Information Systems Frontiers (2023) 25:743–773
https://doi.org/10.1007/s10796-021-10234-5

Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection

Christian Meske[1] · **Enrico Bunde**[1]

Accepted: 13 December 2021 / Published online: 2 March 2022

Abstract
Hate speech in social media is an increasing problem that can negatively affect individuals and society as a whole. Moderators on social media platforms need to be technologically supported to detect problematic content and react accordingly. In this article, we develop and discuss the design principles that are best suited for creating efficient user interfaces for decision support systems that use artificial intelligence (AI) to assist human moderators. We qualitatively and quantitatively evaluated various design options over three design cycles with a total of 641 participants. Besides measuring perceived ease of use, perceived usefulness, and intention to use, we also conducted an experiment to prove the significant influence of AI explainability on end users' perceived cognitive efforts, perceived informativeness, mental model, and trustworthiness in AI. Finally, we tested the acquired design knowledge with software developers, who rated the reusability of the proposed design principles as high.

Keywords Design science research · Design principles · Hate speech detection · Explainable artificial intelligence · Local explanations

1 Introduction

Social media platforms connect users worldwide and allow them to exchange opinions on such topics as politics, finances, or social issues (Kapil & Ekbal, 2020; Shin et al., 2020). In this context, it is difficult to consistently enforce policies regarding undesirable content, such as *hate speech* (Matamoros-Fernández, 2017; Nienierza et al., 2019), that poses a potential risk of psychological harm for affected users (Ullmann & Tomalin, 2020). The developers of social media platforms usually rely on human moderators, who investigate and review potentially offensive content (Plaza-del-Arco et al., 2021; Ullmann & Tomalin, 2020). Recently, decision support approaches based on artificial intelligence (AI) have received much attention in relation to hate speech detection. For example, AI-based models can be used to detect different concepts of unwanted contents communicated through speech, such as hate speech, racism, or offensive language (Kapil & Ekbal, 2020). In addition to research projects that focus on using AI for hate speech detection, there are studies on decision support via software artifacts. For instance, such artifacts can help visualize aggressive comments on a user's timeline (Modha et al., 2020) or treat hate speech as malware by quarantining it and informing the targeted user (Ullmann & Tomalin, 2020).

In addition to researchers, institutions, and developers of social media platforms, large companies are also concerned with hate speech. For instance, Intel Corporation is developing an AI-based application for detecting and redacting audio material based on user preferences to filter hate speech and similar content, such as racism or sexism (Intel, 2021). However, modern AI models provide powerful predictions while being opaque and offering little transparency (Adadi & Berrada, 2018). This opaqueness characterizes many state-of-the-art AI models and is known as the black box problem (Adadi & Berrada, 2018; Kaplan & Haenlein, 2020). The black box problem represents the lack of explainability of the internal learning and decision-making processes of AI models, which is caused, for example, by a high complexity of underlying AI models (Arrieta et al., 2020; Meske et al., 2020). The research field of explainable AI (XAI) tackles the black box problem by introducing transparent models as well as techniques for generating different types of explanations

✉ Christian Meske
christian.meske@rub.de

[1] Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany

for black box models (Adadi & Berrada, 2018; Arrieta et al., 2020; Meske & Bunde, 2020). Consequently, modern AI-based decision support systems (DSSs) can provide powerful decision support while also explaining the outcome via user interfaces (UIs) (Lamy et al., 2019; van der Waa et al., 2021).

An effective UI design can integrate features to visualize or support the interaction with the inner decision-making and learning processes of an underlying algorithm, leading to an increased objective comprehension for the users (Cheng et al., 2019). Moreover, data-driven decisions may be useful to stakeholders, such as managers, who may rely on AI-provided explanations to understand the outcomes of various problems of interest (Martens & Provost, 2014). Furthermore, XAI can help monitor and ensure the fairness and transparency of AI-based systems, improve the management of such systems or support the maintenance of faulty systems (Kim et al., 2020; Meske et al., 2020; Tschandl et al., 2020). Despite active research in this context, there is a lack of user evaluation studies in the XAI field regarding the perception and effects of explanations on the targeted stakeholders (van der Waa et al., 2021). Moreover, different explanation goals and information needs, as well as varying backgrounds and/or expertise, can influence users' perceptions of XAI-based explanations, which further underlines the relevance of evaluations with targeted users (Barda et al., 2020; Meske et al., 2020; van der Waa et al., 2021). More specifically, we have identified two interconnected research gaps. On the one hand, there is a lack of applicable and generalizable UI design knowledge in the hate speech domain. The majority of DSS in the application domain of automated hate speech detection is evaluated based on technical metrics from the field of machine learning and do not involve users in the evaluation of the designed UIs (e.g., Modha et al., 2020; Paschalides et al., 2020; Pereira-Kohatsu et al., 2019). On the other hand, there is a lack of focus on users' and decision-makers' evaluations and perceptions of XAI-based explanations and their effects. To address these research gaps, we posed the following research questions:

1. *What are the essential design principles when designing XAI-based UIs to support moderators on social media platforms in detecting hateful content?*
2. *How are such UIs perceived by relevant stakeholders, and how influential are local explanations?*

To answer these research questions, we conducted a design science research (DSR) project with three consecutive design cycles, following the DSR process of Peffers et al. (2007). The proposed design principles (DPs) were evaluated qualitatively (interviews) as well as quantitatively (survey and experiment). Moreover, we evaluated the DPs in terms of reusability with the help of practitioners (i.e., software developers) (Ivari et al., 2018, 2021). We summarized the general requirements and general components into an explanatory design theory (EDT) that emphasizes general design features (DFs) and their effect on the environment (Baskerville & Pries-Heje, 2010; Baskerville et al., 2018; Gregor et al., 2020). These research activities were accompanied by an assessment of users' perceptions of the instantiated DPs and an investigation of local explanations' influence on the constructs perceived cognitive effort, perceived informativeness, mental model, and trustworthiness, which addresses the need to account for individual users' evaluations in the XAI research field (Meske et al., 2020; van der Waa et al., 2021).

The rest of the article is structured as follows: In the next section, we present the problem identification and motivation. Afterwards, we describe the DSR project. Then, we specify the adapted DRs as well as the derivation and justification of the DPs and DFs. This is followed by the demonstrations and evaluations of the three design cycles. In the subsequent section, we discuss the results, the theoretical implications, the limitations, and future research opportunities. The article ends with a conclusion.

2 Problem Identification and Motivation

2.1 Hate Speech on Social Media Platforms

Social media platforms play an integral role in the contemporary digitized world (Celik, 2019; Kapil & Ekbal, 2020; MacAvaney et al., 2019; Meske & Amojo, 2020). The data generated on these platforms enable data analytics and are valuable to companies, institutions, and individuals (Arapostathis, 2021; Shin et al., 2020; Vallejos et al., 2021). However, social media platforms also pose risks. For instance, scholars have highlighted the role of social media platforms in hate speech dissemination (Celik, 2019). Hate speech can harm individuals and societies and has been described as a threat to social media platforms themselves (Celik, 2019; Fortuna & Nunes, 2018; Kunst et al., 2021; Ullmann & Tomalin, 2020). The United Nations (2019) defines hate speech as follows: "[…] any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language concerning a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor […]" (p. 2).

Such content is not rare on social media platforms and in the digital sphere in general. A survey in the US found that 37% of the pupils surveyed (between the ages of 12 and 17 years) had experienced hate speech, which affected them personally (Hinduja & Patchin, 2019). Being frequently confronted with hate speech can seriously affect users

emotionally and behaviorally (Bilewicz & Soral, 2020). Moreover, hate speech was found to result in emotional damage or even suicide among young people (Kaplan & Haenlein, 2020; Smith et al., 2008). Many users of social media platforms have reported experiencing and being affected by hate due to their ethnicity, religion, politics, or gender (Celik, 2019). Nevertheless, users can also intervene and help reduce hateful content by reporting it or even engaging in counter speech to fight disruptive behavior (Kunst et al., 2021). To combat hate speech, developers of social media platforms have established policies that enable them "[…] to delete comments, such as hate comments, that do not follow these guidelines" (Wilhelm et al., 2020, p. 924). The next section discusses the existing literature on DSSs for hate speech detection.

2.2 Decision Support Systems for Hate Speech Detection on Social Media Platforms

AI-based hate speech detection is receiving a lot of attention in research, a development that is also reflected in the scientific literature (e.g., Ayo et al., 2020; Fortuna & Nunes, 2018; MacAvaney et al., 2019). However, despite the great interest in automated detection of hate speech, there is scant research on the design of DSSs for supporting both end users and human moderators. Scholars have mainly addressed this problem by focusing on end users (i.e., social media users)—for example, Modha et al. (2020) trialed a software artifact based on deep learning techniques in the form of a web-browser plugin that visualizes different nuances of aggressiveness on a user's timeline. This plugin functions primarily as decision support for end users, as human moderators would still have to screen the content manually because the visualizations are displayed along with user-generated content (Modha et al., 2020; Plaza-del-Arco et al., 2021; Ullmann & Tomalin, 2020). Ultimately, the web-browser plugin was evaluated only based on technical metrics from the field of machine learning-based AI (Modha et al., 2020), whereas no evaluation of the design was performed with end users. Thus it is unclear, how the design affect or is perceived by end users. Additionally, the end user may be confronted with the classification results without being able to understand or comprehend why the corresponding prediction was made.

Using big data approaches, Paschalides et al. (2020) developed MANDOLA, a system for monitoring, detecting, visualizing, and reporting the spread of hateful content online. MANDOLA offers visualizations to present detected hate speech to users, with filters based on time, context, and location that allow users to identify, for example, correlations between the development of hate speech and its potential triggers (e.g., events). Despite the interesting approach of this system, it is not suitable for the moderation of individual social platforms. Moreover, MANDOLA is evaluated exclusively based on technical metrics from the field of machine learning-based AI (Paschalides et al., 2020). The designed UIs are not evaluated with end users, thus lacking knowledge about the perception of the design. Another system is HaterNet, which is used by the Spanish National Office Against Hate Crimes at the Spanish State Secretariat for Security to detect and monitor hate speech on Twitter (Pereira-Kohatsu et al., 2019). As with the previous examples, HaterNet was evaluated based solely on technical metrics from the field of machine learning (Pereira-Kohatsu et al., 2019), leaving it unclear how the UI is perceived by end users. Lastly, the quarantining framework proposed by Ullmann and Tomalin (2020) is another approach to protecting individuals in online social spheres. This approach starts with the identification of harmful content, which is then temporarily quarantined; then, an alert is sent to the intended recipient to protect them from harmful content, such as hate speech. Nevertheless, the quarantining framework was not evaluated with end users, leading to missing insights on the perception of end users (Ullmann & Tomalin, 2020). The impression, that the design of automated hate speech detection systems and the involvement of end users to evaluate the design is underrepresented, is also confirmed in surveys about automated hate speech detection, since the beforementioned aspects are not adequately represented (e.g., Ayo et al., 2020; Fortuna & Nunes, 2018).

In general, there have been highly innovative attempts to develop and deploy DSSs to aid hate speech detection. However, most of this work (i) has focused on decision support for end users, (ii) does not provide applicable and reusable prescriptive design knowledge, and (iii) does not adequately involve users in the evaluation of such information systems and their design. In the next section, we provide an overview of AI technology in relation to hate speech detection.

2.3 Artificial Intelligence for Hate Speech Detection on Social Media Platforms

AI has received enormous attention in research (Kaplan & Haenlein, 2020). We understand AI as machine-learning-based systems with the "[…] ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 17). In our study, we focused on text-based hate speech detection, which is frequently addressed using AI-based models (Fortuna & Nunes, 2018; Kapil & Ekbal, 2020; MacAvaney et al., 2019; Modha et al., 2020; Plaza-del-Arco et al., 2021). Such models can be applied as a single or a hybrid method (Ayo et al., 2020). Single methods are represented by models such as logistic regression, an example of machine learning, whereas convolutional neural networks represent a more complex deep

learning models (Fortuna & Nunes, 2018; MacAvaney et al., 2019; Plaza-del-Arco et al., 2021). Hybrid methods combine different machine learning or deep learning models for a classification problem such as text classification (Plaza-del-Arco et al., 2021). When it comes to AI-based hate speech detection, besides AI models, integrated data features are relevant as well. For example, the integration of users' psychological features into the input features for the underlying machine learning model enables the detection of concepts related to hate speech, such as cyberbullying (Balakrishnan et al., 2020).

