings underscore the importance of integrating user feedback early in the design process and highlight the nuanced trade-offs between transparency and user experience in XAI design. This study offers a foundation for future studies seeking to advance the theoretical and practical understanding of XAI application in the disinformation domain.

Our general research approach using DSR is presented in the "Research approach" Sect. (1). In the "Designing a disinformation detection tool on digital discussion platforms"

Sect. (2), we detail our two design cycles, discuss the theoretical background from a structured literature review, and present empirical findings that inform guidelines for designing a responsible XAI-based disinformation detection system. Finally, the "Conclusion" Sect. (3) summarizes our findings and suggests future research directions.

Research background

In recent years, the rapid advancement and integration of AI into critical applications have raised significant concerns regarding transparency, trust, and usability. XAI has emerged as a promising response, aiming to make AI systems more understandable to human users by providing insights into their decision-making process. At its core, XAI seeks to open the "black box" of AI models, offering meaningful, interpretable, and actionable explanations for various stakeholders (Angelov et al., 2021). However, despite its potential, much remains to be explored in effectively operationalizing XAI features and addressing the challenges of balancing transparency with user-centric design (Adadi & Berrada, 2018; Minh et al., 2022). These challenges are particularly salient in digital platform contexts, where AI-powered decision-making intersects with economic, regulatory, and ethical considerations (Alt, 2021; Herm et al., 2022). In such environments, trust-building is not only a technical concern but also a business imperative.

Explanations delivered via XAI systems are operationalized through explainability features, which supply reasoning for a model's decisions. These features can be classified based on their method of generation and their scope of explanation. A key distinction is made between model-agnostic and model-specific approaches. Model-agnostic methods, such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations), are versatile tools capable of explaining the behavior of any black-box model by emphasizing feature importance in classifications and predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). In contrast, model-specific methods, like Grad-CAM (Selvaraju et al., 2017), tailor explanations to the unique characteristics of particular algorithms. Moreover, explainability features differ in scope: local explanations focus on individual outputs, while global explanations elucidate the model's overall behavior (Confalonieri et al., 2021; Linardatos et al., 2020). Regarding transparency, both types of features play complementary roles, with local explanations often addressing immediate user concerns and global explanations enhancing broader trust and understanding. Recent work has proposed frameworks that combine technical explanation methods with business model implications, identifying XAI archetypes applicable to online platforms (Gerlach et al., 2022).

The increasing prevalence of disinformation has underscored the need for transparent AI systems, particularly in the context of detection and intervention. Research has shown that tailored explanations can significantly enhance trust and perceived reliability (Schmitt et al., 2024). However, challenges persist, as overly detailed explanations can lead to cognitive overload (Linder et al., 2021) and over-reliance on incorrect system outputs (Gorwa et al., 2020; Mohseni et al., 2021b). Furthermore, the effectiveness of XAI in improving users' mental models and decision-making has yet to be fully explored. Studies like those of Nguyen et al. (2018) and Mohseni et al. (2021b) demonstrate that XAI can enhance users' ability to assess AI predictions. However, the practical implications for real-world systems remain unclear.

A central challenge in designing XAI lies in balancing transparency with usability. Transparency—revealing a system's inner workings—is a prerequisite for understandability but does not guarantee user comprehension (Haque et al., 2023). Effective explanations must account for the target audience's cognitive abilities, expertise, and expectations (Adadi & Berrada, 2018; Gilpin et al., 2018). Research suggests that a user-centric approach, emphasizing interpretability over mere transparency, is particularly critical for non-expert users (Cirqueira et al., 2020). In disinformation detection, this challenge is amplified by the inherent complexity of the task and the ethical considerations surrounding content moderation. Researchers have proposed various explanation modalities, such as attention-based visualizations and natural language explanations, to address concerns about fairness and censorship (Guo et al., 2022). These concerns are especially relevant for digital platforms that rely on algorithmic content curation, where platform legitimacy and business model sustainability depend heavily on users' trust in moderation systems (Wanner et al., 2022). While these efforts align with regulatory frameworks like the European Union's AI Act, the real-world impact on user understanding and trust has yet to be comprehensively evaluated.

Moreover, most existing studies on XAI focus on technical metrics such as fidelity, feature importance accuracy, or computational efficiency (Wells & Bednarz, 2021). These metrics, however, do not adequately address how users perceive explanations in real-world contexts. There is a clear gap in the literature regarding comprehensive evaluation frameworks incorporating user-centered metrics such as comprehensibility, trust, and usability. Additionally, few studies consider the influence of demographic or social background on how explanations are understood and trusted. As highlighted by Binder et al. (2022), integrating linguistic rules or domain-specific context can enhance explainability in real-world systems like online review platforms, offering a parallel to disinformation detection tools.

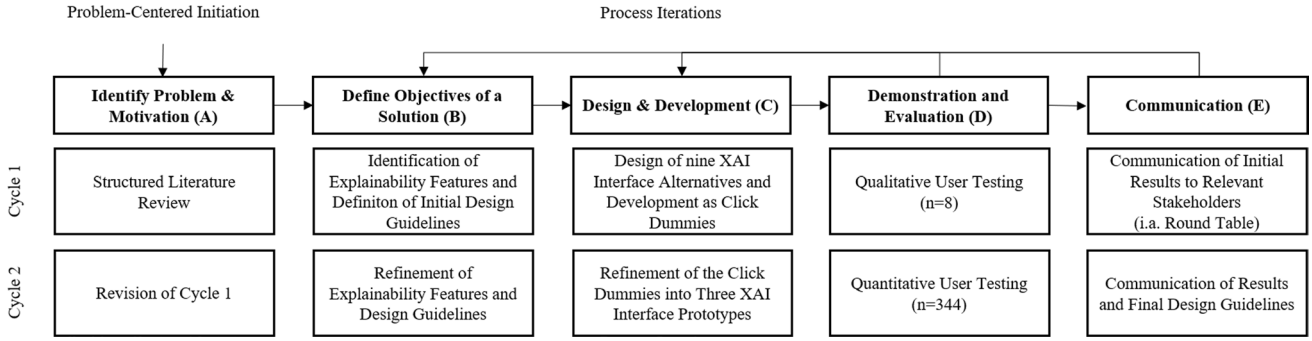

Fig. 1 Overview of the DSR approach

To address these research gaps, this study designs and evaluates an XAI artifact tailored to disinformation detection, guided by theoretically grounded design principles and rigorous user feedback. By combining qualitative and quantitative evaluations, including a large-scale online study, we aim to contribute new insights into how explainability features can be more effectively communicated and evaluated from a user-centered perspective. Our work builds on existing XAI frameworks but emphasizes the importance of integrating user feedback into the design and evaluation process to ensure that AI systems are transparent but also comprehensible and trustworthy.

Research approach

As a problem-solving paradigm, DSR focuses on the creation of artifacts to provide both descriptive and prescriptive knowledge and innovative solutions (March & Smith, 1995; vom Brocke et al., 2020). In the HCI community, DSR is an established method to support the iterative development of technical artifacts focusing on effective human use (Adam et al., 2021; Herm et al., 2022). With their six-step research procedure, Peffers et al. (2007) introduce a structured approach to problem-centered DSR projects. To thoroughly answer our research question, we conduct two DSR cycles following their established procedure of problem identification, definition of objectives, design and development, demonstration and evaluation, and communication (Peffers et al., 2007). While our first DSR cycle focuses on the artifact's relevance (Hevner, 2007), rigorously evaluating the problem space by conducting a structured literature review and an in-depth qualitative analysis of user feedback on initial design guidelines (Gurzick & Lutters, 2009), the second cycle strengthens the evaluative rigor (Hevner, 2007) by quantitatively evaluating refined design guidelines and associated hypotheses in an online experiment (Peffers et al., 2012) with fully functioning XAI prototypes (see Fig. 1).

Conduction of the first DSR cycle

First, to evaluate the problem space thoroughly and motivate potential solutions (Peffers et al., 2007), we conducted a structured literature review following Webster and Watson (2002) (A). Implementing the PRISMA workflow (Page et al., 2021), we structurally identified and screened literature dealing with applying XAI in front-end design, resulting in the analysis of 57 literature endeavors. The literature review's results informed the second and third research activities of our first cycle: To define preliminary objectives for a solution (Peffers et al., 2007), we derive initial design guidelines for developing a disinformation detection tool on digital discussion platforms (Gurzick & Lutters, 2009), emphasizing the critical role of end-user perspectives in the successful design and adoption of such systems (B). The design and development of DSR artifacts comprises the derivation of functionality and architecture based on solution objectives and the artifact's creation (Peffers et al., 2007). Thus, we implement the guidelines (Lukyanenko et al., 2017) in nine mockups for an XAI disinformation detection tool (C). Finally, to demonstrate the artifact's usability and evaluate the extent to which the solution objectives are met (Peffers et al., 2007), we cover the fourth and fifth DSR activities simultaneously in the conduction of an on-site qualitative user study in the form of a focus group (Tremblay et al., 2010) with $n = 8$ users (D). We conclude the first DSR cycle by communicating the initial findings to practicing professionals (Peffers et al., 2007), among other things, through a practitioners' round table (E).

Conduction of the second DSR cycle

Following the iterative nature of DSR research (Hevner, 2007; Peffers et al., 2007), we revise our initial DSR cycle and its insights from the qualitative study and the practitioners' feedback (A) to refine our solution objectives (B). Building on our revised design guidelines (Prat et al., 2015), we further develop the XAI interface click-dummies into three

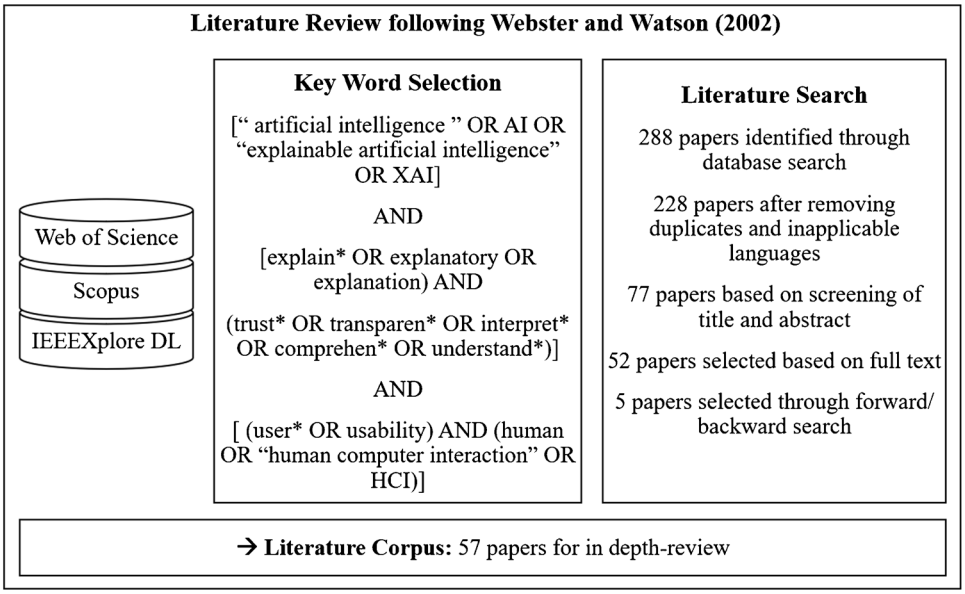

Fig. 2 Workflow guiding through the review process

fully functioning XAI prototypes (C). We then set out to quantitatively demonstrate and evaluate our solution artifact (Peffers et al., 2012; Venable et al., 2016) by designing and conducting an online experiment with $n = 344$ participants (D). Using a between-subject experimental design (Sonnenberg & vom Brocke, 2012), the online study compared the artifacts' suitability to improve comprehensibility, usability, and trust compared to a baseline AI system with no explanations. Finally, the study's findings inform the development of integrated design guidelines for XAI-based systems in disinformation detection (E).

Designing a disinformation detection tool on digital discussion platforms

First DSR cycle

Problem awareness (A)

In this work, we set out to design an XAI-based system to foster user trust, comprehension, and usability in online disinformation detection. Research has shown that XAI offers promising opportunities to provide interpretable insights into AI decision-making processes. However, evaluations predominantly emphasize technical metrics, such as fidelity and computational efficiency, while overlooking how human users perceive and use explanations in the frontend. This gap is especially pressing in disinformation detection, where explanations must balance transparency with usability while navigating ethical concerns like bias and fairness. Moreover, current evaluation frameworks inadequately address how frontend designs influence user understanding, trust, and satisfaction. To address these critical gaps, there is a need to systematically investigate how frontend designs of explainability features can be optimized to support responsible and user-centric AI systems. This study responds to this need by focusing on designing and evaluating explainability interfaces tailored to disinformation detection.

To gain a structured overview of the current state of frontend design in XAI research and its application for disinformation detection, we thus conducted a structured literature review based on Webster and Watson (2002). Figure 2 represents the workflow implemented in this paper, resulting in 57 papers included in the final review. An overview of the results will be given below.

The scientific domain of XAI, particularly the front end, is highly recent, with over 70% of relevant articles published between January 2021 and August 2023. Healthcare (15.8%) and deception detection (14.0%) are the most prominent domains. However, only three studies in the latter domain focus on the detection of online disinformation. Image classification tasks receive special emphasis, while textual data classification is sparse. Visual explainability features are the most common (50.9%), followed by multimodal (29.8%) and textual (15.8%) features. Local explanations (52.6%) are more prevalent than global explanations (7.0%), with 40.4% of sources combining both. In line with this paper's focus, 80.7% of the literature targets inexperienced end-users.

Subsequently, the literature corpus was analyzed and organized into systematic clusters based on the sources' main foci. Table 1 summarizes the investigated literature and highlights key findings. Fourteen explainability features are presented as representatives of their variations and individual modifications, along with a brief description.