Plaza-del-Arco et al. (2021) investigated multilingual as well as monolingual pre-trained language models and compared them with machine learning models. According to their results, transfer learning outperformed the other models. Most of the AI-based hate speech detection systems are black boxes. First attempts have been presented to solve the black box problem of state-of-the-art AI models (Arrieta et al., 2020; Ayo et al., 2020). The HaterNet system, which we described before, illustrates relevant terms, receivers, and emitters within identified hate speech texts. Yet these explanatory features are again not evaluated with end users leaving the affect as well as usefulness for end users open (Pereira-Kohatsu et al., 2019). Another example is provided by MacAvaney et al. (2019) who applied the transfer learning model Bidirectional Encoder Representations from Transformers (BERT) and used the self-attention weights of the model to evaluate the informativeness of relevant words for the classification outcome. Moreover, the visualization of hate speech for end users within social media is described as an emerging area, with proposed systems being very limited (Modha et al., 2020). We further argue that explanations are valuable in the context of automated hate speech detection for example to identify biased algorithms, false classifications, or comprehend and validate the classification outcome to initiate appropriate actions such as deleting hateful comments (Adadi & Berrada, 2018; Arrieta et al., 2020; Meske et al., 2020; Wilhelm et al., 2020).

Current scientific literature, including the above-described contributions, neglect the perspective of human moderators on social media platforms, generalizable design knowledge, and the perception of the targeted stakeholders towards the design as well as explanatory features. Hate speech detection is frequently investigated from a technical perspective—for example, by proposing new AI-based models (e.g., Ayo et al., 2020; Fortuna & Nunes, 2018; MacAvaney et al., 2019). As discussed in the previous section, AI models are also being increasingly integrated into DSSs when it comes to hate speech detection. However, many studies only perform technical evaluations and do not involve real users (e.g., Modha et al., 2020; Paschalides et al., 2020; Pereira-Kohatsu et al., 2019; Ullmann & Tomalin, 2020). In the following section, we discuss the subject of XAI, local explanations, and their importance for the design of UIs.

2.4 Explainable Artificial Intelligence and Local Explanations for User Interfaces

AI is becoming increasingly complex and powerful (Kaplan & Haenlein, 2020). However, AI developments are accompanied by challenges, such as the black box problem, which refers to the tradeoff between complexity-based performance gains and decreasing explainability of AI models internal learning as well as decision-making processes (Adadi & Berrada, 2018; Arrieta et al., 2020; Kaplan & Haenlein, 2020; Meske & Bunde, 2020). Consequently, XAI research has attempted to tackle this problem by introducing techniques for developing explainable high-performance models to enable humans to understand, trust, and manage AI-based systems (Adadi & Berrada, 2018; Arrieta et al., 2020). Various motivations can drive XAI's integration into DSSs, such as using XAI for management, justification, and improvement of AI-based systems or to control them (Adadi & Berrada, 2018; Meske & Bunde, 2020). Scholars generally distinguish two dimensions of interpretability. *Global* interpretability enables users to understand the whole logic of a model by following the reasoning that leads to different possible outcomes, whereas *local* interpretability describes the capability of explaining the reasons behind a specific outcome (Adadi & Berrada, 2018; van der Waa et al., 2021). In our study, we focused on local interpretability because it is difficult for end users (in our case, decision-makers) to understand the overall mechanism of the whole machine learning model. Therefore, we employed post-hoc explainability techniques and generated local explanations that "[…] tackle explainability by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model" (Arrieta et al., 2020, p. 88).

Techniques from the XAI field can be integrated into explanation interfaces (i.e., UIs), which are an essential component for aiding users in their tasks (Gunning & Aha, 2019). Focusing on visual explanations in the context of visual case-based reasoning for breast cancer detection, Lamy et al. (2019) developed a UI for medical experts that the experts found interesting. Another UI was developed by Cheng et al. (2019) to support end users in understanding the algorithms for making university-admission decisions; the UI was found to improve the users' comprehension of the underlying algorithm. Barda et al. (2020) developed an explanatory display for predictions based on a pediatric intensive care unit in-hospital mortality risk model, the users found the display useful.

In general, there have been active efforts to develop UIs in the XAI context in different domains based on various

approaches and perspectives (Gunning & Aha, 2019). However, researchers have described how stakeholders in different organizational roles or working at different knowledge levels can have different explanation goals and information needs, which can be further affected by different backgrounds in terms of training, experience, or demographic characteristics (Barda et al., 2020; Meske et al., 2020; Motorny et al., 2021). Moreover, scholars have pointed out that the XAI field does not focus enough on user evaluations (van der Waa et al., 2021). By developing applicable design knowledge for the UIs of DSSs used for hate speech detection, we contribute useful design knowledge that has been evaluated by relevant stakeholders (Barda et al., 2020) and which is complemented by user evaluations, thus providing insights into the effects of local explanations (van der Waa et al., 2021). In the next section, we provide an overview of our DSR project and methodology.

3 Design Science Research Project

3.1 The Design Science Research Process

In our study, we developed practical DPs for UIs in AI-based DSSs for human moderators of social media platforms based on the DSR methodology. DSR enables scholars to create knowledge that is transferable to real-world scenarios (vom Brocke et al., 2020; Gregor & Hevner, 2013; Gregor et al., 2020). DPs also represent nascent design theories, or knowledge as operational principles (Gregor & Hevner, 2013). The DPs were instantiated in UIs with varying degrees of maturity, which were evaluated qualitatively and quantitatively in three consecutive design cycles. In all evaluations, we provided introductory materials on AI-based DSSs for hate speech detection. During the experiments, participants had to consider multiple exemplary hate speech cases and then fill out a survey. For practitioners' evaluation of the DPs, we did not provide an example of the UI. We uncovered potential for optimization, measured the positive perception of the UI by users, examined the influence of local explanations on them, and the reusability of DPs with practitioners.

We have followed the DSR methodology of Peffers et al. (2007). The first activity, (i) problem identification and motivation, was discussed in the previous sections. The subsequent activities—(ii) defining the objectives of the proposed solution, (iii) design and development, (iv) demonstration, (v) evaluation, and (vi) communication of our results—will be presented in the following sections. Figure 1 provides an overview of the DSR project and is followed by a summary of the individual design cycles.

The *first design cycle* began with a literature review. We discovered that existing research has not adequately addressed human moderators and their role in the context of AI-based hate speech detection. Moreover, we discovered a lack of prescriptive design knowledge for UIs in the domain of hate speech detection. We then identified generic requirements for DSSs, which we adopted in our own work (Meth et al., 2015). We developed the appropriate DPs, DFs and implemented the transfer learning model known as Universal Language Model Fine-Tuning (ULMFiT) (Howard & Gugger, 2020; Howard & Ruder, 2018). Using ULMFiT, we generated predictions and local explanations for the UI. The initial design was implemented as a static UI. The design was evaluated qualitatively by 11 participants who had experience as moderators on social media platforms. We analyzed the resulting data using thematic analysis (Braun & Clarke, 2006) and operationalized the insights as input knowledge for the second design cycle.

The *second design cycle* started with an investigation of DSSs for natural language processing tasks and involved further desk research on UI design. We refined the DRs, DPs, and DFs based on the insights from the first evaluation. The revised design was implemented as an interactive UI and was quantitatively evaluated by means of an experiment with 190 participants recruited via CloudResearch (Litman et al., 2017) and Amazon Mechanical Turk (MTurk). Using

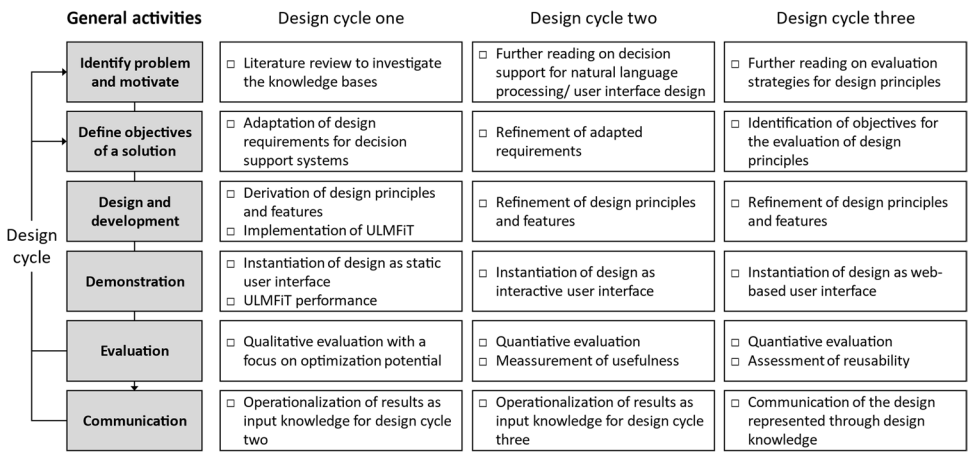


Fig. 1 The design science research process adapted from Peffers et al. (2007)

Adobe XD, the prototype was integrated into the web survey. The overarching goal of this evaluation was to assess users' perceived usefulness (Davis, 1989; Greven et al., 2003), perceived ease of use (Davis, 1989; Greven et al., 2003), and intention to use (Venkatesh et al., 2003), which allowed us to evaluate the artifact's valuable utility (Gregor & Hevner, 2013; Venable et al., 2016). For this and the following experiments, we set the following admission criteria: (i) participation was only possible from the United States and the European Union's member states, (ii) participants had to have experience as moderator on social media platforms (in the evaluation with practitioners, this requirement was replaced by experience as software developers), and (iii) participants had to pass CloudResearchs' attention and engagement checks. Moreover, participants had the opportunity to provide feedback in open text fields. We analyzed the resulting data using thematic analysis (Braun & Clarke, 2006). The obtained insights were operationalized as input knowledge for the third design cycle.

We initiated the *third design cycle* by investigating strategies for evaluating the DPs. The artifact was refined and implemented in a production-ready environment in the form of a web-based UI. The overarching goals were twofold: (i) to assess the impact of explainability (local explanations) on the constructs perceived cognitive effort (Wang & Benbasat, 2009), perceived informativeness (Zhang et al., 2014), mental model (Vitharana et al., 2016), and trustworthiness (Carter & Bélanger, 2005) by means of an experiment with 360 participants; and (ii) to evaluate the reusability of the DPs (Ivari et al., 2018, 2021) by consulting 80 practitioners. In sum, we assessed both the quality of the implemented design from users' perspectives and how well prescriptive statements help practitioners to develop corresponding artifacts in practice (Gregor et al., 2020; Ivari et al., 2018, 2021).

3.2 Hate Speech Detection Using Transfer Learning and Artifact Development

Regarding the transfer learning model, we used ULMFiT (Howard & Gugger, 2020; Howard & Ruder, 2018). We used the Google CoLab environment, Python FastAI library for ULMFiT, ULMFit's interpretation module, and scikit-learn to generate the performance metrics (FastAI, 2021; Pedregosa et al., 2011; Howard & Gugger 2020). All implementations were done using Python 3. We strictly followed FastAI's documentation during the implementation, the fine-tuning process, and the generation of the local explanations (FastAI, 2021). The dataset for hate speech detection was identified in MacAvaney et al. (2019) and is publicly accessible on Kaggle (Kaggle, 2012). The dataset comprised 3,947 samples and consisted of the following two classes: hate speech (1,049 samples) and no hate speech (2,898 samples). We used 80% (3,157 samples) of the data for training and fine-tuning the ULMFiT model and 20% (790 samples) for the final test. Local explanations were generated using ULMFiT's interpretation module (FastAI, 2021). The UIs for the first and second design cycles were developed with Adobe XD, a vector-based graphics software. The last UI was implemented as a web-based prototype using Python Django, CSS-Bootstrap, and JavaScript.

4 Objectives of the Proposed Solution: Adaptation and Justification of Design Requirements

DRs represent the goodness criteria, which should consist of a rich mix of goals from different categories, such as technology, information quality, or human interaction (vom Brocke et al., 2020). Moreover, DRs are part of the problem space, aid in the evaluation of the designed solutions, and are an integral component in EDTs that aim to explain how general design components address general requirements (Baskerville & Pries-Heje, 2010; vom Brocke et al., 2020; Venable et al., 2016). To adapt and justify the developed DRs, we used descriptive and, especially prescriptive knowledge (Gregor & Hevner, 2013; vom Brocke et al., 2020; Hevner, 2020). In scientific literature, knowledge that was contributed via prior research projects and is used in a new DSR project is also denoted as input knowledge (vom Brocke & Maedche, 2019). For DSSs to aid human decision-makers' various goals, we identified the following three DRs as input knowledge (Meth et al., 2015): (i) increase decision quality by providing high-quality advice, (ii) reduce human decision-maker's cognitive effort by providing decision support, and (iii) minimize system restrictiveness by allowing users to control strategy selection.

We transferred these generic DRs into our application domain (i.e., automated hate speech detection for human moderators) to establish an anchor in this domain as well as in the associated knowledge bases. For example, the generic design requirements from Meth et al. (2015) are part of a design theory that the authors developed in their DSR project and are therefore part of the prescriptive knowledge base on which we build on (vom Brocke et al., 2020; Hevner, 2020). Another component of the prescriptive knowledge base that we rely on are design entities (vom Brocke et al., 2020; Hevner, 2020). The prescriptive knowledge base in our DSR project consists of contributions on AI-based hate speech detection with transfer learning (e.g., Ayo et al., 2020; Plaza-del-Arco et al., 2021), XAI-based UI design (e.g., Barda et al., 2020; Cheng et al., 2019), and prior research on information systems for hate speech detection (e.g., Modha et al., 2020; Paschalides et al., 2020). By investigating the before described knowledge bases and operationalizing the

insights as well as contributions, we aimed to adequately transfer the generic design requirements into our application domain. Consequently, we consumed existing knowledge that informed the proposed design knowledge (vom Brocke & Maedche, 2019; Gregor & Hevner, 2013; Hevner, 2020). The discussed DRs were refined based on the insights gained during the evaluations of the three design cycles. Figure 2 provides an overview of the relation between the generic DRs and the DRs for our application domain. To provide a comprehensive overview of the design knowledge up front, we decided to present the final set of DRs right away and illustrate how the DRs evolved over the course of our DSR journey.

AI-based systems can surpass the performance of human experts and can help users make better decisions (Kaplan & Haenlein, 2020; Tschandl et al., 2019). Transfer learning models are part of state-of-the-art AI models and are applicable in hate speech detection (Ayo et al., 2020; Howard & Ruder, 2018; Kim et al., 2020; Peng et al., 2020; Plaza-del-Arco et al., 2021). The transfer learning models such as ULMFiT offer significant benefits to researchers and practitioners (FastAI, 2021; Howard & Gugger, 2020), as these approaches have the potential to save computing power, time and require less data for training (i.e., fine-tuning) while reaching high performance levels (Cadavid et al., 2020; Howard & Gugger, 2020; Howard & Ruder, 2018; Peng et al., 2020). Therefore, we established:

DR1. When users need to identify hate speech, the system should use transfer learning for text classification to provide high-quality advice.

Explanations can be relevant to users if they want to control a DSS's decision-making process or detect possible biases and can be integrated in UIs (Arrieta et al., 2020; Barda et al., 2020; Caliskan et al., 2017; Kaplan & Haenlein, 2019). When XAI techniques are used adequately—for example, in UIs—users' understanding of a system's output can be improved (Barda et al., 2020; Cheng et al., 2019). Furthermore, adequate XAI use can lead to users achieving a high level of trust in a DSS (Adadi & Berrada, 2018; Arrieta et al., 2020; van der Waa et al., 2021). Scholars have emphasized the need for more transparent decision-making processes to construct well-designed tools for decision support (Li & Gregor, 2011). Moreover, XAI provides opportunities for improving underlying AI-systems—for instance, with the goal of identifying errors, faults, or biases in such systems (Adadi & Berrada, 2018; Arrieta et al., 2020). Based on these arguments, we established:

DR2. When users are provided with automated hate speech detection, the system should provide a local explanation for the present case to enable users to interpret the outcome.

Decision-makers' need for information that assists them in their decision-making processes has been clearly identified in the existing literature (e.g., Gönül et al., 2006; Kaplan & Haenlein, 2019, 2020; Li & Gregor, 2011). Therefore, the following DR aims to provide appropriate information and knowledge, which are integral parts of decision-making (Zack, 2007). This goal can be achieved through basic user features, such as historical messages (Fortuna & Nunes, 2018). Therefore, we established:

DR3. When investigating a case of potential hate speech, the system should provide the user with relevant information about the accused user to get a holistic view of that user's behavior.

A DSS should aid decision-making instead of ruling over the user and making autonomous decisions (Akata et al., 2020; Kaplan & Haenlein, 2019). An important element of XAI's overall purpose is to keep humans in the loop (Adadi & Berrada, 2018). This element offers various benefits—for instance, human decision-makers can spot DSSs' biases or faults (Adadi & Berrad, 2018; Arrieta et al., 2020; Meske et al., 2020). Hybrid intelligence, which refers to a "[…] combination of human and machine intelligence, augmenting human intellect and capabilities instead of replacing them, to make meaningful decisions, perform appropriate actions […]" (Akata et al., 2020, p. 20), is a related concept. Consequently, we established:

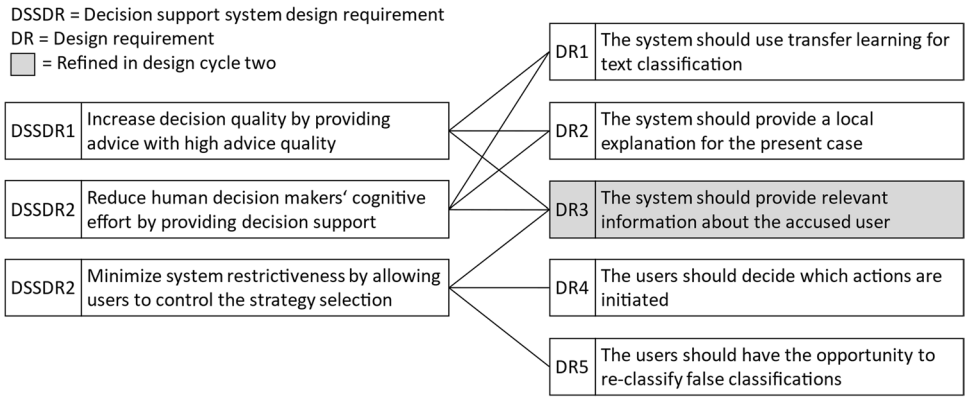


Fig. 2 The transfer of the generic DRs for DSSs (Meth et al., 2015) into our application domain

DR4: When investigating a case, users should decide which actions are initiated to retain the power of decision-making.

It should be possible for users to correct false classifications—for instance, by re-classifying them. By retaining decision-making power, the human moderator evolves into an empathic data-driven decision-maker (Kaplan & Haenlein, 2019, 2020). This retainment of decision-making power also refers to the human in the loop concept (Adadi & Berrada, 2018). An interactive machine learning loop emerges, for instance, when humans generate new data based on reclassified examples, which, in turn, can be used for fine-tuning the underlying AI (Howard & Ruder, 2018; Ramos et al., 2020). Based on these examples, we established:

DR5. When encountering false classifications by the system, users should have the opportunity to re-classify such cases to initiate a feedback loop for fine-tuning the system's performance.

The next section provides information on our design and development.

5 Design and Development

5.1 Derivation and Justification of Design Principles

In this section, we explain how we derived the DPs that were translated into DFs. Both DPs and DFs are part of the solution space and have to address the DR in the problem space (vom Brocke et al., 2020). DPs are used to communicate design knowledge in an accessible format (Gregor et al., 2020). They are translated into specific DFs, which can be implemented in a prototype artifact (Seidel et al., 2018). Moreover, DPs and DFs are part of the general components of an EDT that explains how specific DRs can be met (Baskerville & Pries-Heje, 2010).

The first DPs focuses on transfer learning as an established approach to text classification (Kim et al., 2020). Social media platforms are repositories of textual data and versatile sources of information (Hu et al., 2019). The amount of available data on such platforms can vary (Stieglitz et al., 2018). Transfer learning can reach state-of-the-art performance levels while requiring as few as 100 labeled examples (Howard & Ruder, 2018). This is important as we encounter challenges specific to social media, such as the acquisition of sufficient data for training AI models (Modha et al., 2020). In addition, the combination of human intelligence with approaches from the AI field can lead to meaningful decisions (Akata et al., 2020). Therefore, we established:

DP1. Provide the system with transfer learning techniques for classifying unstructured data so that users can make decisions based on the provided decision support, given that users expect high-quality advice.

XAI techniques can be used to address the emerging black box problem that characterizes modern AI models (Adadi & Berrada, 2018; Arrieta et al., 2020; Gunning & Aha, 2019; Meske et al., 2020). We used XAI for justifying individual outcomes and for detecting potential errors or biases (Adadi & Berrada, 2018; Gupta et al., 2021). Research has shown that explainability features in DSSs can positively affect the users' satisfaction with the decision-making process (Li & Gregor, 2011). In addition, explanations can affect the users' acceptance of a DSS (Gönül et al., 2006) and can improve users' comprehension of the underlying algorithm (Cheng et al., 2019). Moreover, humans may expect a DSS to provide explanations for outcomes (Adadi & Berrada, 2018; van der Waa et al., 2021). Consequently, we established:

DP2. Provide the system with features based on XAI to generate suitable explanations so that users can interpret and comprehend the provided decision support, given that users want to trust and validate received advice.

Research has highlighted that social media data are versatile and can be used for different analytic purposes (Cheng et al., 2019; Hu et al., 2019). Through social media analytics, we can gain useful knowledge on, for example, who creates content or who is an influential driver of communication (Stieglitz et al., 2018). Research with a focus on DSSs has shown that when decision-makers are provided with advice, information, or explanations, they often attempt to use all the different sources of knowledge available to them (Gönül et al., 2006). In addition, a DSS should allow users to dynamically interact with, explore, or manipulate the provided data (Park et al., 2016). To support decision-making, the provided information as well as the interactions with data should be intuitive and flexible for users (Jimenez-Marquez et al., 2019; Li & Kettinger, 2021). Therefore, we established:

DP3. Provide the system with the capability to present relevant, case-based contextual information so that users can develop a holistic understanding of the current case, given that users want to initiate appropriate and informed decisions.

Human moderators must act quickly to minimize potential psychological harm; social media analytics can support their decision-making processes (Stieglitz et al., 2018; Ullmann & Tomalin, 2020). We aimed to enable decision-makers to take proper actions based on the information and visualizations provided (Jimenez-Marquez et al., 2019). Moreover, scholars have provided empirical evidence that to improve decision-making, decision aids should not be too restrictive (Wang & Benbasat, 2009). Social media platform developers establish guidelines and can enable protective actions, such as deleting comments, which can be carried out by

human moderators (Ullmann & Tomalin, 2020; Wilhelm et al., 2020). Consequently, we established:

DP4. Provide the system with capabilities that enable users to initiate case-related actions so that users can incorporate their own social and cultural knowledge, given that users want to make fair and accountable decisions.

Figure 3 summarizes the DPs and DRs as well as their relationships. These DPs were based on state-of-the-art knowledge bases and were refined using insights gained during the evaluations in the three design cycles. In the next subsection, we present the DFs.