Table 1 Summary of the literature review's key findings

Explainability Feature	Description	Cue type	Literature
Heatmap (saliency)	Regions of images, functions, or text are highlighted graded after their importance	Visual	Selvaraju et al. (2017); Hudon et al. (2021); Rieger and Hansen (2020); Kim et al. (2020); Lewis et al. (2021); Kumar et al. (2021)
Display and exploration of similar instances	Depending on the type of data, similar instances offer additional insights into the feature space	Any	Rjoob et al. (2021); Hwang et al. (2022)
Superpixels (saliency)	Meaningful segments of images or functions are emphasized	Visual	Guillemé et al. (2019); Trinh et al. (2021); Heimerl et al. (2020); Baur et al. (2020); Apicella et al. (2021)
Simple plots (feature importance, partial dependence, other)	Visualizing relationships between input and classification	Visual	Banerjee et al. (2023); Ekanayake et al. (2023); Nguyen et al. (2018); Dey et al. (2021)
Counterfactuals	Visualizing the extent of alteration required in input features to change the model's output	Visual	Confalonieri et al. (2021), Le et al. (2023); Cheng et al. (2020); Vermeire et al. (2022); Singla et al. (2023)
Confidence score	Certainty of a specific classification	Textual or visual	Le et al. (2023)
Concepts (normative, comparative, other)	Cluster of pixels that conveys an idea	Visual	Cai et al. (2019); Huang et al. (2022)
Generative representations (various)	Individually visualized relationships or impact of input features	Visual	Kumar and Sharma (2021); Alves et al. (2020); Kubat and Kubat (2017); Linse et al. (2022); Schreiber and Bock (2019)
Natural language explanations (template, predefined, collaborative)	Human-like textual or auditory explanation of the model's workings	Textual	Das et al. (2023); Mencar and Alonso (2019); Wang et al. (2019); Dong et al. (2021); Zhang et al. (2022)
Conversational agent	Multi-way interactive natural language explanations	Textual or auditory	Malandri et al. (2023); Hepenstal et al. (2021); Khurana et al. (2021)
Auditory	Spoken natural language explanations	Auditory	Schuller et al. (2021)
Haptic (feature importance, rules)	Physicalization of other explainability features	Haptic	Colley et al. (2022)
Multimodal (combination through an interface)	Individual combination of multiple explainability features	Any combination	Chromik (2021); Schultze et al. (2023); Park et al. (2022); Finzel et al. (2021); Weitz et al. (2021); Zytek et al. (2021); Kadir et al. (2023); Cirqueira et al. (2020); Hoque and Mueller (2021); Tamagnini et al. (2017); Salako et al. (2021); Zhu et al. (2022); Kerzel et al. (2022); Mohseni et al., (2021a, 2021b)
Keyword contribution	Degree of contribution of a single keyword on the classification	Textual or visual	Mohseni et al. (2019); Linder et al. (2021)

Among other things, the reviewed literature examines various XAI models designed for detecting forms of deception. A notable commonality among these approaches is the absence of visual data as input features, except for one approach tailored explicitly to identifying deepfake videos (Trinh et al., 2021). The emphasis on the textual dimension is apparent, leading to the recommendation to prioritize this input type when developing an XAI approach for disinformation detection. Consequently, visual data is not considered for heatmap overlays, which are exclusively applied to highlight the contribution of keywords in the textual input. The systematic, iterative software development approach, as proposed by Basil and Turner (1975), advocates beginning with a relatively simple application and gradually introducing new features and enhancements iteratively. This iterative process ensures the delivery of high-quality solutions. The initial focus is on the textual dimension, with the potential implementation of extensions or enhancements in subsequent iterative cycles. Moreover, Mohseni et al. (2021a) highlight the importance of carefully balancing explanations in terms of simplicity and information content. Overly dense explanations may lead to rejection by end-users, potentially harming a trustworthy human–machine relationship. This observation further supports the advocated systematic software development process.

The representation of confidence in a prediction is deemed simple and valuable for building trust between humans and AI (Le et al., 2023). However, the relevance of the confidence score may be significant only when it surpasses a specific threshold, especially for inexperienced end-users. Therefore, low scores indicating low confidence in a classification may be streamlined. Shu et al. (2017) propose incorporating diverse metadata input features into a disinformation classification model to improve performance. While the expected benefit of input metadata on performance is acknowledged, it remains uncertain whether end-users perceive an explainability feature relying on metadata as helpful and contributive. Thus, in line with Basil and Turner (1975), it is suggested that metadata explainability features be excluded in the initial approach. Natural language explanations fully expand only on demand and summarize the most influential features in a classification that aligns with the criteria outlined by Mohseni et al. (2021a) for simple yet effective explanations. While often expected to emulate human behavior, conversational agents may face challenges when primarily dedicated to specific applications due to their limited functionalities and knowledge (Brendel et al., 2020; Hepenstal et al., 2021). Consequently, the potential for user frustration arises, which may be detrimental to trust in human–machine interaction. In the context of disinformation, Mohseni et al. (2019) distinguish two kinds of interpretability: algorithmic interpretability and human interpretability. Algorithmic interpretability assists machine learning experts in visualizing model parameters, inspecting behavior, and improving performance. Human interpretability aims to provide transparency for inexperienced end-users by offering comprehensible explainability features to elucidate how a model works and how decisions are made. This form of interpretability is crucial for fostering trust in the human–machine relationship, aligning with the objectives of this work.

In summary, the literature underscores the importance of simplicity and clarity in explainability features to build trust among inexperienced end-users. However, this focus reveals a gap in user-centered research on how these explanations are best delivered and experienced on the front end of XAI applications. Addressing this gap is crucial for developing XAI systems that are not only transparent but also user-friendly across diverse application domains, including disinformation detection.

Solution objectives (B)

Building upon the literature review, the findings can be distilled into solution objectives in the form of design guidelines (Gurzick & Lutters, 2009) generalized for constructing an XAI model to detect disinformation on digital platforms:

1. **Preserve the original GUI.** Maintain the existing platform's GUI to ensure a seamless transition for users and uphold their established interaction habits. This helps avoid disruption and maintains usability and comfort (Garaialde et al., 2020).
2. **Balance simplicity and clarity.** Strive to balance simplicity with an effective explanation of the model's decisions. Use iterative evaluations to refine explanations, ensuring they are clear and comprehensible without becoming overly complex (Mohseni et al., 2021a).
3. **Empower inexperienced users.** Design features to be accessible to inexperienced users, ensuring they retain decision-making authority and can effectively navigate and understand content. This supports user empowerment and fosters trust (Mohseni et al., 2019).
4. **Supplement confidence scores.** Use confidence scores as a supplementary feature to indicate prediction certainty. Simplify the presentation to avoid overwhelming users while providing essential information (Le et al., 2023).
5. **Implement colored saliency for critical insights.** Highlight significant keywords in the text. Use clear color schemes and balance complexity to maintain clarity (Selvaraju et al., 2017; Chromik, 2021).
6. **Provide expendable natural language explanations.** Design natural language explanations to be concise and initially hidden, expanding upon user interaction. This approach keeps the interface clean while allowing users

Fig. 3 Design alternatives for different initial flaggings of classified posts

Fig. 4 Design alternatives for confidence score displays

to access detailed information as needed (Das et al., 2023).

7. **Exclude conversational agents.** Avoid integrating conversational agents in the initial model to prevent potential user frustration. Focus on delivering clear and direct explanations through other features (Brendel et al., 2020; Hepenstal et al., 2021).
8. **Evaluate explainability features.** Conduct practical evaluations of explainability features to assess their effectiveness and impact in real-world scenarios. This evaluation is essential for understanding how well the features meet user needs and improve the overall user experience (Mohseni et al., 2021a).
9. **Iterative development and improvement.** Follow an iterative development approach to enhance the application continuously. Incorporate user feedback and adapt to evolving needs and technological advancements to ensure ongoing improvement and high quality (Basil & Turner, 1975).

Click-dummies of an XAI interface (C)

In the subsequent phase of our design process, we systematically implemented the solution objectives outlined for developing our XAI-based disinformation detection tool. This process began with preserving the platform's original GUI to ensure a smooth integration of new features (Guideline 1). For embedding the system's initial warning, we designed three alternatives (see Fig. 3): an overlay hiding the classified post, a banner above the post, and a banner below the post—all expandable upon desire (Das et al., 2023).

We balanced simplicity with clarity, ensuring that each explainability feature was effective and easy to understand for users with varying levels of expertise (Guidelines 2 & 3). The development focused on integrating confidence scores and text highlighting to enhance the transparency of the model's predictions (Guidelines 4 & 5). Our designs for the display of confidence scores (Fig. 4) either showed a display in percentages (Schmidt et al., 2020) or, to provide an even more simplified concept that may cater to especially inexperienced users, a gradation of "low," "medium," and "high" (Mohseni et al., 2019, 2021a).

In order to emphasize text parts that were relevant for the system's prediction (see Fig. 5), we prepared a design displaying highlighted parts directly in the classified post (Selvaraju et al., 2017; Chromik 2021). As an alternative, another design suggests citations of relevant passages in the explanatory text to keep the initial post clean and simple.

To address user needs for understandable explanations, we designed expandable natural language explanations (Guideline 6). To ensure the provision of critical information while striving to avoid information overload, a longer, more detailed explanation and a shorter explanation were developed (see Fig. 6).

In alignment with the literature review's findings, conversational agents were excluded from the initial design to prevent potential frustration (Guideline 7). These considerations culminated in nine distinct design suggestions, which were visualized in mockups to illustrate the proposed features and their integration into the XAI systems. The mockups serve as a foundation for further refinement and practical evaluation in qualitative user testing (Guideline 8), guiding the ongoing development of a robust and user-centric disinformation detection tool (Guideline 9).

Qualitative user testing (D)

To demonstrate and evaluate the effectiveness of an XAI tool detecting online disinformation, it is essential to understand the perspectives of end-users, which are crucial for

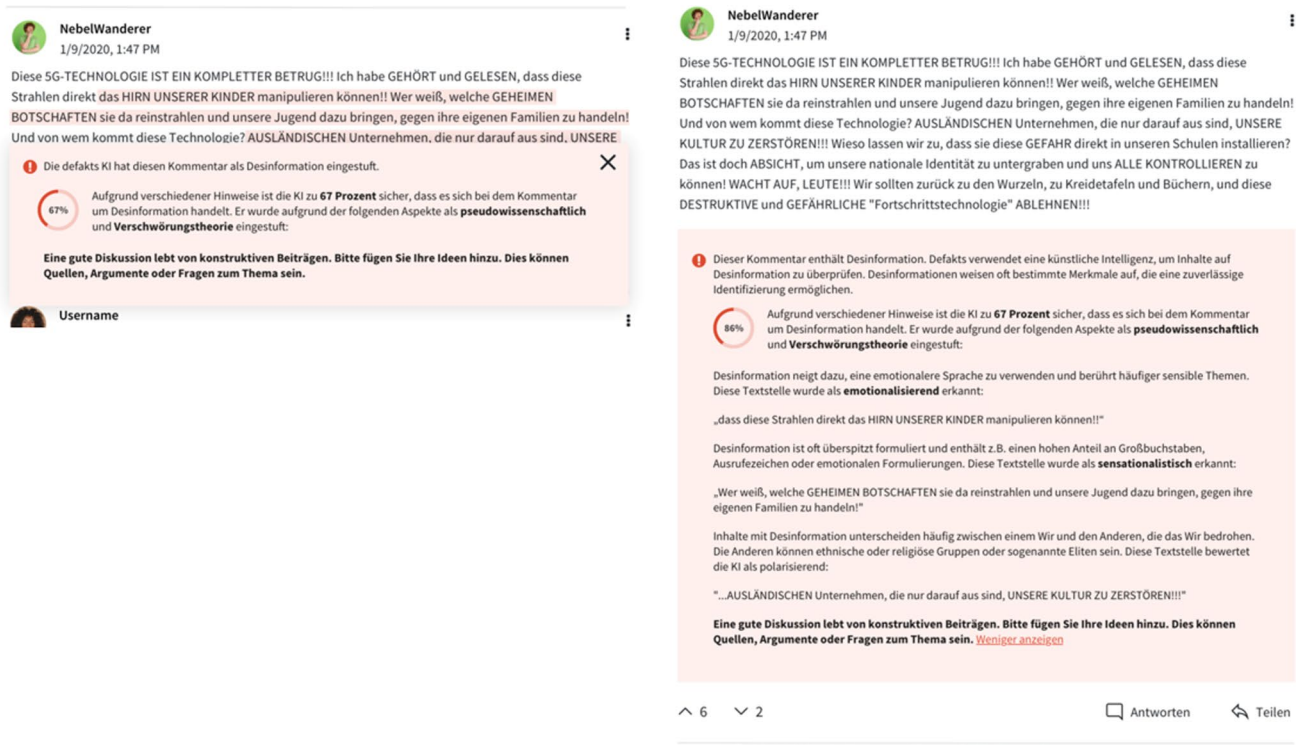

Fig. 5 Design alternatives for highlighting parts relevant for the system's classification

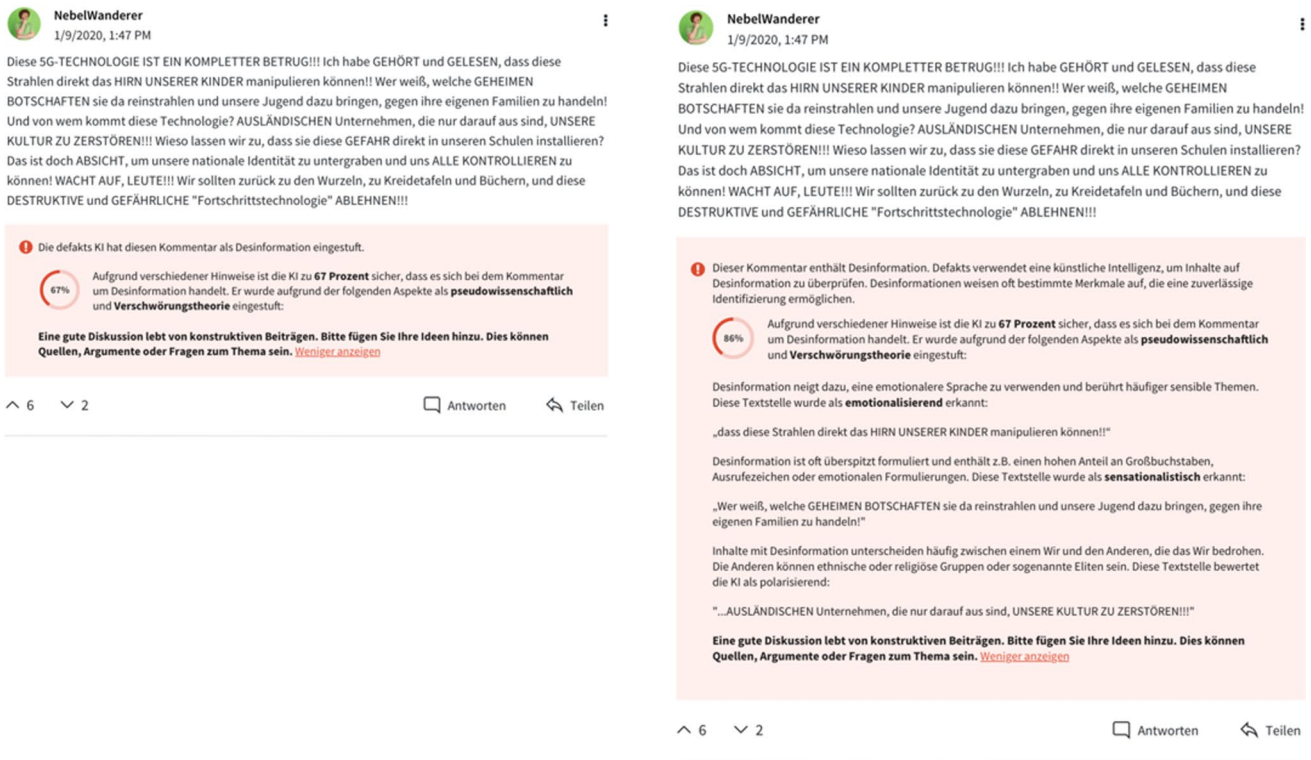

Fig. 6 Design alternatives for different explanation lengths

the successful application of such tools. By focusing on the target group's perspectives in a qualitative focus group (Tremblay et al., 2010), we seek to ensure that the design of these systems aligns with their preferences and enhances their trust and understanding. Such alignment is pivotal for the responsible development and effective integration of AI-based disinformation detection tools. In the following section, we will first elaborate on the design and conduct of the study before presenting its results in detail.