5.2 Derivation and Justification of Design Features

In this sub-section, we describe the DFs that we implemented in the prototype artifact (Seidel et al., 2018). We identified transfer learning as an appropriate approach to hate speech detection (e.g., Kunst et al., 2021; Modha et al., 2020) and underlined the importance of clearly communicating AI-suggested outcomes via the UI (Barda et al., 2020; Lamy et al., 2019; Schneider et al., 2020) (DF1: Provide the outcome of the classification). A key aspect of DSSs' design is to communicate the quality of the provided decision support (Gönül et al., 2006) (DF2: Provide the confidence for the classification). We used case-specific local explanations that had already been investigated by previous studies (Li & Gregor, 2011; van der Waa et al., 2021) (DF3: Provide local explanations for specific cases). To provide relevant case-related information, we used data visualization in the UI, which scholars consider to be an influential feature of UIs for DSSs (Park et al., 2016) (DF4: Provide the offending users' history in relation to hate speech). Contextual variables can be relevant when it comes to explainable decision support (Adadi & Berrada, 2018) (DF5: Provide the history of previously initiated actions against the offending user). Another important DSS aspect is users' sense of control and the opportunity to review different cases, which enables users to evaluate larger amounts of information more rapidly (Huang, 2003; Li & Gregor, 2011) (DF6: Provide an overview of all current cases). Final decision-making power should lie with the human moderator. We also addressed the challenge of responsibility, as the human moderator was left responsible for the initiated actions (Kaplan & Haenlein, 2020). Moreover, users on social media platforms must comply with the guidelines of the platform, otherwise moderators can initiate protective actions (Wilhelm et al., 2020) (DF7: Provide a feature to initiate case-based actions; DF8: Provide a feature to contact the offending user). The option for re-classification was implemented to incorporate concepts such as hybrid intelligence and human in the loop (Adadi & Berrada, 2018; Akata et al., 2020; Arrieta et al., 2020) (DF9: Provide an option for re-classification). Lastly, it is important for users to be able to navigate and manipulate the UI (Huang, 2003; Park et al., 2016) (DF10: Provide navigation).

Figure 4 provides an overview of our DFs and DPs. The DFs were based on state-of-the-art knowledge bases and were refined using insights gained during the evaluations in the three design cycles. In the next section, we discuss the artifact's demonstration and evaluation across the three design cycles.

6 Demonstration and Evaluation

6.1 Demonstration and Evaluation: First Design Cycle

During the first design cycle, we implemented ULMFiT and fine-tuned our model to hate speech detection. The model reached an accuracy of 86.46% on the test data (790 samples). Appendix 1 provides an overview of further performance metrics (precision, recall, and f1-score). Figure 5 provides a confusion matrix and shows that 140 examples of hate speech and 543 examples of no hate speech were classified correctly. In addition, 79 examples of no hate were classified as hate speech and 28 examples of hate speech as no hate speech. Despite the dataset's relatively low sample size, our model exhibited solid performance, which is one of the strengths of transfer learning (Howard & Gugger, 2020; Howard & Ruder, 2018). We used this model to generate the classifications, confidence values, and local explanations for

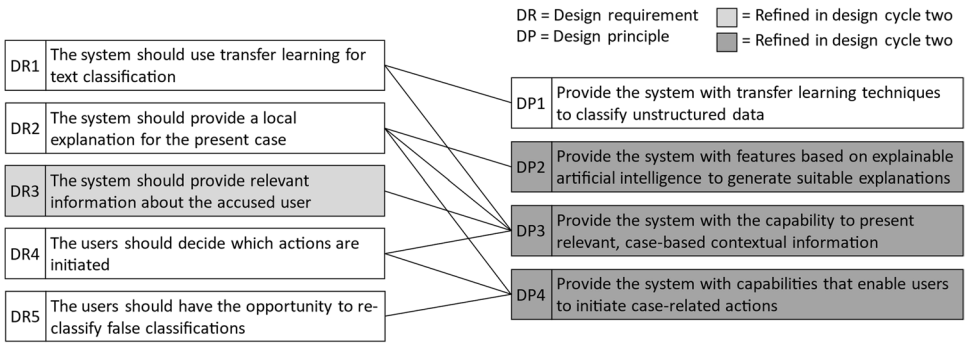


Fig. 3 Overview of the DPs and their relationships with the DRs

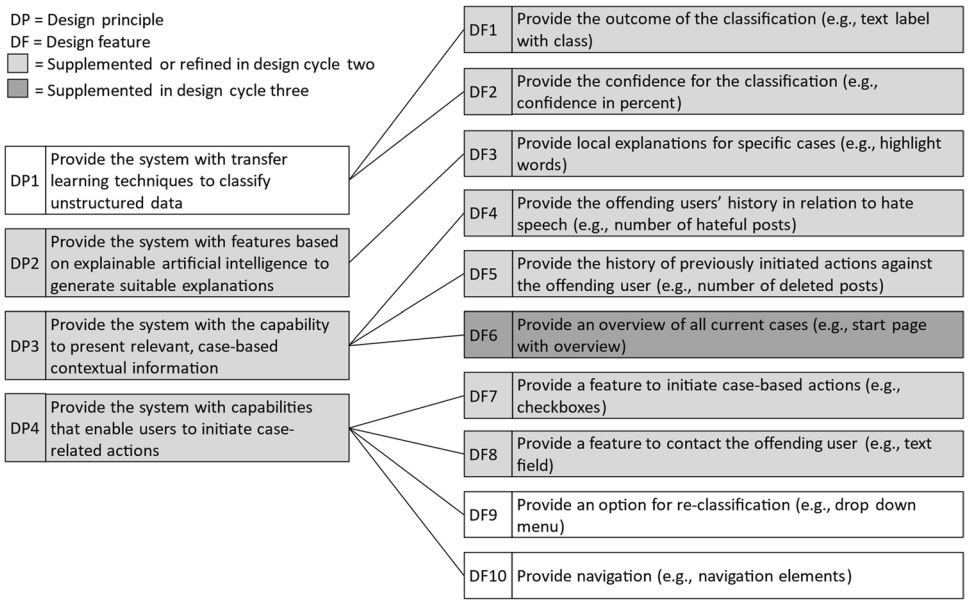


Fig. 4 Overview of the DFs and their relationships with the DPs

the UI. The model served as a baseline for the artifacts and remained unchanged.

Figure 6 explains the UI and the implemented design. We prepared identical UIs for five different cases. The hate speech example with the hate speech confidence barometer and the confidence of the AI classification addressed DP1. To address DP2, we combined different techniques to consider users' different information and explanation needs (Barda et al., 2020; Meske et al., 2020). We highlighted the most important words for the classification and added a bar chart with the most important features and their weighting. We provided different graphical illustrations that summarized the most relevant information regarding the case and the offending user as contextual variables (Adadi & Berrada, 2018)—for instance, we provided a pie chart for the distribution of hate speech by the particular offending user.

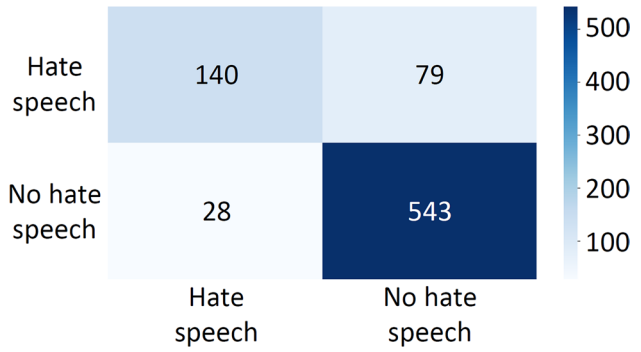


Fig. 5 Confusion matrix for the ULMFiT model based on test data

These DFs addressed DP3. Lastly, we implemented DP4 by providing checkboxes for initiating actions and the possibility of re-classifying specific cases.

For the first evaluation, we conducted semi-structured interviews. The participants were recruited from a university environment. Interview participants recommended further interviewees, which led to a snowball sampling approach (Patton, 2014). The participants for this evaluation were eight men and three women aged 19–31 years ($M = 26.2$, $SD = 3.3$). The participants had 0.5–6.5 years of experience as moderators on social media platforms ($M = 2.5$, $SD = 0.9$). In addition, eight of the 11 participants stated that they had been affected by hateful content. The interviews were conducted virtually via Skype and lasted between 14 and 23 min ($M = 19.1$, $SD = 2.6$). The interviews were recorded and transcribed, and the data were analyzed using thematic analysis (Braun & Clarke, 2006). The overarching goal was to assess users' perceptions and to identify optimization potential. The participants expressed positive sentiments toward the design: "The interface looks easy to use" (Interviewee 4); "[…] such an application could enhance my productivity […]" (Interviewee 6); or "[…] I would like to work with such intelligent systems […] (Interviewee 10). Table 1 contains illustrative quotations that represent recurring themes throughout the interviews and descriptions of future optimizations based on users' responses.

With these insights, we concluded the first design cycle and used the optimization potential as input knowledge for the second design cycle, which is described in the next section.

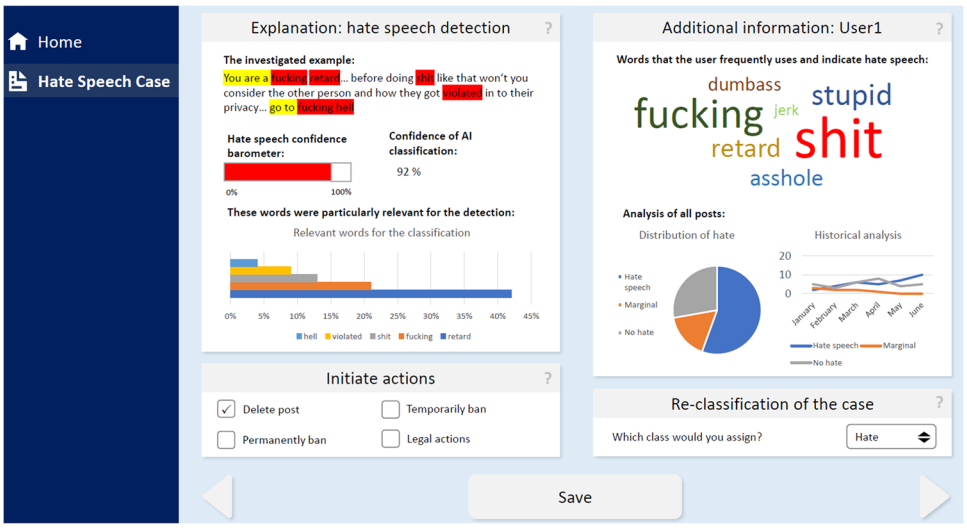


Fig. 6 Implementation of the initial design during the first design cycle

6.2 Demonstration and Evaluation: Second Design Cycle

Based on the insights gathered during the first design cycle, we revised the design and the UI. During the second design cycle, we supplemented the design with interactive capabilities, such as the selection of checkboxes, mouse-over effects for the charts, and navigation between cases. Besides these changes, we removed the two charts that the users had described as redundant, namely the feature importance bar chart and the word cloud. Instead of using the hate speech barometer, we chose to clearly communicate the AI-based classification outcome (i.e., the class, hate speech or no hate speech). We also revised the actions that the human moderators could initiate, eliminating the "legal actions" option. Lastly, we added historical information on the actions that had been initiated against a specific user and the option to contact the offending user directly. Figure 7 depicts the refined UI used for the second evaluation.

The refined prototype was evaluated by means of an experiment whose participants were recruited via CloudResearch and MTurk. Appendix 2 provides a summary of the demographic data for the 190 participants. We used a set of established constructs for the evaluation. Perceived usefulness (Davis, 1989; Greven et al., 2003) is an integral aspect when evaluating the contribution to prescriptive and technological design knowledge bases (Baskerville et al., 2018; Venable et al., 2016). We included this construct to provide evidence for the usefulness for the proposed solution (Gregor & Hevner, 2013). In addition, usefulness is an important measurement in the context of XAI (Arrieta et al., 2020). The second construct was perceived ease of use (Davis, 1989; Greven et al., 2003). Scholars have shown that perceived ease of use is an important aspect for users' acceptance of information systems (Davis, 1989). The last construct was the intention to use the system (Venkatesh et al., 2003). Appendix 3 provides an overview of the survey.

We used IBM Statistics 27 for all statistical evaluations. The constructs were measured using a 5-point Likert scale (1 = completely disagree; 5 = completely agree). First, we measured the results in terms of mean values, standard deviation, and Cronbach's alpha. The constructs were rated as follows: perceived ease of use ($M = 3.99$, $SD = 0.59$; $\alpha = 0.824$), perceived usefulness ($M = 3.97$, $SD = 0.53$, $\alpha = 0.780$), and intention to use ($M = 4.03$, $SD = 0.62$, $\alpha = 0.760$). For all constructs, Cronbach's alpha was satisfactory at > 0.70. These measurements indicate that all constructs were evaluated positively, which confirms users' positive sentiments toward the UI during the first evaluation. Appendix 4 provides a box plot that summarizes the measurements. To examine the results more closely, we calculated the frequencies and percentages by summing the responses for all items of the constructs, an established method for describing ordinal and quantitative data (Blaikie, 2003). These measurements are provided in Table 2. The constructs of perceived ease of use and perceived usefulness consisted of six items, which resulted in 1,140 responses per construct from the 190 participants. The same participants generated 570 responses for the construct intention to use with three items. Our results show that for the constructs perceived ease of use and intention to use, more than 75% of the participants chose either agree or completely agree, and for the construct perceived usefulness, the values were only slightly lower and over 73%. Therefore, we conclude that the proposed UI design was perceived positively by the participants.

Table 1 Optimization of the DPs after the first design cycle

Design principle	Illustrative quotations	Descriptions of optimizations for the second design cycle
DP1	"It bothers me that the actual result is not communicated clearly." (Interviewee 9) "If the performance of the underlying algorithms is adequate, I would like to use such an application." (Interviewee 11)	We provided the confidence value of the AI model for this classification represented through a graphical "confidence barometer." We replaced this feature by clearly communicating the confidence value of the AI model as a percentage value
DP2	"The chart for the relevant words seems superfluous to me […] as I cannot see the added value of this chart for my decision-making." (Interviewee 1) "The colors for the highlighted words are too intrusive. These should be weakened somewhat. Especially, I imagine working with this application for a longer period, that would bother me." (Interviewee 3)	We combined two different XAI techniques: local explanations via highlighted words and a bar chart to show which words were important (feature importance) for a specific classification. Participants criticized the colors of the highlighted words and described the bar chart as redundant; therefore, we modified our design and dropped the bar chart
DP3	"I would be interested in the actions initiated against this user in the past." (Interviewee 2) "For me, the combination of so many charts is overwhelming and the tech cloud, for example, does not offer me much added value." (Interviewee 5)	We combined different charts to present case-related information. The participants found this to be overwhelming; therefore, we refined our design. The participants stated that the history of previous actions against a user constituted relevant knowledge
DP4	"If I want to contact the user, for instance, to warn him informally and point out the wrongdoing, then I would currently be missing a function." (Interviewee 8) "Legal actions are confusing […] what exactly is triggered?" (Interviewee 10)	We included the option of "legal actions," which was rather confusing for the participants. Therefore, we removed this option. Moreover, the participants expressed the need for an option that would enable them to contact the offending user directly

The participants were asked to provide feedback on their perceptions of the UI. Of the 190 participants, 114 provided textual feedback. To examine the resulting data, we used thematic analysis (Braun & Clarke, 2006). Table 3 presents positive perceptions and thought-provoking reflections that we identified in the provided feedback. The table also presents illustrative quotations and descriptions of insights and optimizations for the third design cycle. Consequently, we validated the participants' positive perceptions of the UI. It was also clear that the participants wanted to use such an application for other content-classification purposes. To summarize, most feedback was constructive and positive, and we identified few optimization opportunities.

With these insights, we concluded the second design cycle and used the optimization potential as input knowledge for the third design cycle, which is described in the next section.

6.3 Demonstration and Evaluation: Third Design Cycle

6.3.1 Production-Ready Environment

For the last design cycle, the prototype was implemented in a production-ready environment as a web-based UI. The first qualitative evaluation revealed the participants' positive perceptions of the implemented artifact and provided us with valuable feedback for further optimization. During the second design cycle, we validated the positive perceptions. The quantitative evaluation of the second design cycle motivated us to add a start page with an overview of all cases to address the thought-provoking reflections. Figure 8 provides an overview of the web-based UI.

Figure 9 provides an overview of the case page with highlighted DPs. DP1 was implemented by clearly communicating the AI-based classification outcome (DF1) and providing a confidence value in percentage for each case (DF2). DP2 was implemented by providing a local explanation for each case in the form of highlighted words (DF3). We implemented DP3 by combining two different charts that summarized the offending user's history in relation to hate speech (DF4), providing the history of the initiated actions against the offending user (DF5), and adding a start page with an overview of all cases (DF6). DP4 was implemented by providing checkboxes to initiate actions (DF7), a feature to contact the offending user (DF8), an option to re-classify the current case (DF9), and navigation elements (DF10). Consequently, we implemented the DPs and the corresponding DFs in the web-based UI (Seidel et al., 2018).

In the following two subsections, we describe the final evaluation, which focused on (i) the importance of AI

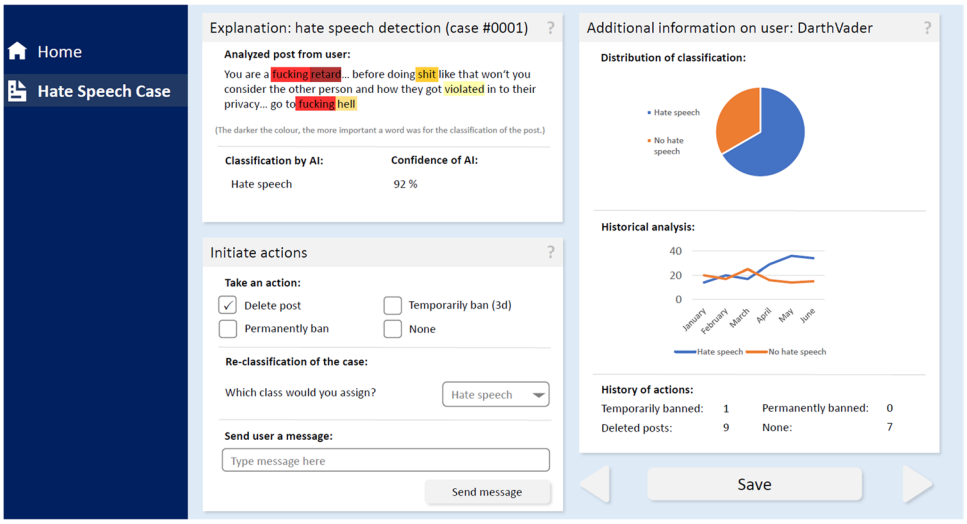


Fig. 7 Implementation of the prototype in the second design cycle

Table 2 Summary of the frequencies and percentages for the measured constructs ($N = 190$)

Likert scale	Perceived ease of use		Perceived usefulness		Intention to use	
	n	%	*n*	%	*n*	%
Completely agree	305	26.75	290	25.44	157	27.54
Agree	582	51.05	552	48.42	286	50.18
Neither agree nor disagree	208	18.25	274	24.04	114	20.00
Disagree	36	3.16	21	1.84	13	2.28
Completely disagree	9	0.79	3	0.26	0	0

Table 3 Positive perception and thought-provoking reflections based on the participants' evaluations

Insights	Illustrative quotations	Description of the insights and optimizations for the third design cycle
Positive perceptions	"I would like to have such an application that supports me in moderating social groups." (Participant 17) "It would be cool if this app would also classify other unwanted content, such as spam, phishing, or sexist statements." (Participant 61) "For me personally, such an application could really improve my efficacy in moderating social groups with larger user numbers." (Participant 65) "I like the highlighted words to explain which words were most important for this classification. It is easy to comprehend and validate." (Participant 74)	The participants' positive perceptions of the UI during the first design cycle were confirmed in this evaluation. The participants wanted to use such an application in their work. They mentioned that such an application should also detect other unwanted content, such as spam or sexist statements, which highlights our design's potential for adaptation. In addition, the participants claimed that the highlighted words eased comprehension of the AI-based decision-making
Thought-provoking reflections	"I find the navigation from case to case, without an overarching navigation structure, suboptimal." (Participant 17) "I am currently missing an overview of all the cases." (Participant 41) "I would like to test such an application in an even more real environment." (Participant 78) "It is difficult to get an overview of the cases or to keep track of them when one has to navigate through the individual cases." (Participant 103)	The participants expressed their need for an overview of all cases to have a holistic perspective on their workload and the cases. In the third design cycle, we addressed this aspect by adding a start page with an overview of all cases to our artifact and design knowledge. Moreover, by implementing the artifact as a web-based prototype in the third design cycle, we trialed the artifact in a more realistic environment

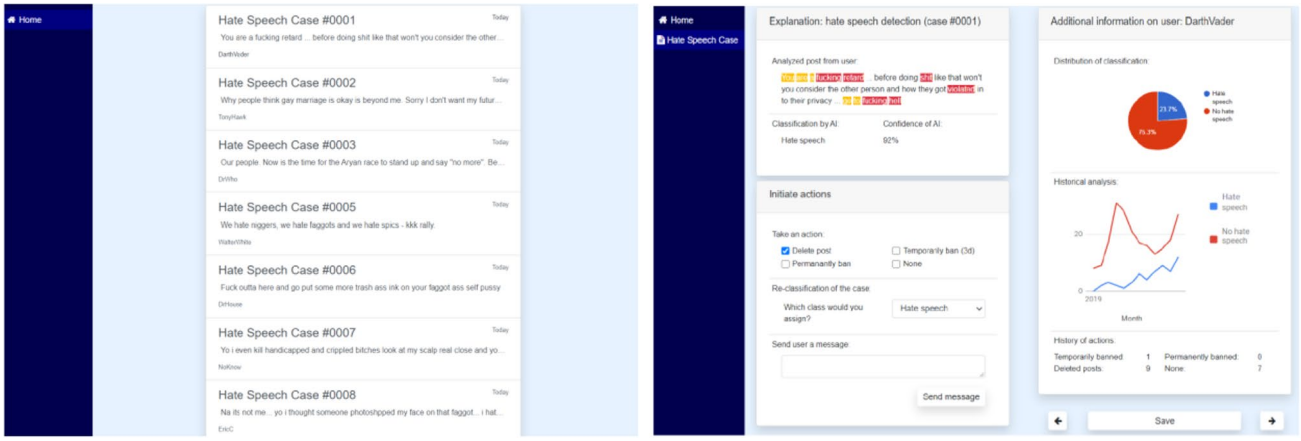


Fig. 8 Implementation of the prototype in the third design cycle (left: start page; right: a specific case)

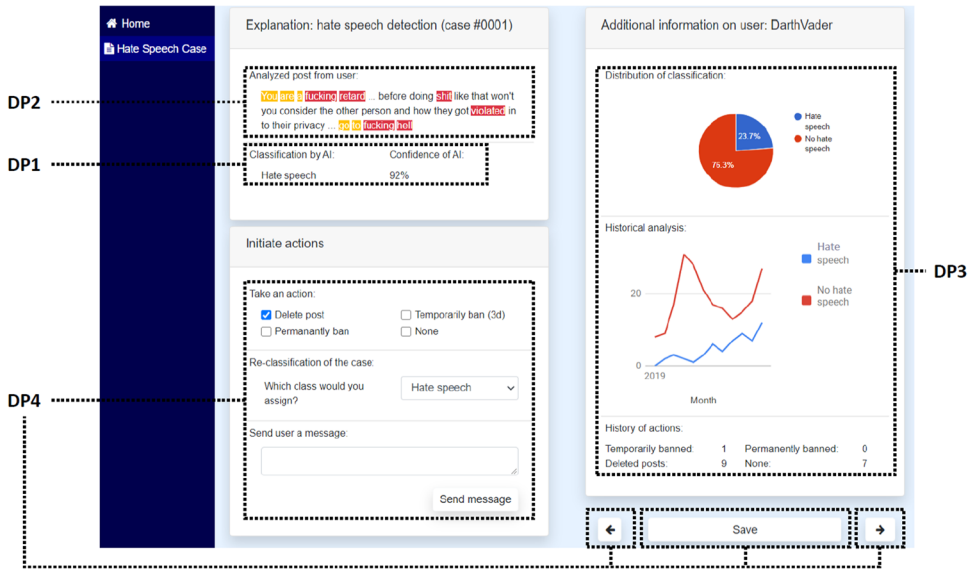


Fig. 9 Final set of the DPs implemented in the web-based UI (exemplary case page)

explainability for end users and (ii) the perceived reusability of our proposed DPs by practitioners.