Procedure

We conducted qualitative user testing to evaluate design preferences for our developed XAI mockups. The goal was to gain an in-depth understanding of how diverse users perceive and interact with the system's output. The study involved eight participants, equally divided by gender and aged 24 to 64, recruited via the recruiting platform Testing-Time to ensure diversity in demographics and professional backgrounds. Two on-site sessions were held in February 2024, each lasting two hours with four participants. Led by two researchers and two practitioners, these sessions assessed responses to our nine different design options for the AI system's output display. The sessions followed a structured format:

1. **Introduction and briefing:** Participants were briefed on the study's purpose and the confidentiality of their participation.
2. **Design presentation:** Nine designs were sequentially presented, with explanations of each format's rationale.
3. **Individual questionnaire:** Participants completed a questionnaire capturing their initial reactions and preferences, with the freedom to review the designs as needed.
4. **Joint discussion:** A moderated group discussion explored participants' thoughts, aiming to uncover deeper insights into usability and preferences.

Data collection included questionnaires, observational notes, and discussion transcripts, which were analyzed using evaluative qualitative content analysis (Kuckartz, 2012). This method involved reviewing and summarizing the data through inductive category formation (Mayring, 2015), focusing on identifying key themes, user preferences, and potential concerns to inform the tool's further development.

Results

These findings provide initial insights into participants' encounters with disinformation, their familiarity with AI technologies, and preferences regarding the presentation of warnings in relation to posts. The following section delves into further details derived from these responses.

In response to the question "Have you already encountered disinformation? If so, where?" five out of the eight participants confirmed that they have encountered instances of disinformation. The platforms most frequently cited for encountering disinformation include social media platforms such as Facebook, X (formerly Twitter), YouTube, and Instagram. Participants noted that these encounters primarily revolved around political discourse, occurring in both public forums and private discussions. When asked about their experience with AI-based systems, specifically where they have consciously gained experience, six out of the eight participants indicated they have used AI-based systems before. Common experiences cited include interacting with generative AI (ChatGPT) and other chatbots.

Before receiving an explanation of the system's classification, users were provided with a brief warning indicating that a post was labeled as potential disinformation. Regarding the placement of warning messages in relation to classified posts, participants were asked, "Where should the warning be placed (before the post, after the post, or post hidden)?" Six out of the eight participants expressed a preference for having the warning displayed above the post. When asked to choose between brief and detailed explanations for AI classifications, all eight participants preferred the longer version of the text. Common explanations for this strong preference were the increased trust and understandability provided by more comprehensive explanations. Participants claimed that it "should be possible to find out why the AI classified a post in this way" [TN2] and that "it makes the reference more credible, and this strengthens trust in the AI" [TN6]. However, it was also posited that detailed explanatory texts could potentially induce fatigue over extended periods:

> "What may also be annoying for some - not for me - is the length of the text. If it feels like it pops up with every post, you definitely lose interest at some point. But on the other hand, I wouldn't really know how to minimize that." [TN4]

Furthermore, participants exhibited diverse perspectives regarding the display of confidence scores used in the context of disinformation detection. Several participants expressed a consistent preference for the utility of confidence scores, with four individuals finding them consistently helpful. Conversely, three participants indicated that they never found confidence scores helpful, while four others believed they were only beneficial if they exceeded a specific threshold. The threshold for what constitutes a helpful confidence score varied considerably among participants, ranging from as low as 20% to as high as 80%. Although the concept of confidence scores was explained to all participants during the briefing and in the questionnaire, it became evident that the comprehension of confidence scores poses challenges for laypersons, rendering them prone to misinterpretation. Consequently, this factor impacts the perceived utility of displaying such scores and the perceived usefulness of the provided information. One participant raised concerns about the clarity of low percentage scores without concrete

examples [TN2], while another participant made the following statement:

> "I don't think measuring in percentages is a suitable unit of measurement for comments in a forum. In reality, every post on the forum will not be 100% compliant, and it just becomes visually annoying that an AI is checking people." [TN7]

Here, it becomes obvious that confidence scores can be easily misunderstood as to what they actually refer to. If users assume, for example, that such a score evaluates the credibility of a person instead of the system's own confidence in its prediction, one can expect a corresponding rejection of its display. Additionally, participants were asked which variant of confidence score display (as a percentage or gradation in low, medium, and high) they preferred if such a score were to be shown. Here, a clear preference became visible: Seven out of eight individuals favored a percentage display. One person stated that they would find a display in percentages "clear and comprehensible—"medium" is kind of vague so I would rather interpret it, hm, that's a bit unclear now. Whereas with "67% certain" I would have the feeling that I have clear information. Seems precise, convincing, as if the AI knows what it's doing." [TN2]. Other participants expressed similar sentiments, claiming that they "can visualize the probability better with percentages" [TN3] and that a display of gradations does "not provide me personally with a basis on which I want to rely" [TN1]. However, one participant offered a contrasting viewpoint, preferring simpler classifications:

> "It's a simpler classification with three levels. At up to 100% everyone assesses the situation for themselves. Some find 60% completely reliable and some only from 90%, for example. The percentage variant offers too much scope for interpretation and making decisions based on gut feeling." [TN4].

These varied responses illustrate a general preference for percentage-based confidence scores, although some participants see value in more straightforward classification. The divided opinions on the usefulness of a confidence score stand in stark contrast to the consensus regarding the importance of highlighting text passages relevant for classification: All eight participants found it helpful to display the text passages that the AI considers indicative of disinformation. In this context, individuals indicated that the highlighting serves multiple functions for them, going beyond the direct interaction with the system:

> "This also makes the AI's advice reliable and ensures that it is given more credence. At the same time, it sensitizes the reader to recognize disinformation more easily in the future." [TN6]

Other participants added that the highlighting helps to "understand and comprehend things better" [TN8], making it "transparent how the AI has assessed what has been classified as 'red' and I can check for myself what I think of it." [TN7]. These unanimous responses highlight the importance of transparency in AI assessments, as displaying specific text passages helps users understand and trust the system's conclusions. Finally, participants were asked how they preferred the AI to display text passages that indicated disinformation. Five out of eight participants favored color highlights in the original posts instead of citing relevant text passages in the explanatory text. One attendee explained their preference for colored highlights in the original post as follows:

> "Striking colors are an eye-catcher. It also reminds me a bit of my school days: important information was marked with a highlighter, here too. So why make it complicated and quote the article again in a large block instead of making the info text short and concise and simply using and including the existing post?" [TN4].

Another participant supported this preference, claiming that "readers are shown even more clearly which passages and statements are involved" [TN5] and "text passages can be found much more quickly" [TN5]. In contrast, participants who preferred citations of the passages in the explanatory text argued that this variant is better structured. One person stated that they find "[colorful highlights] too confusing, as you have to constantly open pop-up windows for an explanation." [TN6]. Consequently, they claimed this "would discourage me from reading the explanations and thus deprive me of the opportunity to gradually recognize disinformation myself" [TN6]. These statements highlight the participants' general preference for color highlights in the original posts, as this method is seen as more intuitive and clear. However, a significant minority preferred citations in the explanatory text for better structure and ease of understanding.

Stakeholder communication (E)

Hevner et al. (2004) stress the importance of effective communication of DSR research results "both to a technical audience (researchers who will extend them and practitioners who will implement them) and to a.

managerial audience (researchers who will study them in context and practitioners who will decide if they should be implemented within their organizations)" (Hevner et al., 2004, p. 82). We followed this approach and presented our artifact through various presentations at practitioners' conferences and through media outlets to provide critical insights into user interactions with XAI-based systems for disinformation detection. Furthermore, we communicated and discussed results focusing on an expert audience,

conducting a round table format together with 16 researchers from various disciplines and practitioners from domains including politics, citizen participation, communication science, machine learning, and fact-checking in March 2024. Through these discussions, we identified key areas that require attention in the development of XAI tools, specifically emphasizing the importance of comprehensibility, user-friendliness, and trustworthiness. The expert feedback underscored the necessity for designing systems that are not only effective in detecting disinformation but also comprehensible and reliable from a user perspective.

Second DSR cycle

Revision of the first cycle and objective refinement (A, B)

The qualitative user testing's findings and the round table discussion underscore the importance of designing AI-based disinformation detection systems that are transparent, user-friendly, and trustworthy. By addressing user preferences and concerns early on, developers can create more effective tools that not only detect disinformation but also educate and empower users to navigate digital spaces more critically. This approach is essential for fostering a more informed and resilient digital public. Accordingly, alongside our formulated guidelines, the results discussed at the round table inform the further development of our prototype by deciding which design choices can be implemented directly (indicated by the participants' consensus) and which design choices may need further testing in the future (indicated by the participants' disagreement or varying preferences). Therefore, the initial warning appears above the article. Users can view the explanation by clicking on "Read more" (Guidelines 1 and 2). The system provides a detailed explanation: The note explains the characteristics on which the classification is based and which text passages the AI is referring to (Guidelines 3 and 6). Although there is a slight observable preference for displaying a confidence score, it may only be displayed above a certain value. Accordingly, several prototype variants are designed to address these diverse needs. One variant will offer explanations without a confidence score, while the other will include it (given in percentage) (Guideline 4). Furthermore, the specific text passages of a post that are relevant to the AI's classification shall be displayed. Participants favored both the option of color highlighting in the original post and the citation of the text passages in the explanatory text of the classification. As there was a slight tendency towards color highlighting in the original post, this tendency will be reflected in the design of the prototypes for the quantitative study (Guideline 5).

Our next step is to test these prototypes through an online study to evaluate the effectiveness of these design choices based on user interactions (Guidelines 8 and 9). To frame our design process within a broader context, we draw on empirical literature examining the impact of XAI on user perceptions. XAI has emerged as a pivotal approach to bridge the gap between complex AI models and user comprehension. Understanding AI systems' working principles is crucial for users to make informed decisions in various contexts (Haque et al., 2023). In particular, XAI plays a vital role in enhancing its understanding. Understandability specifies whether the features and attributes of a model are easily recognizable by users without knowing its inner composition. XAI ensures that AI systems are not just accurate but also interpretable and transparent, making their operations more comprehensible to users (Arrieta et al., 2020). When explanations are presented appropriately, user understandability significantly increases (Bussone et al., 2015; Cai et al., 2019; Eiband et al., 2019; Hudon et al., 2021). Experimental research shows that a user's knowledge about the system's interactions results in better understandability of the system (Bove et al., 2021; Branley-Bell et al., 2020; Cheng et al., 2020). Users who can grasp how an AI system functions are more likely to find it user-friendly and reliable (Górski & Ramakrishna, 2021). For non-technical stakeholders, clear, concise, and comprehensive information is essential to avoid cognitive overload (Hudon et al., 2021). Properly labeled and explained attributes, along with well-reasoned decisions, are crucial for increasing user understandability (Li et al., 2021). Accordingly, we hypothesize the following: H_1: XAI leads to a higher degree of perceived understandability compared to AI without an XAI component.

Trust in AI systems can be bolstered by providing contextual information and transparent decision-making processes (Bove et al., 2021; Cirqueira et al., 2020; Wang et al., 2019). Moreover, a high confidence level for predictions helps users build trust in the system (Bussone et al., 2015; Ehsan et al., 2021). The explanation should contain enough details regarding the prediction and decision-making procedure so that users can feel confident and trust the system. Too much information could create cognitive overload and decrease users' understanding and trust (Cramer et al., 2008; Hudon et al., 2021; Schmidt et al., 2020). To promote trust in the system, it is recommended to reduce the knowledge gap between the user and the system by collaborating with users during the XAI development lifecycle (Chromik, 2021; Hong et al., 2020; Park et al., 2021). Therefore, we formulate the following hypothesis: H_2: XAI leads to a higher degree of trust in the system compared to AI without an XAI component.

XAI systems are also shown to positively impact usability (Oh et al., 2018), potentially leading to higher technology acceptance (Davis & Grani, 1989; Venkatesh & Davis, 2000). Furthermore, to increase usability, accessible and interactive interfaces should be designed and developed for non-technical stakeholders (Andres et al., 2020; Brennen, 2020). Involving the stakeholders in the development

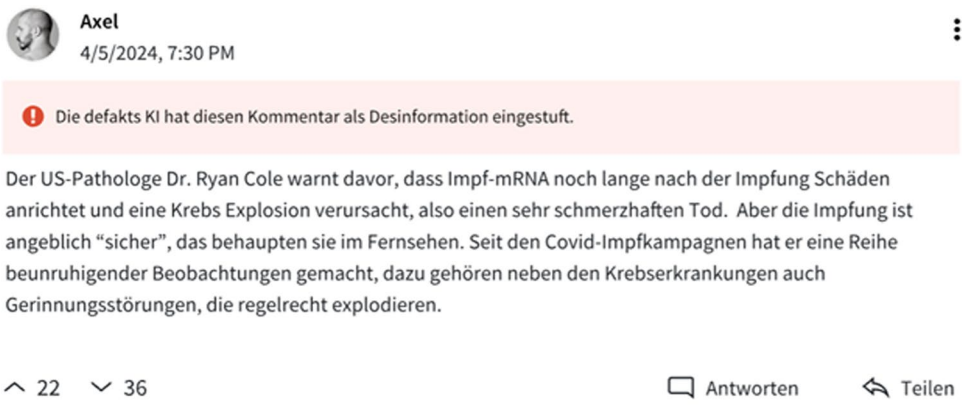

Fig. 7 First design prototype without explanations

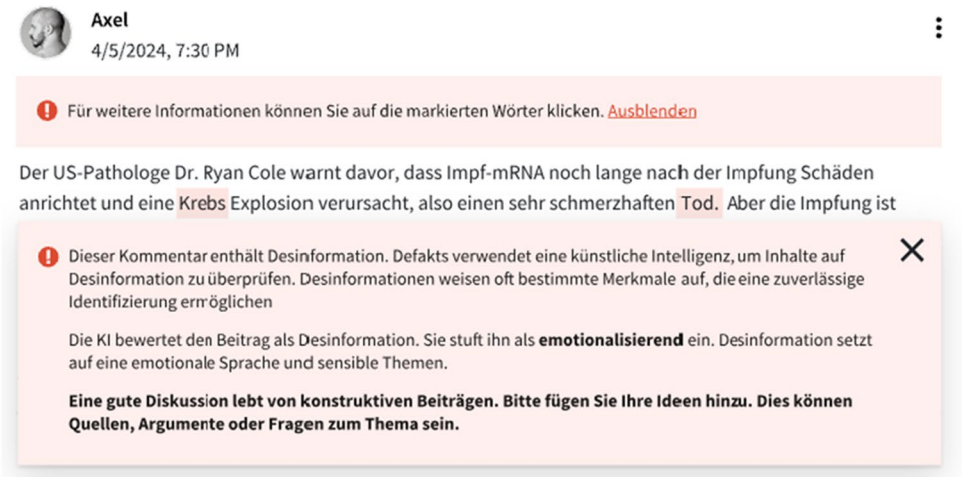

Fig. 8 Second design prototype with explanations

lifecycle may also increase a system's usability (Chromik 2021). Therefore, we formulate the following hypothesis: H_3: XAI leads to a higher degree of perceived usability compared to AI without an XAI-component.

These findings and the iterative development process help refine our prototype and establish a framework for the subsequent demonstration and evaluation phase. As we proceed with an online study to assess the impact of these design choices, this approach is situated within the broader context of XAI's influence on user perceptions. Through systematic testing and iteration, the aim is to critically assess the final system's effect on the transparency metrics of understandability, trust, and usability (Haque et al., 2023).

Prototypes of an XAI interface (C)

Building on the qualitative study's user feedback, we refined our interface prototypes to enhance user experience and functionality. This iterative development process has led to the creation of three interactive design prototypes for a discussion platform. Each prototype integrates an AI-based system that monitors contributions to a digital discussion and flags suspicious content as potential disinformation. The prototype (Fig. 7), used for the first treatment, serves as our baseline system. It shows the system's binary classification of suspicious content without providing further explanations on the system's reasoning behind its prediction.

The second prototype (Fig. 8), used for the second treatment, shows users the system's classification and provides them with additional explanations in an expendable window. Similar to our baseline prototype, a banner appears above each classified post. Upon clicking on "Read more", users are provided with an explanation of why the system recognized the post as disinformation as well as in which way the recognized characteristics are indicative of disinformation.

Our third prototype (Fig. 9), later used for the third treatment, provides explanations identical to the ones of our second prototype but supplements them with a confidence score. In this part of the explanation, the system communicates its confidence in its own prediction.

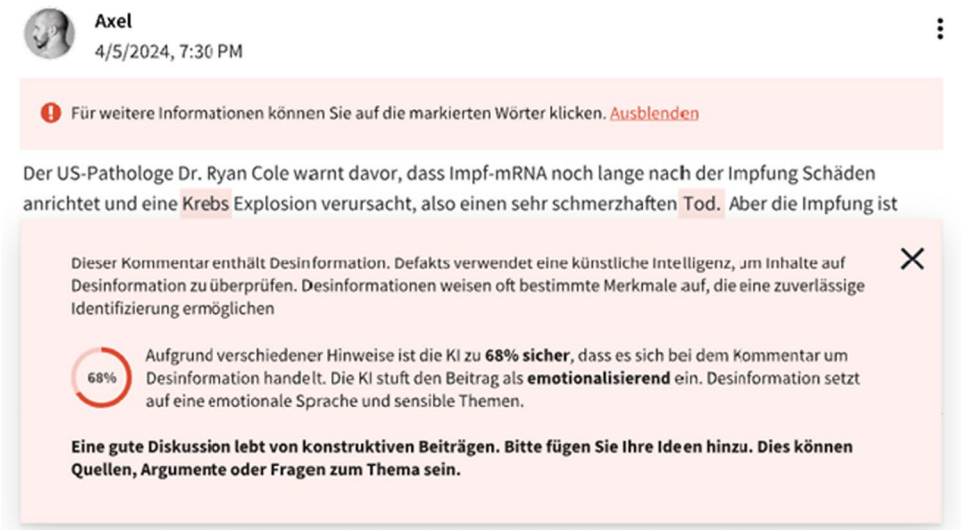

Fig. 9 Third design prototype with explanations and confidence score

Quantitative online study (D)

Strengthening the evaluative rigor of the first DSR cycle, we demonstrate and evaluate our solution artifact in a quantitative experimental approach. In the following, we will first present our approach to designing and conducting the online study before delving into its results.

Procedure

The study, conducted in July 2024, involved the recruitment of 400 participants through the online panel provider Prolific. To ensure data quality, a pre-test was conducted, and two attention-check (AC) questions in the form of instructional manipulation checks (IMCs) were included in the questionnaire. These questions were designed to identify inattentive respondents. The first AC question was positioned in the middle of the questionnaire, while the second was placed toward the end. Participants who failed one or both of these questions were excluded from further analysis, resulting in the removal of 56 respondents. After this exclusion process, a total of 344 participants remained in the dataset for analysis. Participants were selected based on their demographic diversity, with considerations for age (mean = 32.02, SD = 10.38) and sex (171 male and 173 female). Each participant was presented with one clickable prototype. Participants were informed that the study aimed to investigate user perceptions of an AI-supported tool for detecting disinformation on digital platforms, such as discussion forums. They were provided with a brief overview of the study, including the expected duration of approximately 30 min, and were encouraged to respond to the questionnaire honestly and carefully via the initial instructions in the questionnaire. In this study, the prototypes featured an online discussion on the topic of new vaccination technologies in the context of the COVID-19 pandemic (Fig. 10). The comment section displayed six comments, two of which were classified as disinformation. The study followed a between-subjects design (Charness et al., 2012) and included three experimental treatments each aligning with one of our three prototypes.

Participants were randomly assigned to one of the three groups to ensure that any observed differences in outcomes could be attributed to the experimental manipulation rather than pre-existing differences among participants. The key concepts under investigation included participants' trust in the presented AI-based system, perceived understandability of the provided information, and their overall usability experience (see Table 2 in Appendix A). On the basis of established theoretical constructs, these concepts were measured on a 1–7 Likert scale (fully disagree to fully agree). Additional measures were taken to assess participants' demographic characteristics, their propensity to trust, and their prior experience with AI. In order to explore the effects of the experimental conditions on participants' perceptions and behaviors while also accounting for demographic variables, we conducted Kruskal–Wallis tests complemented by Dunn–Bonferroni post-hoc tests and linear regression analyses.

Reliability analyses were conducted for each of the constructs used in the study (see Table 2). The *understandability* construct, consisting of five items (Madsen & Gregor, 2000), showed very good reliability. Further, the *trust* construct, consisting of six items (Merritt, 2011), indicated excellent internal consistency among the items. The *usability* construct, measured by five items (Benbasat & Wang, 2005), demonstrated good reliability. These findings indicate that the items within each scale are sufficiently consistent to be considered reliable measures of their respective constructs.

Before conducting the primary analyses, the normality of the data was tested using the Shapiro–Wilk test. The results ($p < 0.001$) indicated a significant deviation from normality,

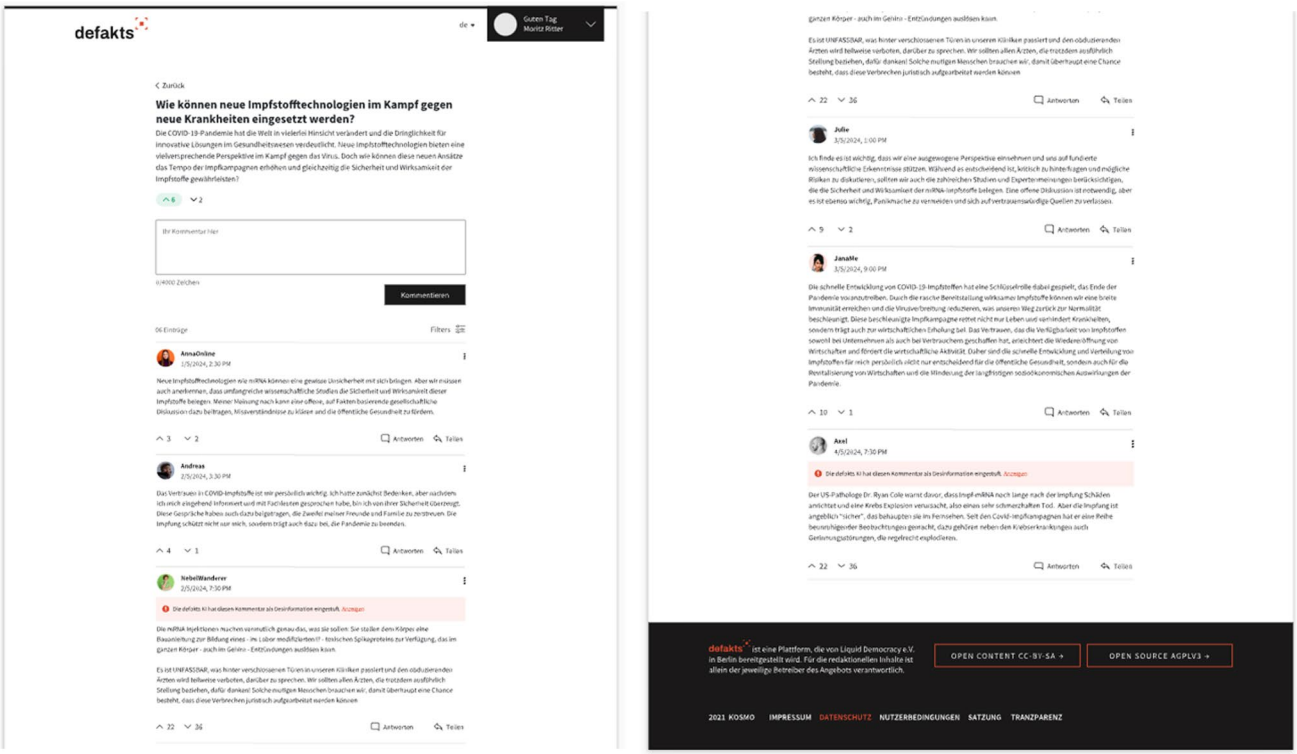

Fig. 10 Clickable user interface of the discussion with two classified posts

Table 2 Summary of reliability analyses for the measured constructs

Construct	Number of items	Cronbach's alpha	Average inter-item correlation	Guttman's Lambda 6	Standard error
Understandability	5	0.88	0.60	0.87	0.010
Trust	6	0.91	0.62	0.90	0.008
Usability	5	0.83	0.50	0.80	0.015

violating one of the key assumptions required for parametric tests such as ANOVA. Given this violation, non-parametric tests were used for the main analyses. To compare the effects of the three treatment conditions, the Kruskal–Wallis test was employed as a non-parametric alternative to the one-way ANOVA. This test was used to assess whether there were statistically significant differences in the dependent variables (e.g., trust, understandability, and usability) across the three experimental groups. In addition, multiple linear regression analyses were conducted to explore the impact of demographic and personal factors (e.g., age, educational background, and previous AI experience) on the dependent variables. These analyses allowed for examining how these characteristics might influence participants' responses independent of the treatment effects. Informed consent was obtained from all participants before their involvement in the study, and they were assured of the confidentiality and anonymity of their responses.

Results

The three explainability levels' effect on participants' perceptions of the system's trustworthiness, perceived usability, and understandability and, as an additional insight, their overall agreement with the displayed classifications was analyzed (see Table 3 and Figs. 11, 12, 13, and 14) and is presented in the following.

Understandability. A Kruskal–Wallis test indicated no significant difference in understandability among the three treatment groups with a negligible effect size (Fig. 11). Median scores were identical at 5.00 across all treatments. These findings suggest that the inclusion of XAI components does not significantly enhance understandability compared to a basic AI system. Given the consistent median scores and small effect sizes, we cannot confirm hypothesis H_1, which proposed that XAI components would improve understandability.