6.3.2 Final Evaluation by End Users: The Importance of Local Explanations

In the last evaluation, we tested the relevance of local explanations by measuring their influence on perceived cognitive effort, perceived informativeness, mental model, and trustworthiness toward the AI-based artifact. Cognitive effort is important to consider when designing or implementing explanations (Arrieta et al., 2020). Perceived informativeness refers to the overall perception of aspects related to information quality (Zhang et al., 2014). Moreover, informativeness is one of XAI's overall goals and is relevant to a broad range of target audiences (Arrieta et al., 2020). Mental models are another important concept in the context of XAI research and the perception of explanations (Arrieta et al., 2020; Kühl et al., 2019). For our construct, we used the mental model's subdimension called "processes," which refers to users' overall understanding of the work processes involved in the artifact (Vitharana et al., 2016). Trustworthiness is an important factor for user acceptance and the intention to use systems or services (Carter & Bélanger, 2005). In addition, trustworthiness is an important goal in XAI research (Adadi & Berrada, 2018; Arrieta et al., 2020; Gunning & Aha, 2019; Meske et al., 2020; van der Waa et al., 2021). Overall, we developed the following hypotheses:

- H1. Providing a UI with local explanations leads to users experiencing reduced perceived cognitive effort compared to a UI without local explanations.
- H2. Providing a UI with local explanations leads to users experiencing an increased (a) perceived informativeness, (b) mental model, and (c) trustworthiness compared to a UI without local explanations.

Figure 10 shows our research model for this experiment.

The UI was evaluated by means of an experiment with participants recruited via CloudResearch and MTurk, using the same admission criteria as in design cycle two. After participating in one of the two experiments (AI or XAI), the participants were automatically excluded from participating in the second experiment. The evaluation was conducted with 360 participants, 180 participants per group. Appendix 5 provides an overview of the demographic data, and Appendix 6 describes the survey used. Ordinally scaled data were collected using a Likert scale. The resulting data was not normally distributed ($p < 0.001$), which was determined through a Kolmogorov–Smirnov test. Appendix 7 shows the Kolmogorov–Smirnov test for normal distribution and Cronbach's alpha. To test our hypotheses, we compared the mean values of the AI and XAI groups using the Mann–Whitney U test (Mann & Whitney, 1947).

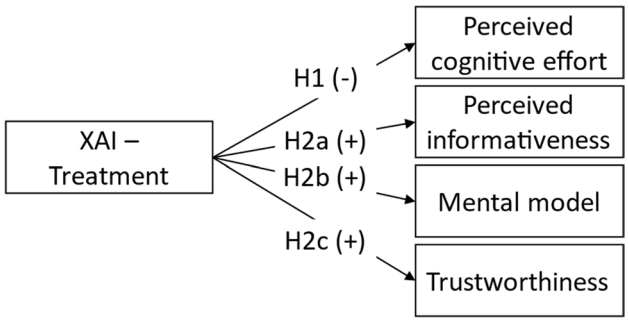


Fig. 10 Research model for the evaluation and comparison of AI and XAI

As our hypotheses were directional, we conducted a one-sided (one-tailed) test. Table 4 provides an overview of the Mann–Whitney U test and summarizes the mean rank, sum of ranks, Mann–Whitney U, z value (examined for significance), and effect size (Pearson correlation coefficient r). Appendix 8 provides the measurements of the constructs represented using a box plot.

The results indicate that the local explanation has a direct significant and positive effect with a strong effect size on perceived cognitive effort. This was also reflected in the measurements for the mean rank and sum of ranks, which were lower for the XAI group than for the AI group. Consequently, we assume that H1 is supported. Moreover, the results indicate that local explanation has a direct significant and positive effect with a small-to-medium effect size on perceived informativeness, mental model, and trustworthiness. The measurements for the mean rank and sum of ranks also reflected this finding, as they were higher for the XAI group than for the AI group. Therefore, we assume that H2(a, b, c) is supported.

6.3.3 Final Evaluation by Developers: Reusability of the Proposed Design Principles

By having practitioners evaluate the DPs, we addressed the risk of proposing DPs that are not applicable or useful in practice (Ivari et al., 2021). We adapted the template of Ivari et al., (2018, 2021), which includes the constructs of accessibility, importance, novelty and insightfulness, actability and guidance, and effectiveness (Likert scale 1–5, with 1 = completely disagree, 5 = completely agree). We complemented the information regarding our DPs with introductory materials on AI-based DSSs, hate speech detection and UI design (see Appendix 9) based on the recommendations of Ivari et al. (2021). Accessibility represents successful communication of DPs to re-users, whereas importance refers to an estimation of the relevance of the addressed problem in the real world (Ivari et al., 2021). Novelty and insightfulness serve as indicators of

Table 4 Summary of the Mann–Whitney U test ($N = 360$)

Construct	Treatment	Mean rank	Sum of ranks	Mann–Whitney U	z	Sig	r
Perceived cognitive effort	AI	239.63	43,134.00	5556.000	-10.829	$p < 0.001$	0.57
	XAI	121.37	21,846.00				
Perceived informativeness	AI	140.22	25,240.00	8950.000	-7.423	$p < 0.001$	0.39
	XAI	240.78	39,740.00				
Mental model	AI	158.80	28,584.00	12,294.000	-4.012	$p < 0.001$	0.21
	XAI	202.20	36,396.00				
Trustworthiness	AI	143.24	25,782.50	9492.500	-6.876	$p < 0.001$	0.36
	XAI	217.76	39,197.50				

whether practitioners are provided with new knowledge and insights (Ivari et al., 2018). Actability and guidance show whether DPs can be implemented in practice due to being actable and providing adequate guidance, while effectiveness refers to the potential relative value of the DPs from the perspective of practitioners (Ivari et al., 2018, 2021).

The DPs were evaluated by 80 practitioners. Of the 80 participants, 64 had a minimum of one year of experience in the domain of software development. Appendix 10 provides an overview of further demographic characteristics of the 80 participants. The evaluation was carried out in the same manner as the evaluation of the second design cycle and is presented accordingly. The constructs were rated as follows: accessibility ($M = 3.90$, $SD = 0.81$, $\alpha = 0.892$), importance ($M = 4.10$, $SD = 0.62$, $\alpha = 0.799$), novelty and insightfulness ($M = 3.82$, $SD = 0.71$, $\alpha = 0.780$), actability and guidance ($M = 3.87$, $SD = 0.61$, $\alpha = 0.826$), and effectiveness ($M = 3.97$, $SD = 0.57$, $\alpha = 0.869$). Cronbach's alpha was > 0.70 for all constructs. All measurements indicated practitioners' positive perceptions of the DPs. Appendix 11 presents the measurements as a box plot. To carry out a nuanced analysis of these measures, we calculated the frequencies and percentages (see Table 5) by summing the responses for all items of the individual constructs. The constructs of accessibility, importance, and novelty and insightfulness consisted of three items each and generated 240 responses each. Actability and guidance consisted of six items and generated 480 responses. Effectiveness consisted of five items and generated 400 responses. This overview of the data emphasizes the practitioners' positive perception of the DPs. The results further showed that more than 80% of the participants chose to agree or completely agree regarding accessibility and importance. For effectiveness, the value was slightly lower at over 78%, followed by actability and guidance at over 73%; for novelty and insightfulness, this value was over 68%. Moreover, 66 participants stated that they would adapt the DPs for a software development project. Therefore, we conclude that the practitioners had a positive perception of the proposed DPs, which exhibited an adequate degree of reusability. In the next section, we discuss our DSR project.

7 Discussion

7.1 Summary of the Findings

In this article, we have provided a comprehensive overview of our DSR project. The overarching goal of our study was to generate prescriptive knowledge that could be used in future research projects or practice (Gregor & Hevner, 2013; Hevner, 2020). In our DSR project, we focused on the implemented artifact and the reusability of the underlying DPs. We now summarize the findings of the DSR project and the conducted evaluations.

The qualitative evaluation during the first design cycle revealed versatile optimization potentials, which we used to refine our design knowledge. In this evaluation, we identified evidence regarding the usefulness of the proposed design via the target group of human moderators. The identification of such evidence is an important part of the evaluation process in DSR projects (Gregor & Hevner, 2013; Venable et al., 2016). We analyzed the gathered data using thematic analysis (Braun & Clarke, 2006). Overall, the participants perceived the designed UI positively and were interested in the application suggested by our design. Moreover, the same participants provided us with constructive criticism, which allowed us to further refine our design knowledge for the second design cycle.

The quantitative evaluation of the second design cycle validated the previously registered positive perceptions. We measured the three constructs of perceived ease of use, usefulness, and intention to use. These constructs are important for both evaluations in DSR contexts (e.g., Gregor & Hevner, 2013; Venable et al., 2016) and examinations based on an information systems perspective (e.g., Davis, 1989; Greven et al., 2003). Moreover, these constructs are used in XAI research (e.g., Arrieta et al., 2020; van der Waa et al., 2021). We reported the constructs' calculated values in terms of the mean, standard deviation,

Table 5 Summary of the frequencies and percentages for the measured constructs ($N = 80$)

Likert scale	Accessibility		Importance		Novelty and insightfulness		Actability and guidance		Effectiveness	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Completely agree	52	21.67	69	28.75	50	20.83	102	21.25	84	21.00
Agree	142	59.16	134	55.83	115	47.92	250	52.08	231	57.75
Neither agree nor disagree	21	8.75	29	12.08	58	24.16	96	20.00	73	18.25
Disagree	21	8.75	8	3.34	16	6.67	31	6.46	12	3.00
Completely disagree	4	1.67	0	0	1	0.42	1	0.21	0	0

and Cronbach's alpha, which we also visualized as a box plot (see Appendix 4). To analyze the results more closely, we calculated the frequencies and percentages by summing the responses for all constructs (Blaikie, 2003). This led us to discover that approximately 75% of the 190 participants rated the constructs using either agree or strongly agree. This is an important finding, as perceived ease of use and usefulness are relevant constructs for assessing the acceptance of an information technology artifact or as measurements of explanations (Arrieta et al., 2020; Davis, 1989; Greven et al., 2003). Moreover, the intention to use a DSS is also an important construct that is established in information systems and XAI literature (Gönül et al., 2006; Wang & Benbasat, 2009). By means of a text field, we collected feedback from the participants. We analyzed the collected data using thematic analysis (Braun & Clarke, 2006). We identified positive perceptions of the UI and final thought-provoking reflections for the third design cycle.

In the third design cycle, we conducted a two-sided evaluation. First, we investigated the influence of local explanations by conducting an experiment with two independent groups (AI vs. XAI, 180 participants per group). We used the constructs of perceived cognitive effort, perceived informativeness, trustworthiness, and mental model (process). All these constructs are relevant for research on information systems and XAI (Arrieta et al., 2020; Gönül et al., 2006; Greven et al., 2003; Meske et al., 2020; Vitharana et al., 2016; Wang & Benbasat, 2009). We derived directional hypotheses H1 and H2(a, b, c), which we evaluated using a one-sided (one-tailed) Mann–Whitney U test. Based on statistical evaluations, we confirmed both hypotheses and uncovered a direct and significant influence of local explanations with varying effect sizes. Then, we evaluated the DPs using the minimum reusability approach (Ivari et al., 2018, 2021). As in the second design cycle, we reported constructs' values in terms of the mean, standard deviation, and Cronbach's alpha, which we visualized as a box plot (see Appendix 11). We also calculated the frequencies and percentages of the responses for all constructs to obtain more nuanced insights. Consequently, we uncovered that approximately 70% of the 80 practitioners rated the constructs using either agree or strongly agree. For the constructs of accessibility and importance, these measurements exceeded 80%. Moreover, 66 out of the 80 practitioners stated that they would use the proposed design knowledge for a software development project. Based on these insights, we conclude that the proposed DPs have been evaluated as helpful by practitioners. Consequently, in this last design cycle, we generated insights into how relevant local explanations are for DSS UI design in the domain of hate speech detection; moreover, we found that that the DPs were evaluated as having a high degree of reusability. Therefore, we answered both of our research questions. In the next subsection, we present the theoretical implications of our work.