Table 3 Summary statistics of Kruskal–Wallis test and post-hoc analyses (Dunn–Bonferroni test, Cohen's *d*)

Statistic	Understandability	Trust	Usability	Classification agreement
Median scores	Treatment 1: 5.00 Treatment 2: 5.00 Treatment 3: 5.00	Treatment 1: 5.17 Treatment 2: 4.58 Treatment 3: 4.67	Treatment 1: 5.40 Treatment 2: 5.60 Treatment 3: 5.40	Treatment 1: 6.00 Treatment 2: 5.00 Treatment 3: 5.00
χ^2(df)	$\chi^2(2) = 1.03$	$\chi^2(2) = 7.91$	$\chi^2(2) = 0.76$	$\chi^2(2) = 15.64$
p-value	$p = 0.60$	$p = 0.02$	$p = 0.68$	$p < 0.001$
e^2_{ordinal} (95 CI)	$e^2_{\text{ordinal}} = 0.003$ (95% CI [0.000446, 1.00])	$e^2_{\text{ordinal}} = 0.02$ (95% CI [0.0474, 1.00])	$e^2_{\text{ordinal}} = 0.002$, (95% CI [0.000441, 1.00])	$e^2_{\text{ordinal}} = 0.05$ (95% CI [0.02, 1.00])
Post-hoc test Treatment 1 vs 2	$Z = -0.09$, p.adj = 1.00, $d = -0.12$	Z = 2.71, p.adj = 0.14, $d =$ 0.26	$Z = -0.82$, p.adj = 1.00, $d = -0.09$	$Z =$ 3.15, p.adj = 0.0048, $d =$ 0.43
Post-hoc test Treatment 1 vs 3	$Z = -0.92$, p.adj = 1.00, $d = -0.12$	$Z =$ 2.00, p.adj = 0.14, $d =$ 0.26	$Z = -0.67$, p.adj = 1.00, $d = -0.07$	$Z =$ 3.65, p.adj = 0.0008, $d =$ 0.43
Post-hoc test Treatment 2 vs 3	$Z = -0.83$, p.adj = 1.00, $d = -0.09$	$Z = -0.72$, p.adj = 1.00, $d = -0.06$	$Z = -0.15$, p.adj = 1.00, $d =$ 0.01	$Z =$ 0.48, p.adj. = 1.00, $d =$ 0.12

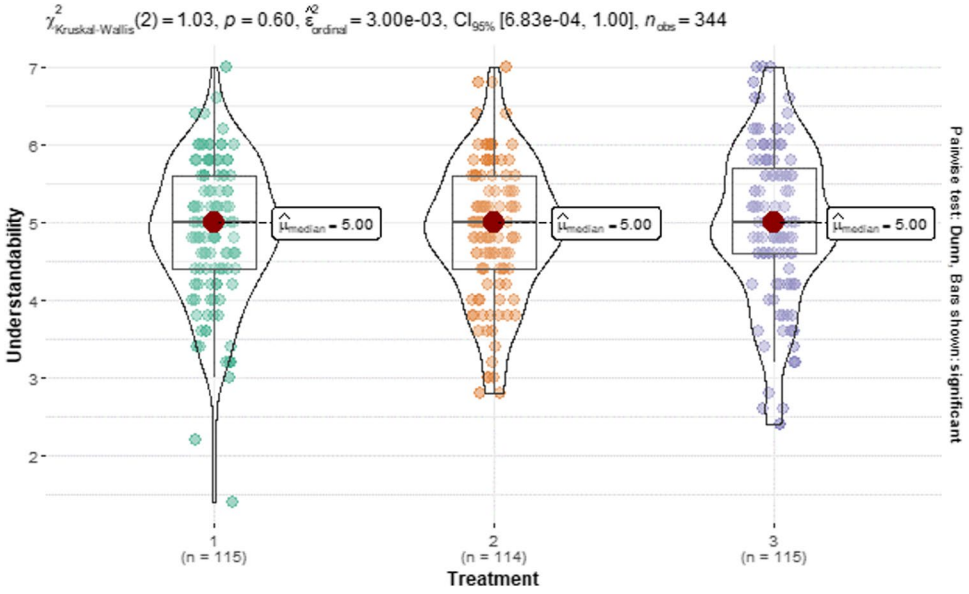

Fig. 11 Kruskal–Wallis test of perceived understandability

Trust. The Kruskal–Wallis test revealed a significant difference in trust scores among the three treatment groups with a small effect size (Fig. 12). Median trust scores were 5.17 for treatment one, 4.58 for the second treatment, and 4.67 for treatment three. Dunn–Bonferroni post-hoc tests showed a significant difference between treatment one and treatment two with a small effect size. However, the difference between treatment one and treatment three was not significant. Furthermore, no significant difference was found between the second and third treatment with a negligible effect size. These results indicate that while there is a small but significant difference in trust between the control group and the group with explanations, the presence of XAI components does not lead to higher trust overall. Consequently, we cannot confirm H_2.

Usability. The results of the Kruskal–Wallis test show no significant differences in usability scores across the treatment groups with a negligible effect size (Fig. 13). Median usability scores were 5.40 for treatment one, 5.60 for treatment two, and 5.40 for treatment three. Confirming the initial observation, Dunn–Bonferroni post-hoc tests also found no significant pairwise differences between treatment one and two, treatment one and three, and treatment two and three. These observations suggest no significant differences in perceived usability among the treatment groups, and the presence of XAI components does not enhance usability over a basic AI system. As such, we cannot confirm H_3.

Classification agreement. For additional insights, we investigated potential differences between the treatments regarding the participants' overall agreement with the

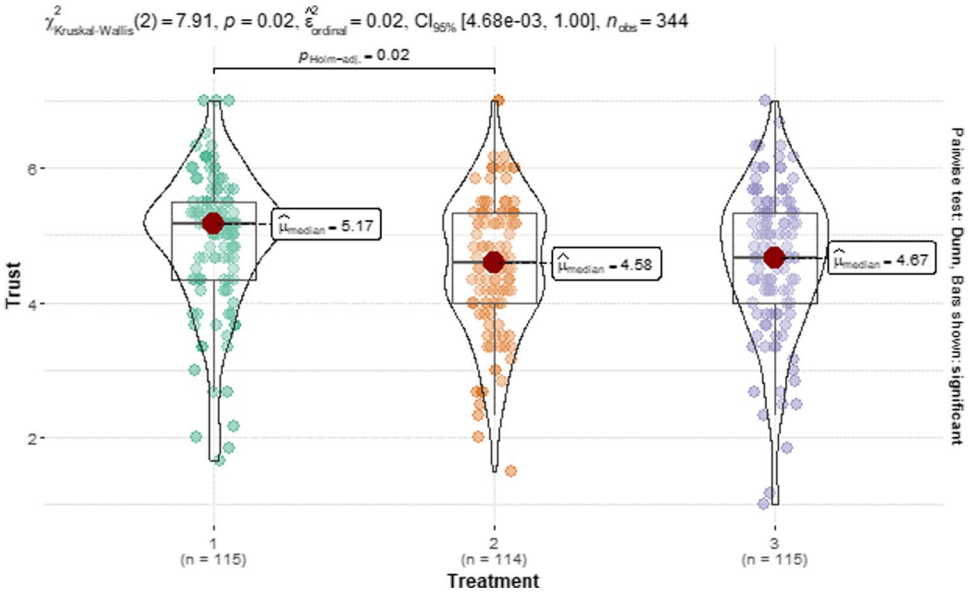

Fig. 12 Kruskal–Wallis test of trust in the system

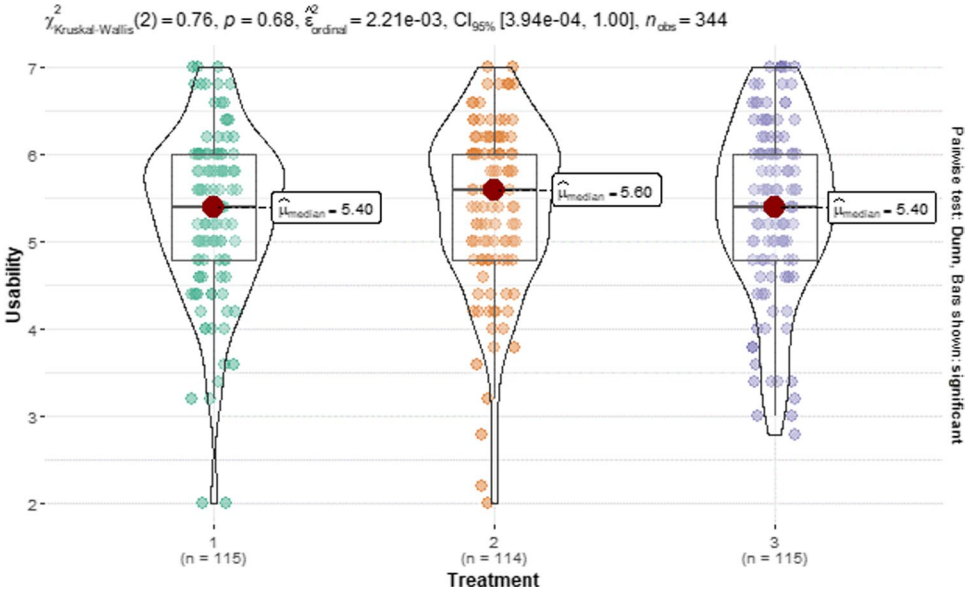

Fig. 13 Kruskal–Wallis test of perceived usability

displayed classifications (Fig. 14). The Kruskal–Wallis test revealed a significant effect with a moderate effect size. Median classification agreement was highest in the control group (6.00), while both treatment groups scored lower (5.00). Dunn–Bonferroni post hoc tests showed that the control group had significantly higher agreement scores compared to both treatment two and treatment three, suggesting small to moderate differences. No significant difference was found between the two XAI treatment groups, with a minor effect size. These findings indicate that the introduction of explanations, with or without confidence scores, actually reduced participants' agreement with the system's classifications.

Impact of demographic and personal characteristics. In previous analyses, we examined the impact of the different treatments, varying in their degrees of explainability, on the measured constructs. This analysis extends our understanding by employing linear regression to explore additional factors associated with these constructs. Table 4 presents the results of four separate linear regressions, each assessing the relationships between various predictors and our four distinct dependent variables: understandability, trust, usability, and classification agreement. The models include the predictors' age, gender, academic background, prior experience with AI, individual propensity to trust, and treatment membership. The model for understandability (1) suggests that

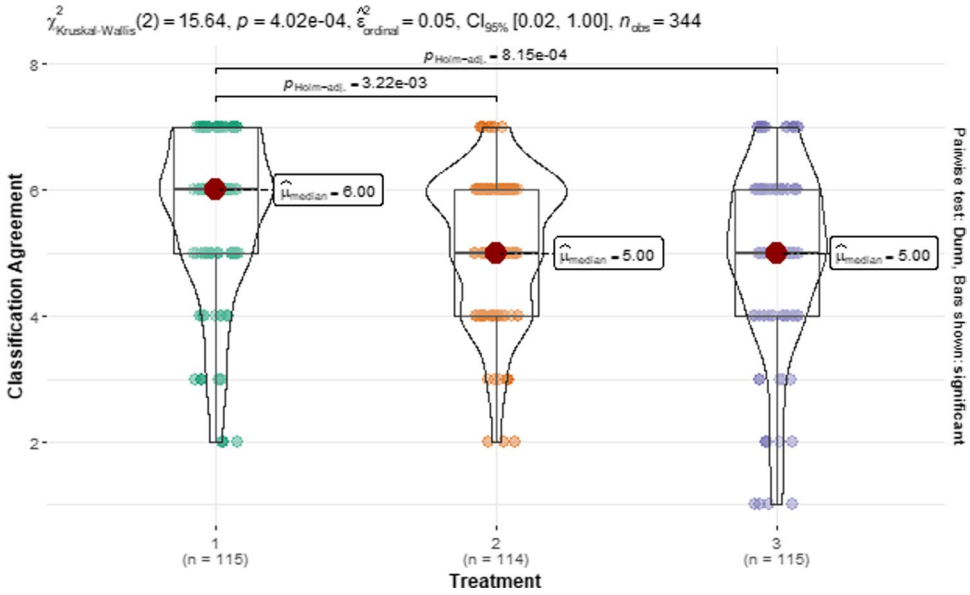

Fig. 14 Kruskal–Wallis test of classification agreement

older individuals tend to perceive understandability as lower ($\beta = -0.014$, p < 0.01). Similarly, individuals with higher levels of general trust (propensity to trust) are more likely to perceive greater understandability ($\beta = 0.101$, $p < 0.01$). Other variables, including gender, academic background, and AI experience, are not significantly associated with understandability in this model. For trust (2), older individuals report lower levels of trust in the system ($\beta = -0.016$, $p < 0.01$), while individuals with a higher propensity to trust exhibit higher levels of reported trust ($\beta = 0.178$, $p < 0.001$). Gender, academic background, and prior AI experience are not significantly associated with trust in this model.

For usability (3), age is negatively associated with perceived usability ($\beta = -0.015$, $p < 0.01$), suggesting that older individuals report a lower perception of usability than younger participants. Notably, gender also shows a significant association, with female participants reporting lower levels of usability compared to male participants ($\beta = 0.339$, $p < 0.001$). In contrast, an individual's propensity to trust has a positive association with usability ($\beta = 0.084$, $p < 0.05$), while academic background and prior experience with AI do not exhibit significant associations. Regarding the overall agreement with the displayed classifications (4), age is negatively associated with agreement with the system's classifications ($\beta = -0.018$, $p < 0.01$). Individuals with a higher propensity to trust exhibit higher agreement levels ($\beta = 0.200$, $p < 0.001$), while female participants report lower levels of agreement compared to male participants ($\beta = -0.303$, $p < 0.05$). Neither academic background nor prior experience with AI is significantly associated with classification agreement. Interestingly, the treatment conditions do not exhibit significant associations with perceived understandability, trust, or usability. However, both treatment two ($\beta = -0.345$, $p < 0.05$) and treatment three ($\beta = -0.546$, $p < 0.001$) are significantly associated with lower classification agreement compared to the control group. Participants provided with explanations from the system exhibited lower agreement rates with its predictions, particularly in treatment three, where the explanatory text included a confidence score.

Discussion

Our study explored how different degrees of explainability, including explanations with or without confidence scores, impact user perceptions across multiple constructs, such as understandability, trust, usability, and classification agreement. The findings reveal that these treatments had a minimal effect on participants' evaluations, suggesting that additional underlying factors, such as demographic and individual characteristics, play a more significant role in shaping user experiences and perceptions (Schemmer, 2022).