7.2 Theoretical Implications

We believe that according to the DSR knowledge contribution framework by Gregor and Hevner (2013), our research project can be categorized as an improvement. This categorization is justified because we have developed a new solution for a known problem. Moreover, we have proposed a set of DPs that can be described as nascent design theory, and we have evaluated their reusability with practitioners, who rated the DPs as reusable (Gregor & Hevner, 2013; Ivari et al., 2018, 2021). Therefore, we have developed applicable prescriptive design knowledge with a focus on UIs for hate speech detection systems with local explanations. This knowledge can be used as input knowledge and potentially extended by future research (vom Brocke & Maedche, 2019; vom Brocke et al., 2020; Hevner, 2020), especially by researchers focusing on the development of hate speech detection systems or related concepts.

With respect to XAI-based explanations and their integration in UIs, our results from the experiment in design cycle three emphasize the high relevance of user evaluations in the corresponding application domain, which is emphasized in scientific literature (Adadi & Berrada, 2018; van der Waa et al., 2021). The need for evaluation by users within the individual application domain is also described as highly relevant due to their varying backgrounds, expertise, information needs, and expectations (Barda et al., 2020). In the experiment that we performed during the third design cycle, we found a direct significant and positive effect of local explanations on the constructs perceived cognitive effort, perceived informativeness, mental model, trustworthiness Therefore, we have illustrated that local explanations can support the achievement of major XAI goals, such as trustworthiness and informativeness (Arrieta et al., 2020). This position is in line with research on online advisory tools: for example, Li and Gregor (2011) emphasized the need for more transparent decision-making processes in the design of UIs of DSSs. Furthermore, Cheng et al. (2019) found that users' trust in algorithmic decisions was not affected by the explanation interface that the researchers had developed. In contrast, we have shown that local explanations have a significant impact on trustworthiness and further constructs, which supports the argument that explanations should be evaluated with corresponding stakeholders from the targeted application domain (Barda et al., 2020; Meske et al., 2020; van der Waa et al., 2021). Consequently, we believe that local explanations are an

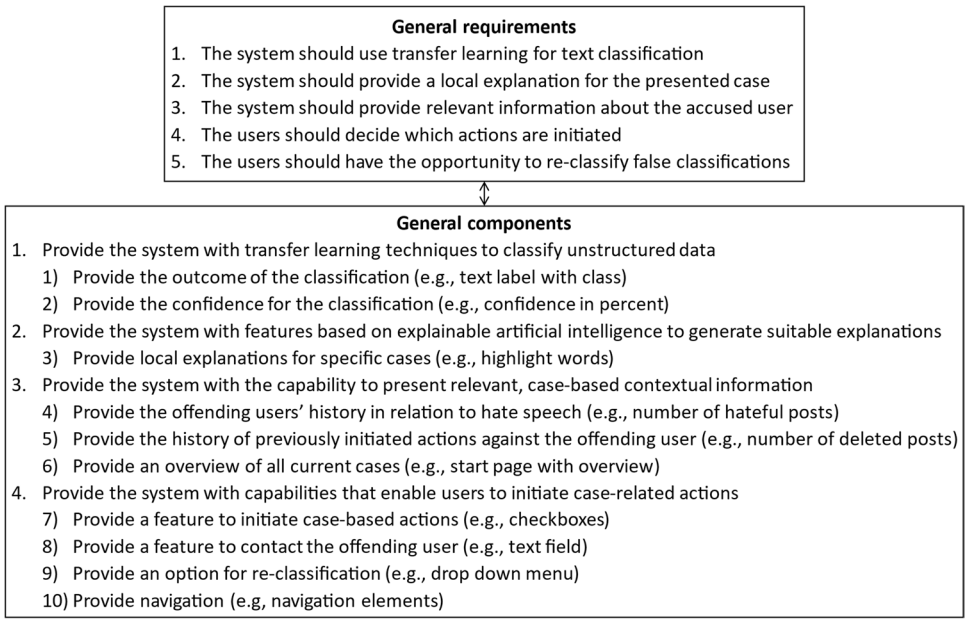


Fig. 11 Summary of the general requirements and general components represented as an EDT (Baskerville & Pries-Heje, 2010)

integral part for the design knowledge proposed in this paper.

Despite the interesting approaches on which related systems for automated hate speech detection are based, we have addressed the following shortcomings. First, we have introduced rigorously evaluated prescriptive design knowledge that can be used in future research projects as input knowledge (vom Brocke & Maedche, 2019). The reusability of the proposed prescriptive design knowledge was evaluated with practitioners and conducted according to the recommendations of Ivari et al., (2018, 2021). Participants of the evaluation round communicated a positive perception of the prescriptive design knowledge and rated it as reusable. Second, we have involved end users in a qualitative as well as quantitative evaluation of the instantiated design knowledge as UI, which was not done in related work (e.g., Modha et al., 2020; Paschalides et al., 2020; Pereira-Kohatsu et al., 2019; Ullmann & Tomalin, 2020). Here, we found that the design was positively perceived by the end users which provided us with valuable feedback for optimizing the underlying design knowledge. Results illustrate that 73% of the participants and more choose the rating of agree or completely agree when rating the constructs perceived ease of use, perceived usefulness, and intention to use. Therefore, we not only introduce prescriptive design knowledge for the design of UIs for automated hate speech detection systems. We also provide empirical evidence for the perception of the proposed design by the end users as well as practitioners who eventually could adapt the prescriptive design knowledge for suitable software development projects.

In the next section, we summarize the general requirements and general components as an EDT, which aims to provide a functional explanation for the implementation of the proposed DPs, DFs and addressed DRs (Baskerville & Pries-Heje, 2010).

7.3 An Explanatory Design Theory as Conditional Functional Explanation

We summarized the general requirements and general components as an EDT in Fig. 11 and used the simple as well as elegant structure proposed by Baskerville and Pries-Heje (2010). General requirements represent the adapted DRs, and general components represent the DPs and DFs. These elements were grounded in the state-of-the-art knowledge bases and refined via the insights gained during the three consecutive design cycles. Moreover, the underlying elements of this EDT were evaluated qualitatively and quantitatively, with a focus on aspects such as usefulness, reusability, or the influence of local explanations (Gregor & Hevner, 2013; Ivari et al., 2021; van der Waa et al., 2021). Consequently, this EDT provides a functional explanation for the related generalized requirements and the related generalized components of the proposed solution (Baskerville & Pries-Heje, 2010).

Through the proposed DPs and EDT, we have addressed the lack of applicable prescriptive design knowledge for the design of DSS UIs in the domain of hate speech detection for human moderators. We have described different DSSs for hate speech detection on social media platforms,

such as a web browser plugin that visualizes aggressiveness (Modha et al., 2020); MANDOLA, a system based on big data approaches (Paschalides et al., 2020); or the quarantining framework proposed by Ullmann and Tomalin (2020). Despite these interesting contributions, scholarship has lacked concrete prescriptive knowledge on how to design UIs in the context of hate speech detection and how the design is perceived by users; in addition, many studies have tended to focus on the end user. HaterNet, a system used by Spanish authorities, has a stronger relation to our design, as HaterNets' design also focuses on hate speech detection and the monitoring of social media (Pereira-Kohatsu et al., 2019). However, the HaterNet system and its design are focused on Twitter, prescriptive design knowledge is not available, and there are no insights into how the design affects users. Therefore, we have developed design knowledge that can be adapted by practitioners and is not limited by the size or type of social media platform. Moreover, by means of three consecutive design cycles, we have generated knowledge about how users (i.e., human moderators) perceive the design and to what extent it is reusable by practitioners. Lastly, we argue that the design can be extended to other concepts related to hate speech, such as cyberbullying, racism, or sexism, on social media platforms (Fortuna & Nunes, 2018; MacAvaney et al., 2019). In the next section, we discuss the limitations and future research opportunities.

7.4 Limitations and Future Research Opportunities

We have reported the process of our DSR project according to established guidelines (Gregor & Hevner, 2013; Gregor et al., 2020; Ivari et al., 2018, 2021; Peffers et al., 2007). Nevertheless, our approach had certain limitations. First, despite the implementation of ULMFiT, we focused on UIs. Consequently, we neglected the backend perspective, as realistic interactions with the UI were only simulated. Second, despite the involvement of 641 participants, we did not investigate the design using a case study. Therefore, we did not assess how this design could influence dimensions such as task performance in a real-world work environment. This limitation could be addressed by taking up our proposed design knowledge and extending to the aforementioned dimensions. Third, we used a dataset with two classes. However, hate speech is a nuanced subject, and there are several related concepts. Therefore, future studies could, for instance, integrate a greater number of detectable concepts. Fourth, the evaluations were conducted in controlled settings and environments. This is related to the implementation of our proposed design using a case study, which could produce empirical knowledge, for example, on how the design is perceived from stakeholders in an enterprise context. In addition, our design could be used as input knowledge (vom Brocke et al., 2020; Hevner, 2020) and extended with such concepts as human in the loop, interactive machine learning, or hybrid intelligence (Adadi & Berrada, 2018; Akata et al., 2020; Meske et al., 2020).

8 Conclusion

We have developed a set of rigorous evaluated DPs (Ivari et al., 2018, 2021) for the development of UIs in the domain of XAI-based hate speech detection. Our evaluations have highlighted users' and practitioners' positive perceptions of the design, which also allowed us to optimize the overall design knowledge. The evaluations of the three design cycles generated valuable insights into how the design is perceived by the target audience and the influence of local explanations. It is important to note that we included the perspectives of both users and practitioners, as the latter could adapt the DPs. From a theoretical perspective, we have addressed an important real-world challenge regarding the design of UIs for AI-based DSSs and its perception by users in the context for hate speech detection, for which we have proposed a scientifically grounded and evaluated design. The developed and evaluated design knowledge was formalized as prescriptive knowledge and summarized as an EDT (Baskerville & Pries-Heje, 2010; vom Brocke et al., 2020; Hevner, 2020). In terms of practical contributions, we have developed a set of reusable DPs. According to our evaluation, 83% of the 80 practitioners stated they would use the proposed DPs in a suitable software development project. Lastly, this study further highlights the need for more research in the fight against hateful content in the digital sphere, something that was emphasized by 86% of 641 participants who had experienced hate speech personally.