The analysis indicates that the presence of explanations, whether with or without a confidence score, did not significantly affect understandability among the different treatment groups. Consequently, the results do not support the notion that XAI components improve understandability compared to a basic AI system without such components. The observed lack of improvement in understandability may be attributed to cognitive overload and issues with the relevance of the explanations provided (Tsai et al., 2021; Liu et al., 2021; Sanneman & Shah, 2022). Specifically, the data suggest that as age increases, perceived understandability tends to decrease, possibly because older participants may find complex or technical explanations more challenging. In contrast, higher levels of general trust are positively associated with

Table 4 Results of our linear regression

Dependent variable:				
	Understandability (1)	Trust (2)	Usability (3)	Classification agreement (4)
Age	− 0.014** (0.005)	− 0.016** (0.006)	− 0.015** (0.005)	− 0.018** (0.007)
Female	− 0.022 (0.098)	− 0.014 (0.111)	− 0.339*** (0.101)	− 0.303* (0.134)
Academic status	− 0.192 (0.103)	− 0.022 (0.117)	− 0.103 (0.106)	− 0.231 (0.142)
AI experience	0.306 (0.208)	− 0.088 (0.237)	0.151 (0.216)	− 0.070 (0.287)
Trust propensity	0.101** (0.036)	0.178*** (0.041)	0.084* (0.038)	0.200*** (0.050)
Treatment two	0.089 (0.120)	− 0.263 (0.136)	0.112 (0.124)	− 0.345* (0.165)
Treatment three	0.155 (0.119)	− 0.214 (0.136)	0.087 (0.123)	− 0.546*** (0.164)
Constant	4.712*** (0.314)	4.707*** (0.357)	5.509*** (0.324)	5.604*** (0.432)
Observations	341	341	341	341
R^2	0.069	0.086	0.077	0.118
Adjusted R^2	0.050	0.067	0.058	0.101

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

greater perceived understandability, indicating that those who are more trustful are likely to find the system's explanations clearer. Additionally, factors such as gender, academic background, and AI experience did not show significant effects on understandability, suggesting that the effectiveness of explanations may be more closely related to cognitive factors and trust rather than demographic or experience-based differences. These findings reinforce the need for platform operators to carefully tailor explanation formats to user profiles to maintain accessibility and perceived value across diverse user segments (Gregor & Hevner, 2013; Rai, 2020).

Moreover, the analysis reveals that the presence of explanations without confidence scores was associated with a lower level of trust compared to the control group. Further, adding confidence scores to the explanations did not significantly enhance trust compared to the control group, indicating that confidence scores alone may not effectively enhance trust unless combined with other supportive elements (Hamm et al., 2023; Schmidt et al., 2020). This suggests that participants might have perceived the explanations as less straightforward or more confusing than simply receiving no explanations at all (Papenmeier et al., 2019; Poursabzi-Sangdeh et al., 2021).

The presence of explanations, whether with or without a confidence score, did not significantly affect perceived usability among the treatment groups. The negligible effect sizes and similar median usability scores across groups suggest that the treatments had no meaningful impact on participants' perception of usability. Consequently, the results do not support the notion that XAI components improve usability compared to a basic AI system without such components. Despite being informed by qualitative user testing, the lack of significant impact on perceived usability from explanations, whether with or without a confidence score, may be attributed to several factors. First, the explanations, even when designed based on user feedback, may not have effectively addressed all aspects of usability or aligned with users' specific interaction needs, e.g., when explanations seem unintuitive (Mohseni et al., 2021b; Schmidt et al., 2020). Second, it is possible that these explanations might not have sufficiently altered users' overall experience or efficiency with the system (Schemmer, 2022; Wanner et al., 2022). These results emphasize the broader challenge in integrating AI-driven features within platform interfaces without disrupting core user-flows, a critical concern in the business design of digital platforms (Lyytinen et al., 2021). Currently, the existing XAI literature lacks a comprehensive set of methodologies and metrics for effectively assessing the quality of explanations (Sanneman & Shah, 2022).

The analysis of participants' classification agreement suggests that the presence of explanations was associated with lower classification agreement compared to the control group. This finding underscores the complex interplay between explainability and user agreement. The significant differences between treatment one and both treatments two and three indicate that the explanations provided in these treatments may have introduced additional uncertainty or complexity (Sanneman & Shah, 2022). Specifically, the inclusion of confidence scores in treatment three and the detailed textual explanations in both treatments may have made the system's decision-making process more transparent but also more challenging to interpret, particularly for users without prior familiarity with AI-based systems. One plausible explanation for this decrease in agreement is that overly detailed or technical explanations might have prompted users to scrutinize the system's classifications more critically, leading to increased doubt or skepticism (Ferguson et al., 2022). While this can be seen as a positive outcome in contexts where critical engagement with AI decisions is desirable, it may not align with the goal of fostering

trust and usability in disinformation detection tools. Furthermore, explanations that incorporate probabilistic or confidence information can introduce cognitive overload for users who may lack the expertise to interpret such data effectively, exacerbating uncertainty. This observation aligns with prior research suggestion that user trust and agreement can be undermined when explanations are perceived as too complex or ambiguous (Miller, 2019). In platform settings, this could translate into reduced conversion, churn, or lack of confidence in AI-generated outputs, particularly in high-stakes domains like e-commerce or content moderation (Benbya et al., 2020; Rai, 2020).

Our linear regression analysis elucidates several significant determinants impacting the constructs of understandability, trust, usability, and classification agreement, independent of treatment variations. The results consistently demonstrate that age has a negative relationship with each of the dependent variables, indicating that older individuals generally displayed higher aversion when interacting with our system. This trend may be attributed to age-related cognitive and perceptual changes, which could affect how older individuals process and evaluate information (Salthouse, 1992, 1994; Zahodne et al., 2011). Older adults might experience greater difficulty in understanding new concepts, trusting new technical systems, or experiencing high usability due to accumulated experience or changes in cognitive functions (Miller & Bell, 2012; Peters et al., 2008; Salthouse et al., 1999).

Conversely, an individual's propensity to trust exerts a positive influence across all constructs, underscoring the role of individual trustfulness not only in enhancing trust in the system but also in perceived understandability, usability, and agreement with the classifications provided by the (X) AI. This finding highlights the importance of inherent trust levels in shaping perceptions. People who naturally exhibit higher trust are likely to approach information and systems with a more positive outlook, which could enhance their overall experience and evaluation (Fan et al., 2020).

Notably, gender differences are evident in our findings as being female is associated with a lower perceived usability and lower classification agreement. This indicates that, with regard to some elements, female participants potentially perceive the system less favorably compared to their male counterparts. The observed discrepancy may stem from varying expectations, experiences, or societal factors that affect how different genders interact with and evaluate systems (Reeder et al., 2023). Further research is needed to explore the underlying causes of these gender-related differences, including potential biases in system design or differences in interaction styles. In contrast, whether someone has an academic degree or prior experience interacting with AI has no significant influence on any of the constructs. This may imply that educational background and prior exposure to AI-based systems are less influential in shaping user experience than other individual characteristics, such as age and trust propensity. From a platform design perspective, these insights suggest that adaptive personalization, based on traits like age and trust propensity, may help mitigate usability barriers and enhance engagement across heterogeneous user bases (Berente et al., 2021; Lyytinen et al., 2021). In summary, this analysis extends our understanding of how individual differences shape user perceptions, highlighting the significance of age and trust propensity while indicating the need for further exploration into gender-related differences. This broader perspective complements our findings related to treatment variations, offering a more comprehensive view of the factors influencing user evaluations while hinting at the need for the design of adaptive systems (Kabudi et al., 2021).

Integrated design guidelines (E)

Finalizing our second DSR cycle by building on our previous design guidelines, we further refine and expand our approach to developing responsible XAI systems for disinformation detection. Considering our empirical findings, the integrated guidelines (Gurzick & Lutters, 2009) highlight user needs and the importance of maintaining a balance between simplicity, clarity, and adaptation while also addressing demographic and individual differences:

1. **Integrate explanations seamlessly into the user experience.** Ensure that explanations are integrated in a way that enhances, rather than disrupts, the overall usability of the system. Since the addition of confidence scores did not significantly improve trust or usability, focus on how explanations are presented and ensure they contribute positively to the user experience without causing confusion.
2. **Simplify explanations to avoid cognitive overload.** Ensure that explanations provided by the XAI system are clear and not overly complex. Given that explanations did not significantly impact understandability, it is crucial to avoid introducing unnecessary complexity. Tailor explanations to be straightforward and relevant to the user's current context to prevent cognitive overload.
3. **Prioritize trustworthiness in design to build credibility for inexperienced users.** Even though confidence scores alone did not significantly enhance trust, ensure that explanations are part of a broader strategy to build system credibility. Develop supportive elements that reinforce trust and reliability, ensuring users perceive the system as trustworthy and effective in detecting disinformation.
4. **Make explanations optional by offering customizable explanation features.** In line with the principle of

user empowerment, explanations should be an optional feature, allowing users to access additional details only when needed. This approach respects the user's autonomy and avoids unnecessary complexity in the overall user experience.

5. **Consider user trust and cognitive factors.** Recognize that inherent trustfulness and cognitive factors may significantly influence how users perceive explanations. Account for cognitive differences, such as those related to age, by simplifying explanations for older users who may struggle with more technical content.
6. **Address demographic and individual differences through adaptability in design.** Design explanations adaptable to different user profiles, acknowledging that factors such as age and trust propensity affect user perceptions and be mindful of potential biases in system design as well as differences in how various demographic groups interact with the system. Consider conducting targeted user research to tailor explanations effectively.
7. **Refine and test explanation mechanisms continuously.** Continuously refine explanation mechanisms based on user feedback and iterative testing. The findings suggest that explanations alone might not improve usability or classification agreement. Regularly test and adjust explanations to better align with user needs and enhance the system's effectiveness.

By adhering to these guidelines, responsible XAI systems for disinformation detection may be developed to better meet user needs, enhance usability, and improve overall effectiveness in combating false information on digital platforms.

Conclusion

Summary

This study addressed the research question of how a responsible XAI-based system for detecting online disinformation should be designed to foster user trust, understandability, and comprehension. By leveraging a Design Science Research (DSR) approach (Peffers et al., 2007), we developed and evaluated explainability features tailored to the high-stakes, sensitive domain of disinformation detection. Through a comprehensive literature review, iterative design cycles, and empirical user testing, we provide both practical design guidelines and important theoretical insights into the limitations and potential of explainable AI (XAI) in real-world applications.

From a theoretical standpoint, this study contributes to an underexplored intersection between XAI and disinformation detection by shifting the focus from purely technical accuracy toward user-centric design principles (Rjoob et al., 2021; Wells & Bednarz, 2021). While transparency is widely recognized as a cornerstone of XAI (Haque et al., 2023), our findings challenge the assumption that greater transparency inherently leads to improved user trust, comprehension, or usability (Schmidt et al., 2020). Contrary to common expectations, the inclusion of XAI components did not significantly enhance participants' understanding or trust in the system and in some cases even introduced confusion or reduced agreement with system outputs. These results emphasize the importance of designing explanations that are not only technically accurate but also cognitively appropriate for the target user group. By demonstrating that explanations can inadvertently increase cognitive load, our study refines existing cognitive load theory and highlights the contextual and individual variability in how users perceive and benefit from XAI. We show that user demographics—particularly age—and individual characteristics like trust propensity significantly influence the effectiveness of explainability features. Older users, for example, reported lower levels of trust, usability, and understanding, suggesting a need for adaptive XAI systems that account for users' cognitive and experiential diversity. Additionally, our application of the DSR methodology underscores the value of integrating theoretical and empirical insights into the iterative development of XAI systems. This study contributes to IS and HCI literature by offering a framework for embedding user feedback early and systematically in the design process, revealing the nuanced trade-offs between transparency, usability, and user trust. Practically, our findings translate into actionable design guidelines for developing responsible, user-aware XAI systems in the disinformation space. These include simplifying explanations to minimize cognitive overload, tailoring them to users' demographic and cognitive profiles, and offering explanations as optional features to preserve user autonomy. Furthermore, we advocate for combining XAI with other trust-enhancing mechanisms, such as user feedback loops, to foster engagement and reliability.

In conclusion, this research advances both theoretical understanding and practical implementation of explainable AI by uncovering the complex interplay between user characteristics, contextual factors, and design choices in disinformation detection systems. While explainability does not universally improve user perceptions, our contributions provide a foundation for future studies to build more adaptive, context-sensitive, and trustworthy XAI systems—crucial for navigating the evolving challenges of disinformation and responsible AI governance in the digital age.

Limitations

While this study offers valuable insights into the responsible design of XAI systems for disinformation detection, some

limitations must be acknowledged to fully contextualize the findings and guide future research. The structured literature review, though comprehensive, is inherently limited by the selection criteria and databases. The focus on specific keywords or publication types may have excluded relevant studies that could provide additional insights or counterpoints. The qualitative user testing's sample allowed for an in-depth exploration of participants' experiences and perspectives; nevertheless, it may not fully represent the diversity of views within the population. We therefore conducted a quantitative study to test the results with a broader range of backgrounds and present more generalizable results. The online study's design was cross-sectional, capturing user perceptions at a single point in time. Longitudinal studies would be beneficial to assess the long-term impact of explainability features on user perceptions. The study observed a reduction in agreement with the system's classifications when explanations were provided. Investigating the content and format of the explanations could reveal whether they contribute to misunderstandings or if alternative presentation methods might improve agreement. Furthermore, focusing on the design of the explanations, rather than also considering their content and providing a broader array of examples with varying textual features, may not fully capture the range of disinformation features users might encounter. Future studies could expand on this by offering participants more diverse examples, which could help identify how different types of explanations interact with varying content and how they affect user perceptions. Finally, our study focuses on the perception of explainability features. Other aspects of algorithmic transparency (such as model accuracy) are also crucial for how users perceive the system and should be considered in future research to develop a more comprehensive approach to responsible AI design. By acknowledging these limitations, future research can deepen our understanding of how to effectively design and implement XAI systems for disinformation detection and other high-stakes applications. Such research can ultimately support platform providers of OSNs in responsibly adopting and integrating AI-based systems for disinformation detection, fostering a more trustworthy and accountable digital platform ecosystem.

Future work

The ethical deployment of AI in cyberspace governance, especially for disinformation detection, requires a thorough examination to safeguard transparency and fairness on digital platforms. Future research may explore several avenues to build on our findings. First, further studies may investigate a broader range of explanation types and their interactions with various user demographics to identify which formats are most effective in different contexts. Second, longitudinal studies could provide insights into how users' perceptions of AI systems develop over time and whether continuous exposure to explanations affects their experience. Third, investigating the integration of explanations with other trust-enhancing features, such as transparency mechanisms and user feedback systems, could offer a holistic approach to improving user interactions with AI in the combat of online disinformation. In conclusion, while explainability is a critical component of responsible AI, its effectiveness in promoting usability, user trust, and comprehension requires careful consideration and tailored implementation. Our study underscores the importance of a nuanced approach to integrating explainability features and highlights the need for ongoing research to refine these mechanisms and better align them with user needs. Building on our findings, future work can contribute to the development of more effective and trustworthy AI-based systems for disinformation detection and beyond.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-025-00799-3.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was carried out with the support of the German Federal Ministry of Education and Research (BMBF) within the project "DeFaktS" (Grant 16KIS1524K).

Data availability The data that support the findings of this study are available from the authors upon reasonable request.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adam, M. T., Gregor, S., Hevner, A., & Morana, S. (2021). Design science research modes in human-computer interaction projects. *AIS Transactions on Human-Computer Interaction, 13*(1), 1–11. https://doi.org/10.17705/1thci.00139

Alt, R. (2021). Electronic markets on digital platforms and AI. *Electronic Markets, 31*(2), 233–241. https://doi.org/10.1007/s12525-021-00489-w

Alves, J., Araújo, T., Marques, B., Dias, P., & Santos, B. S. (2020). Deeprings: A concentric-ring based visualization to understand deep learning models. *2020 24th International Conference Information Visualisation (IV)*, 292–295. https://ieeexplore.ieee.org/abstract/document/9373251/.

Andres, J., Wolf, C. T., Cabrero Barros, S., Oduor, E., Nair, R., Kjærum, A., Tharsgaard, A. B., & Madsen, B. S. (2020). Scenario-based XAI for humanitarian aid forecasting. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3334480.3382903.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: An analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(5), e1424. https://doi.org/10.1002/widm.1424

Ansar, W., & Goswami, S. (2021). Combating the menace: A survey on characterization and detection of fake news from a data science perspective. *International Journal of Information Management Data Insights, 1*(2), 100052. https://doi.org/10.1016/j.jjimei.2021.100052

Apicella, A., Giugliano, S., Isgrò, F., & Prevete, R. (2021). Explanations in terms of Hierarchically organised Middle Level Features. *XAI. It-2021 Italian Workshop on Explainable Artificial Intelligence, CEUR Workshop Proceedings*. https://www.iris.unina.it/retrieve/63969e84-67ca-4b25-887e-a7d8418c96b3/paper4.pdf.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Banerjee, J. S., Mahmud, M., & Brown, D. (2023). Heart rate variability-based mental stress detection: An explainable machine learning approach. *SN Computer Science, 4*(2), 176. https://doi.org/10.1007/s42979-022-01605-z

Basil, V. R., & Turner, A. J. (1975). Iterative enhancement: A practical technique for software development. *IEEE Transactions on Software Engineering, 4*, 390–396.

Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M. F., Schuller, B., & André, E. (2020). Explainable cooperative machine learning with NOVA. *KI - Künstliche Intelligenz, 34*(2), 143–164. https://doi.org/10.1007/s13218-020-00632-3

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems, 6*(3), 4. https://doi.org/10.17705/1jais.00065

Benbya, H., Nan, N., Tanriverdi, H., & Yoo, Y. (2020). Complexity and information systems research in the emerging digital world (SSRN Scholarly Paper 3539079). Social Science Research Network.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly, 45*, 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First Monday*, *21*(11–7). https://doi.org/10.5210/fm.v21i11.7090

Binder, M., Heinrich, B., Hopf, M., & Schiller, A. (2022). Global reconstruction of language models with linguistic rules – Explainable AI for online consumer reviews. *Electronic Markets, 32*(4), 2123–2138. https://doi.org/10.1007/s12525-022-00612-5

Blackman, R., & Ammanath, B. (2022). When—and why—you should explain how your AI works. *Harvard Business Review*, *31*.

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2021). Contextualising local explanations for non-expert users: An XAI pricing interface for insurance. *Joint Proceedings of the ACM IUI 2021 Workshops*, *2903*. https://hal.science/hal-03844389.

Branley-Bell, D., Whitworth, R., spsampsps Coventry, L. (2020). User trust and understanding of explainable AI: Exploring algorithm visualisations and user biases. In M. Kurosu (Ed.), *Human-Computer Interaction. Human Values and Quality of Life* (pp. 382–399). Springer International Publishing. https://doi.org/10.1007/978-3-030-49065-2_27.

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets, 33*, 26. https://doi.org/10.1007/s12525-023-00644-5

Brendel, A. B., Greve, M., Diederich, S., Bührke, J., & Kolbe, L. M. (2020). You are an Idiot!-How conversational agent communication patterns influence frustration and harassment. *AMCIS*. https://core.ac.uk/download/pdf/326836248.pdf.

Brennen, A. (2020). What do people really want when they say they want "Explainable AI?" We asked 60 stakeholders. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3334480.3383047.

Brocke, J. vom, Hevner, A., spsampsps Maedche, A. (2020). *Introduction to Design Science Research* (pp. 1–13). https://doi.org/10.1007/978-3-030-46781-4_1.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *International Conference on Healthcare Informatics, 2015*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Cai, C. J., Jongejan, J., & Holbrook, J. (2019). The effects of example-based explanations in a machine learning interface. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. https://doi.org/10.1145/3301275.3302289.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Cheng, F., Ming, Y., & Qu, H. (2020). Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1438–1447. https://doi.org/10.48550/arXiv.2008.08353

Chromik, M. (2021). Making SHAP rap: Bridging local and global insights through interaction and narratives. In C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, spsampsps K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021* (Vol. 12933, pp. 641–651). Springer International Publishing. https://doi.org/10.1007/978-3-030-85616-8_37.

Cirqueira, D., Nedbal, D., Helfert, M., spsampsps Bezbradica, M. (2020). Scenario-based requirements elicitation for user-centric explainable AI: A case in fraud detection. In A. Holzinger, P. Kieseberg, A. M. Tjoa, spsampsps E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (Vol. 12279, pp. 321–341). Springer International Publishing. https://doi.org/10.1007/978-3-030-57321-8_18.

Colley, A., Väänänen, K., & Häkkilä, J. (2022). Tangible explainable AI - an initial conceptual framework. *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*, 22–27. https://doi.org/10.1145/3568444.3568456.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(1), e1391. https://doi.org/10.1002/widm.1391

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

Das, D., Nishimura, Y., Vivek, R. P., Takeda, N., Fish, S. T., Plötz, T., & Chernova, S. (2023). Explainable activity recognition for smart home systems. *ACM Transactions on Interactive Intelligent Systems, 13*(2), 1–39. https://doi.org/10.1145/3561533

Davis, F. D., & Grani, A. (1989). *The Technology Acceptance Model.*

Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy*

of Sciences, 113(3), 554–559. https://doi.org/10.1073/pnas.1517441113

Dey, A., Radhakrishna, C., Lima, N. N., Shashidhar, S., Polley, S., Thiel, M., & Nürnberger, A. (2021). Evaluating reliability in explainable search. *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)*, 1–4. https://ieeexplore.ieee.org/abstract/document/9582653/.

Dong, J., Chen, S., Zong, S., Chen, T., & Labi, S. (2021). Image transformer for explainable autonomous driving system. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)* (pp. 2732–2737). https://ieeexplore.ieee.org/abstract/document/9565103

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3411764.3445188.

Eiband, M., Buschek, D., Kremer, A., & Hussmann, H. (2019). The impact of placebic explanations on trust in intelligent systems. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312787.

Ekanayake, I. U., Palitha, S., Gamage, S., Meddage, D. P. P., Wijesooriya, K., & Mohotti, D. (2023). Predicting adhesion strength of micropatterned surfaces using gradient boosting models and explainable artificial intelligence visualizations. *Materials Today Communications, 36*, 106545.

European Commission. (2018). *A multi-dimensional approach to disinformation. Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union.

Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. (2020). Investigating the impacting factors for the healthcare. Professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research, 294*(1), 567–592. https://doi.org/10.1007/s10479-018-2818-y

Ferguson, A. N., Franklin, M., & Lagnado, D. (2022). Explanations that backfire: Explainable artificial intelligence can cause information overload. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/3d97g0n3.

Finzel, B., Tafler, D. E., Thaler, A. M., & Schmid, U. (2021). Multimodal explanations for user-centric medical decision support systems. *HUMAN@ AAAI Fall Symposium*.

Garaialde, D., Bowers, C. P., Pinder, C., Shah, P., Parashar, S., Clark, L., & Cowan, B. R. (2020). Quantifying the impact of making and breaking interface habits. *International Journal of Human-Computer Studies, 142*, 102461.

Gerlach, J., Hoppe, P., Jagels, S., Licker, L., & Breitner, M. H. (2022). Decision support for efficient XAI services—A morphological analysis, business model archetypes, and a decision tree. *Electronic Markets, 32*(4), 2139–2158. https://doi.org/10.1007/s12525-022-00603-6

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). *Explaining explanations: An overview of interpretability of machine learning*. 80–89.

Górski, Ł., & Ramakrishna, S. (2021). Explainable artificial intelligence, lawyer's perspective. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 60–68. https://doi.org/10.1145/3462757.3466145.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly, 37*, 337–356. https://doi.org/10.25300/MISQ/2013/37.2.01

Guillemé, M., Masson, V., Rozé, L., & Termier, A. (2019). Agnostic local explanation for time series classification. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 432–439. https://ieeexplore.ieee.org/abstract/document/8995349/.

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Guo, L., Daly, E. M., Alkan, O., Mattetti, M., Cornec, O., & Knijnenburg, B. (2022). *Building trust in interactive machine learning via user contributed interpretable rules*. 537–548.

Gurzick, D., & Lutters, W. G. (2009). Towards a design theory for online communities. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology - DESRIST '09*, 1. https://doi.org/10.1145/1555619.1555634.

Hamm, P., Klesel, M., Coberger, P., & Wittmann, H. F. (2023). Explanation matters: An experimental study on explainable AI. *Electronic Markets, 33*, 17. https://doi.org/10.1007/s12525-023-00640-9

Hanley, H. W. A., & Durumeric, Z. (2023). *Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites* (arXiv:2305.09820; Version 1). arXiv. https://doi.org/10.48550/arXiv.2305.09820.

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change, 186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Heimerl, A., Weitz, K., Baur, T., & André, E. (2020). Unraveling ml models of emotion with nova: Multi-level explainable ai for non-experts. *IEEE Transactions on Affective Computing, 13*(3), 1155–1167.

Hepenstal, S., Zhang, L., Kodagoda, N., & Wong, B. L. W. (2021). Developing conversational agents for use in criminal investigations. *ACM Transactions on Interactive Intelligent Systems, 11*(3–4), 1–35. https://doi.org/10.1145/3444369

Herm, L.-V., Steinbach, T., Wanner, J., & Janiesch, C. (2022). A nascent design theory for explainable intelligent systems. *Electronic Markets, 32*(4), 2185–2205. https://doi.org/10.1007/s12525-022-00606-3

Hevner, A. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems, 19*, 4.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 75–105.

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1–68. https://doi.org/10.1145/3392878

Hoque, M. N., & Mueller, K. (2021). Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making. *IEEE Transactions on Visualization and Computer Graphics, 28*(12), 4728–4740.

Huang, J., Mishra, A., Kwon, B. C., & Bryan, C. (2022). ConceptExplainer: Interactive explanation for deep neural networks from a concept perspective. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 831–841.

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., spsampsps Sénécal, S. (2021). *Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence*. 237–246.

Hwang, J., Lee, T., Lee, H., & Byun, S. (2022). A clinical decision support system for sleep staging tasks with explanations from artificial intelligence: User-centered design and evaluation study. *Journal of Medical Internet Research, 24*(1), e28659.

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence, 2*, 100017. https://doi.org/10.1016/j.caeai.2021.100017

Kadir, M. A., Mohamed Selim, A., Barz, M., & Sonntag, D. (2023). A user interface for explaining machine learning model explanations. *28th International Conference on Intelligent User Interfaces*, 59–63. https://doi.org/10.1145/3581754.3584131.

Kerzel, M., Ambsdorf, J., Becker, D., Lu, W., Strahl, E., Spisak, J., Gäde, C., Weber, T., & Wermter, S. (2022). What's on your mind, NICO?: XHRI: A framework for explainable human-robot interaction. *KI - Künstliche Intelligenz, 36*(3–4), 237–254. https://doi.org/10.1007/s13218-022-00772-8

Khurana, A., Alamzadeh, P., & Chilana, P. K. (2021). ChatrEx: Designing explainable chatbot interfaces for enhancing usefulness, transparency, and trust. *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 1–11. https://ieeexplore.ieee.org/abstract/document/9576440/.

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

Kubat, M., Kubat, J. (2017). *An introduction to machine learning* (Vol. 2). Springer. https://doi.org/10.1007/978-3-319-63913-0

Kuckartz, U. (2012). *Qualitative Inhaltsanalyse: Methoden, Praxis*. Computerunterstützung. Beltz Juventa.

Kumar, A., Manikandan, R., Kose, U., Gupta, D., & Satapathy, S. C. (2021). Doctor's dilemma: Evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. *ACM Transactions on Multimedia Computing, Communications, and Applications, 17*(3s), 1–26. https://doi.org/10.1145/3457187

Kumar, P., & Sharma, M. (2021). Feature-importance feature-interactions (FIFI) graph: A graph-based novel visualization for interpretable machine learning. *2021 International Conference on Intelligent Technologies (CONIT)*, 1–7. https://ieeexplore.ieee.org/abstract/document/9498467/.

Le, T., Miller, T., Singh, R., & Sonenberg, L. (2023). *Explaining model confidence using counterfactuals*(arXiv:2303.05729). arXiv. http://arxiv.org/abs/2303.05729.

Lehrer, C., Wieneke, A., vom Brocke, J., Jung, R., & Seidel, S. (2018). How big data analytics enables service innovation: Materiality, affordance, and the individualization of service. *Journal of Management Information Systems, 35*(2), 424–460. https://doi.org/10.1080/07421222.2018.1451953

Lewis, M., Li, H., & Sycara, K. (2021). Deep learning, transparency, and trust in human robot teamwork. In Trust in Human-Robot Interaction (pp. 321–352). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128194720000149.

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (2021). Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176. https://doi.org/10.1145/3461702.3462531.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Linder, R., Mohseni, S., Yang, F., Pentyala, S. K., Ragan, E. D., & Hu, X. B. (2021). How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters, 2*(4), e49. https://doi.org/10.1002/ail2.49

Linse, C., Alshazly, H., & Martinetz, T. (2022). A walk in the black-box: 3D visualization of large neural networks in virtual reality. *Neural Computing and Applications, 34*(23), 21237–21252. https://doi.org/10.1007/s00521-022-07608-4

Liu, H., Lai, V., & Tan, C. (2021). Understanding the Effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–45. https://doi.org/10.1145/3479552

Lukyanenko, R., Parsons, J., Wiersma, Y., Wachinger, G., Huber, B., & Meldt, R. (2017). Representing crowd knowledge: Guidelines for conceptual modeling of user-generated content. *Journal of the Association for Information Systems, 18*, 1–50. https://doi.org/10.17705/1jais.00456

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Lyytinen, K., Nickerson, J. V., & King, J. L. (2021). Metahuman systems = humans + machines that learn. *Journal of Information Technology, 36*(4), 427–445. https://doi.org/10.1177/0268396220915917

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian Conference on Information Systems*, *53*, 6–8. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b8eda9593fbcb63b7ced1866853d9622737533a2.

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based digital assistants. *Business & Information Systems Engineering, 61*(4), 535–544. https://doi.org/10.1007/s12599-019-00600-8

Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). ConvXAI: A system for multimodal interaction with any black-box explainer. *Cognitive Computation, 15*(2), 613–644. https://doi.org/10.1007/s12559-022-10067-7

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., & Mukherjee, A. (2020). Hate begets hate: A temporal study of hate speech. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 92:1–92:24. https://doi.org/10.1145/3415163.

Mayring, P. (2015). Qualitative Content Analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, spsampsps N. Presmeg (Eds.), *Approaches to Qualitative Research in Mathematics Education: Examples of Methodology and Methods* (pp. 365–380). Springer Netherlands. https://doi.org/10.1007/978-94-017-9181-6_13.

Mencar, C., spsampsps Alonso, J. M. (2019). Paving the way to explainable artificial intelligence with fuzzy modeling: Tutorial. In R. Fullér, S. Giove, spsampsps F. Masulli (Eds.), *Fuzzy Logic and Applications* (Vol. 11291, pp. 215–227). Springer International Publishing. https://doi.org/10.1007/978-3-030-12544-8_17.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors, 53*(4), 356–370. https://doi.org/10.1177/0018720811411912

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

Miller, L. M. S., & Bell, R. A. (2012). Online health information seeking: The influence of age, information trustworthiness, and search challenges. *Journal of Aging and Health, 24*(3), 525–541. https://doi.org/10.1177/0898264311428167

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, 1–66.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021a). A multidisciplinary survey and framework for design and evaluation of explainable

AI systems. *ACM Transactions on Interactive Intelligent Systems, 11*(3–4), 1–45. https://doi.org/10.1145/3387166

Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., Hu, X., Ji, S., & Ragan, E. (2021b). Machine learning explanations to prevent overtrust in fake news detection. *Proceedings of the International AAAI Conference on Web and Social Media, 15*, 421–431. https://doi.org/10.1609/icwsm.v15i1.18072

Mohseni, S., Ragan, E., & Hu, X. (2019). Open issues in combating fake news: Interpretability as an opportunity. *arXiv Preprint* arXiv:1904.03016. *https://arxiv.org/abs/1904.03016*

Murphy, Hannah (2023): Israel conflict lets loose a deluge of falsehoods on social media. In Financial Times, 2023. Available online at https://web.archive.org/web/20231121164908/https://www.ft.com/content/01650afb- dab4–4668-b16a-6add6ade0c04, checked on November 11th, 2023.

Nguyen, A., Kharosekar, A., Lease, M., & Wallace, B. (2018). An interpretable joint graphical model for fact-checking from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://ojs.aaai.org/index.php/AAAI/article/view/11487.

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3174223.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*. https://doi.org/10.1136/bmj.n71

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). *How model accuracy and explanation fidelity influence user trust*(arXiv:1907.12652). arXiv. http://arxiv.org/abs/1907.12652.

Park, J., Gu, J., & Kim, H. Y. (2022). "Do not deceive me anymore!" interpretation through model design and visualization for Instagram counterfeit seller account detection. *Computers in Human Behavior, 137*, Article 107418.

Park, H., Ahn, D., Hosanagar, K., & Lee, J. (2021). Human-AI interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445304.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Peffers, K., Rothenberger, M., Tuunanen, T., spsampsps Vaezi, R. (2012). Design science research evaluation. In K. Peffers, M. Rothenberger, spsampsps B. Kuechler (Eds.), *Design Science Research in Information Systems. Advances in Theory and Practice* (Vol. 7286, pp. 398–410). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-29863-9_29.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Peters, E., Diefenbach, M. A., Hess, T. M., & Västfjäll, D. (2008). Age differences in dual information-processing modes: Implications for cancer decision making. *Cancer, 113*(12 Suppl), 3556–3567. https://doi.org/10.1002/cncr.23944

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). *Manipulating and Measuring Model Interpretability* (arXiv:1802.07810; Issue arXiv:1802.07810). arXiv. http://arxiv.org/abs/1802.07810.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems, 32*(3), 229–267. https://doi.org/10.1080/07421222.2015.1099390

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Reeder, S., Jensen, J., spsampsps Ball, R. (2023). Evaluating Explainable AI (XAI) in Terms of User Gender and Educational Background. In H. Degen spsampsps S. Ntoa (Eds.), *Artificial Intelligence in HCI* (pp. 286–304). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35891-3_18.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *" Why should I trust you?" Explaining the predictions of any classifier* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rieger, L., & Hansen, L. K. (2020). *Aggregating explanation methods for stable and robust explainability*(arXiv:1903.00519). arXiv. http://arxiv.org/abs/1903.00519.

Rjoob, K., Bond, R., Finlay, D., McGilligan, V., Leslie, S. J., Rababah, A., Iftikhar, A., Guldenring, D., Knoery, C., McShane, A., spsampsps Peace, A. (2021). Towards explainable artificial intelligence and explanation user interfaces to open the 'black box' of automated ECG interpretation. In T. Reis, M. X. Bornschlegl, M. Angelini, spsampsps M. L. Hemmje (Eds.), *Advanced Visual Interfaces. Supporting Artificial Intelligence and Big Data Applications* (Vol. 12585, pp. 96–108). Springer International Publishing. https://doi.org/10.1007/978-3-030-68007-7_6.

Salako, A. A., Jiansu, P., Jinlun, Z., Guanqun, L., & Agbley, B. L. Y. (2021). Exsumm-VIZ: Visual interpretation of attention model in text summarizations. *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, 225–229. https://ieeexplore.ieee.org/abstract/document/9674053/.

Salthouse, T. A. (1992). Why do adult age differences increase with task complexity? *Developmental Psychology, 28*(5), 905–918. https://doi.org/10.1037/0012-1649.28.5.905

Salthouse, T. A. (1994). The nature of the influence of speed on adult age differences in cognition. *Developmental Psychology, 30*(2), 240–259. https://doi.org/10.1037/0012-1649.30.2.240

Salthouse, T. A., McGuthry, K. E., & Hambrick, D. Z. (1999). A framework for analyzing and interpreting differential aging patterns: Application to three measures of implicit learning. *Aging, Neuropsychology, and Cognition, 6*(1), 1–18. https://doi.org/10.1076/anec.6.1.1.789

Sanneman, L., & Shah, J. A. (2022). The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *International Journal of Human-Computer Interaction, 38*(18–20), 1772–1788. https://doi.org/10.1080/10447318.2022.2081282

Schemmer, M. (2022). *A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making*. https://doi.org/10.48550/arXiv.2205.05126

Schlagwein, D., & Hu, M. (2017). How and why organisations use social media: Five use types and their relation to absorptive capacity. *Journal of Information Technology, 32*(2), 194–209. https://doi.org/10.1057/jit.2016.7

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems, 29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Schmitt, V., Villa-Arenas, L.-F., Feldhus, N., Meyer, J., Spang, R. P., & Möller, S. (2024). The role of explainability in collaborative

human-AI disinformation detection. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2157–2174. https://doi.org/10.1145/3630106.3659031.r

Schreiber, A., spsampsps Bock, M. (2019). Visualization and exploration of deep learning networks in 3D and virtual reality. In C. Stephanidis (Ed.), *HCI International 2019—Posters* (Vol. 1033, pp. 206–211). Springer International Publishing. https://doi.org/10.1007/978-3-030-23528-4_29.

Schuller, B. W., Virtanen, T., Riveiro, M., Rizos, G., Han, J., Mesaros, A., & Drossos, K. (2021). Towards sonification in multimodal and user-friendlyexplainable artificial intelligence. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 788–792. https://doi.org/10.1145/3462244.3479879.

Schultze, S., Withöft, A., Abdenebaoui, L., & Boll, S. (2023). Explaining image aesthetics assessment: An interactive approach. *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 20–28. https://doi.org/10.1145/3591106.3592217.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T. H., Ding, K., Karami, M., & Liu, H. (2020). Combating disinformation in a social media age. *Wires Data Mining and Knowledge Discovery, 10*, 1–23. https://doi.org/10.1002/widm.1385

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*. https://doi.org/10.1145/3137597.3137600

Siering, M., Koch, J.-A., & Deokar, A. V. (2016). Detecting fraudulent behavior on crowdfunding platforms: The role of linguistic and content-based cues in static and dynamic contexts (SSRN Scholarly Paper 2866922). Social Science Research Network.

Singla, S., Eslami, M., Pollack, B., Wallace, S., & Batmanghelich, K. (2023). Explaining the black-box smoothly—A counterfactual approach. *Medical Image Analysis, 84*. https://doi.org/10.1016/j.media.2022.102721

Sonnenberg, C., spsampsps vom Brocke, J. (2012). Evaluation patterns for design science research artefacts. In M. Helfert spsampsps B. Donnellan (Eds.), *Practical Aspects of Design Science* (pp. 71–83). Springer. https://doi.org/10.1007/978-3-642-33681-2_7.

Stitini, O., Kaloun, S., & Bencharef, O. (2022). Towards the detection of fake news on social networks contributing to the improvement of trust and transparency in recommendation systems: Trends and challenges. *Information, 13*(3), 128. https://doi.org/10.3390/info13030128

Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication, 13*, 18.

Tamagnini, P., Krause, J., Dasgupta, A., & Bertini, E. (2017). Interpreting black-box classifiers using instance-level visual explanations. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, 1–6. https://doi.org/10.1145/3077257.3077260.

Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

Thuan, N. H., Drechsler, A., & Antunes, P. (2019). Construction of design science research questions. *Communications of the Association for Information Systems, 44*(1), 20. https://doi.org/10.17705/1CAIS.04420

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). Focus groups for artifact refinement and evaluation in design research. *Communications of the Association for Information Systems, 26*, 27. https://doi.org/10.17705/1CAIS.02627

Trinh, L., Tsang, M., Rambhatla, S., & Liu, Y. (2021). Interpretable and trustworthy deepfake detection via dynamic prototypes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1973–1983.

Truong, B. T., Lou, X., Flammini, A., & Menczer, F. (2024). Quantifying the vulnerabilities of the online public square to adversarial manipulation tactics. *PNAS Nexus, 3*(7), pgae258. https://doi.org/10.1093/pnasnexus/pgae258

Tsai, C.-H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and promoting diagnostic transparency and explainability in online symptom checkers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3411764.3445101.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204.

Vermeire, T., Brughmans, D., Goethals, S., De Oliveira, R. M. B., & Martens, D. (2022). Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications, 25*(2), 315–335. https://doi.org/10.1007/s10044-021-01055-y

Voigt, P., & Von dem Bussche, A. (2017). The EU general data protection regulation (gdpr). In *A practical guide* (1st ed., Vol. 10, No. 3152676, pp. 10–5555). Springer International Publishing. https://doi.org/10.1007/978-3-319-57959-7

Vraga, E. K., & Bode, L. (2020). Correction as a solution for health misinformation on social media. *American Journal of Public Health, 110*(S3), S278–S280. https://doi.org/10.2105/AJPH.2020.305916

Wang, X., Yuan, S., Zhang, H., Lewis, M., & Sycara, K. (2019). Verbal explanations for deep reinforcement learning neural networks with attention on extracted features. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 1–7. https://ieeexplore.ieee.org/abstract/document/8956301/.

Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets, 32*(4), 2079–2102. https://doi.org/10.1007/s12525-022-00593-5

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Wiriting a literature review. *MIS Quarterly, 26*(2), 13–23.

Wei, X., Zhang, Z., Zhang, M., Chen, W., & Zeng, D. D. (2019). Combining crowd and machine intelligence to detect false news on social media. *MIS Quarterly*.

Weinhardt, C., Fegert, J., Hinz, O., & van der Aalst, W. M. P. (2024). Digital democracy: A wake-up call. *Business & Information Systems Engineering, 66*(2), 127–134. https://doi.org/10.1007/s12599-024-00862-x

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). "Let me explain!": Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces, 15*(2), 87–98. https://doi.org/10.1007/s12193-020-00332-0

Wells, L., & Bednarz, T. (2021). Explainable AI and reinforcement learning—A systematic review of current approaches and trends. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.550030.

Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., Chung, Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2024). *Large language models can consistently generate high-quality content for election disinformation operations* (arXiv:2408.06731). arXiv. https://doi.org/10.48550/arXiv.2408.06731.

World Economic Forum (2024). Global Risks Report. Retrieved August 2024, from https://www.weforum.org/publications/global-risks-report-2024/.

Zahodne, L. B., Glymour, M. M., Sparks, C., Bontempo, D., Dixon, R. A., MacDonald, S. W. S., & Manly, J. J. (2011). Education does not slow cognitive decline with aging: 12-year evidence from the Victoria longitudinal study. *Journal of the International Neuropsychological Society, 17*(6), 1039–1046. https://doi.org/10.1017/S1355617711001044

Zhang, Z., Tian, R., Sherony, R., Domeyer, J., & Ding, Z. (2022). Attention-based interrelation modeling for explainable automated driving. *IEEE Transactions on Intelligent Vehicles, 8*(2), 1564–1573. https://doi.org/10.1109/TIV.2022.3229682

Zhu, H., Yu, C., spsampsps Cangelosi, A. (2022). Affective human-robot interaction with multimodal explanations. In F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, spsampsps S. S. Ge (Eds.), *Social Robotics*(Vol. 13817, pp. 241–252). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24667-8_22.

Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2021). Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics, 28*(1), 1161–1171. https://doi.org/10.48550/arXiv.2103.02071