Appendix 1

Performance Metrics for the ULMFiT Model, Generated with Scikit-Learn

Table 6 Metrics of the ULMFiT model on test data split

Class/ Metric	Precision	Recall	F1-Score	Examples
Hate Speech	83.33%	63.93%	72.35%	219
No Hate Speech	87.30%	95.10%	91.03%	571
Macro Average	85.32%	79.51%	81.69%	790
Weighted Average	86.20%	86.46%	85.85%	790

Appendix 2

Demographic Data for Participants of the Evaluation within Design Cycle Two

Table 7 Overview of demographic data for the participants (N = 190)

Characteristic	N	%
Self-identified gender		
Female	65	34.21%
Male	125	65.79%
Age		
< 20 years	0	0%
20 – 29 years	63	33.16%
30 – 39 years	72	37.89%
40 – 49 years	46	24.21%
50—59 years	9	4.74%
> 59 years	0	0%
Have you ever been affected by hateful content on a social media platform?		
Yes	172	90.53%
No	18	9.47%

Appendix 3

Questionnaire for the Second Design Cycle

Table 8 Overview of the questionnaire for the second design cycle

Construct/ Item	Statement	Reference
Perceived ease of use		Davis, 1989; Greven et al., 2003
PEOU1	The user interface for hate speech detection is easy to use	
PEOU2	It is easy to become skillful at using the user interface for hate speech detection	
PEOU3	Learning to operate the user interface for hate speech detection is easy	
PEOU4	The user interface for hate speech detection is flexible to interact with	
PEOU5	My interaction with the user interface for hate speech detection is clear and understandable	
PEOU6	It is easy to interact with the user interface for hate speech detection	
Perceived usefulness		
PU1	The user interface for hate speech detection is useful for detecting hateful content	
PU2	The user interface for hate speech detection improves my performance in detecting hateful content	
PU3	The user interface for hate speech detection enables me to identify hateful content faster	
PU4	The user interface for hate speech detection enhances my effectiveness in identifying hateful content	
PU5	The user interface for hate speech detection makes it easier to detect hateful content	
PU6	The user interface for hate speech detection increases my productivity in detecting hateful content	
Intention to use		Venkatesh et al., 2003
ITU1	If available, I intent to use the user interface for hate speech detection in the next six months	
ITU2	If available, I predict I will use the user interface for hate speech detection in the next six months	
ITU3	If available, I plan to use the user interface for hate speech detection in the next six months	
Question open text field		
Q1	Does something bother you about the user interface, are there things that you would like to change? If so, please describe your ideas	

Appendix 4

Box Plot for the Measurements of the Second Design Cycle

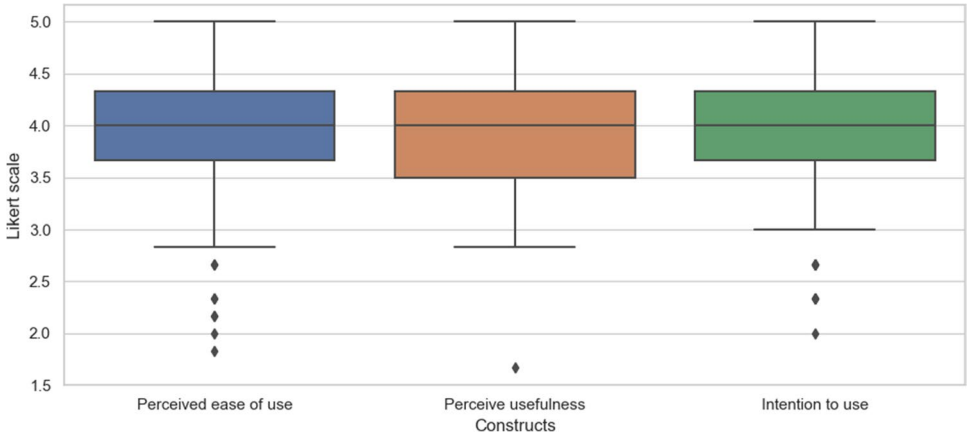


Fig. 12 Graphical summary of the measurements for the second design cycle

Appendix 5

Demographic Data for Participants of the Evaluation within Design Cycle Three with End Users

Table 9 Overview of demographic data for the participants (N = 360)

Characteristic	N	%
Self-identified gender		
Female	147	40.83%
Male	213	59.17%
Age		
< 20 years	0	0%
20 – 29 years	119	33.06%
30 – 39 years	131	36.38%
40 – 49 years	83	23.06%
50—59 years	27	7.50%
> 59 years	0	0%
Have you ever been affected by hateful content on a social media platform?		
Yes	317	88.06%
No	43	11.94%

Appendix 6

Questionnaire for the Third Design Cycle with End Users

Table 10 Overview of the questionnaire for the third design cycle

Construct/ Item	Statement	Reference
Perceived cognitive effort		Wang & Benbasat, 2009
PCE1	The task of identifying hateful content using the user interface for hate speech detection is very frustrating	
PCE2	Using the user interface for hate speech detection, it is hard to find the information I need to help me decide what to do	
PCE3	The task of identifying hateful content using this user interface for hate speech detection takes too much time	
PCE4	The task of identifying hateful content using the user interface for hate speech detection is difficult	
PCE5	Identifying hateful content using the user interface for hate speech detection requires too much effort	
PCE6	The task of identifying hateful content using the user interface for hate speech detection is too complex	
Perceived informativeness		Zhang et al., 2014
PIN1	The system provides relevant information about the present hate speech case	
PIN2	The system provides complete information about the present hate speech case	
PIN3	The system provides timely information about the present hate speech case	
Trustworthiness		Carter & Bélanger, 2005
TRUSTW1	I think I can trust the user interface for hate speech detection	
TRUSTW2	The user interface for hate speech detection can be trusted to carry out hate speech detection faithfully	
TRUSTW3	I trust the user interface for hate speech detection to keep my best interest in mind	
Mental model (process)		Vitharana et al., 2016
MMP1	I characterize my understanding of the types of processes (e.g., detect hate speech and initiate appropriate actions) involved in the user interface for hate speech detection as high	
MMP2	I characterize my understanding of the sequence of processes involved in the user interface for hate speech detection as high	
MMP3	I characterize my overall understanding of the processes involved in the user interface for hate speech detection as high	
Question open text field		
Q1	Does something bother you about the user interface, are there things that you would like to change? If so, please describe your ideas	

Appendix 7

Test for Normal Distribution and Cronbach's Alpha

Table 11 Summary for the Kolmogorov Smirnov test for normal distribution and Cronbach's alpha

Construct	Treatment	Median	Mean ± SD	Statistics	Df	Sig	Cronbach's alpha
Perceived cognitive effort	AI	3.50	3.56 ± 0.63	0.150	180	$p < 0.001$	0.830
	XAI	2.50	2.51 ± 0.82	0.106		$p < 0.001$	0.936
Perceived informativeness	AI	3.34	3.25 ± 0.81	0.184		$p < 0.001$	0.822
	XAI	3.96	3.89 ± 0.68	0.123		$p < 0.001$	0.734
Mental model	AI	3.67	3.61 ± 0.74	0.152		$p < 0.001$	0.814
	XAI	4.00	3.92 ± 0.58	0.116		$p < 0.001$	0.717
Trustworthiness	AI	3.67	3.39 ± 0.78	0.176		$p < 0.001$	0.787
	XAI	4.00	3.94 ± 0.69	0.158		$p < 0.001$	0.767

Appendix 8

Box Plot for the Measurements of the Third Design Cycle with End Users

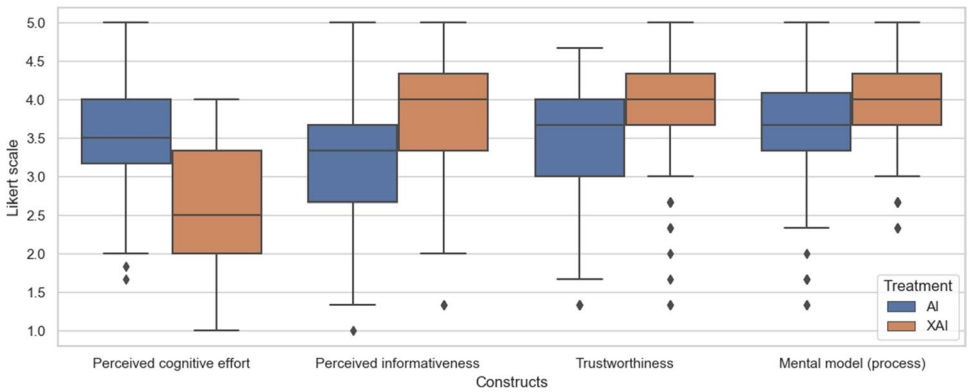


Fig. 13 Graphical summary of the measurements for the third design cycle

Appendix 9

Information for the Third Design Cycle with Developers

Here, we provide the survey design for the evaluation with the developers within the third design cycle.

In a design science research project on hate speech detection supported by artificial intelligence (AI), we have derived four "design principles" for developers. These design principles shall help developers in building systems that eventually support social media moderators in detecting hate speech and act accordingly. In this survey we ask you as a developer to evaluate the derived design principles (e.g., are they useful, understandable, effective, …?).

The survey is grouped into three parts. Within the first part we ask a few general and demographic questions. Afterwards you are provided with reading material, and we kindly ask you to read this material carefully. The last and third part is a short survey which is related to the mandatory reading material in part two.

Thank you for taking part in this survey.

Please read the following text carefully. The text contains all relevant information for the survey. Thank you very much.

The focus of this research project and the application domain of hate speech detection

Within this research project we focus on the user interfaces of decision support systems for human moderators in social media platforms. We are particularly interested in the phenomenon of hate speech detection. Scientific studies have already proven the far-reaching consequences of hate speech, including, for example: (i) the psychological, or (ii) emotional damage as well as (iii) suicide among young people. Hateful content is often examined by human moderators and removed if necessary. Here, the opportunity emerges to support the human moderators with a decision support system. Approaches from the field of artificial intelligence (based on machine learning) are currently the state-of-the-art for automated hate speech detection in textual data.

However, these state-of-the-art artificial intelligence techniques have also weak spots. One of these weaknesses is the so-called black box problem. This problem refers to the circumstance that these approaches are complex and therefore, the decision-making process often remains hidden. In other words, such systems do not justify their recommendations or outputs. This can be problematic for different reasons, for example, if human decision-makers must initiate actions, how can they trust or comprehend the systems output? This is one of the problems that the research field of explainable artificial intelligence addresses. Researchers of this discipline introduce methods to generate explanations for such black box systems. An exemplary explanation in a hate speech detection scenario could be the highlighting of the most relevant words for the underlying artificial intelligence. Therefore, the human moderator could easily comprehend which words lead to the classification of the system.

This is where our research and proposed design comes in, which we will describe in more detail below. We aim to contribute scientifically grounded design knowledge, and your view as a developer is of utmost importance for us to optimize the design knowledge.

We now describe how the four overarching design principles could inform the development and design of decision support systems in the context of hate speech detection.

The aim of the proposed design

We focus on the user interface of decision support systems to detect hate speech. The proposed design is rather abstract so that the core functionality could be adapted in different systems as well as domains. It was derived from scientific literature, instantiated as lightweight web application and evaluated with end users. This design is scientifically communicated as design principles. As a contribution, such design principles can be reused or extended in other research projects as well as practice. The here illustrated design principles have the aim to describe how a user interface for decision support in hate speech detection can be designed. It is important to note, that these design principles allow the developer or designer a certain degree of freedom in how they can be implemented. In the next section we describe the design principles and ask you to read the description carefully. Thank you for your effort, the survey questions will follow, after the design principles have been presented.

The design principles

Design principle 1: Provide the system with a transfer learning basis to classify unstructured data.

The first design principle focuses on a specific artificial intelligence technique, namely transfer learning. It is a state-of-the-art approach for text classification and requires relatively small data sets and computing time for training. As described before, artificial intelligence can outperform human experts and therefore it is more frequently investigated in the context of decision support systems. By combining the high performance of transfer learning with the advantage of requiring less data for training, such approaches could be easier adapted by smaller providers of social media platforms, with lower amounts of users and hence data. This design principle could be instantiated, for example, by integrating a transfer learning technique to classify text data, provide the outcome of the classification in the user interface and it could be complemented by confidence metrics in percent (or probability) of the transfer learning classification. The following figures depicts this easy design.

Design principle 2: Provide the system with features based on explainable artificial intelligence.

The second design principle focuses on an explanatory feature. Such explainability features should justify the system's output (for instance the classification of text as being hate speech or not). Through such explanations the user can control the system and answer questions such as: does it work correctly, why was a classification made, or can I trust the system? By providing explanations for classifications, we can also provide the human decision-maker with an opportunity to detect a potential bias within the system. This design principle could be instantiated, for example, by utilizing a method that highlights the most relevant words for the classification and provide the information as a heatmap within the user interface. Methods that allow for such highlighting are, for instance, GradCAM, Local-Interpretable Model-Agnostic Explanations or Heatmaps.

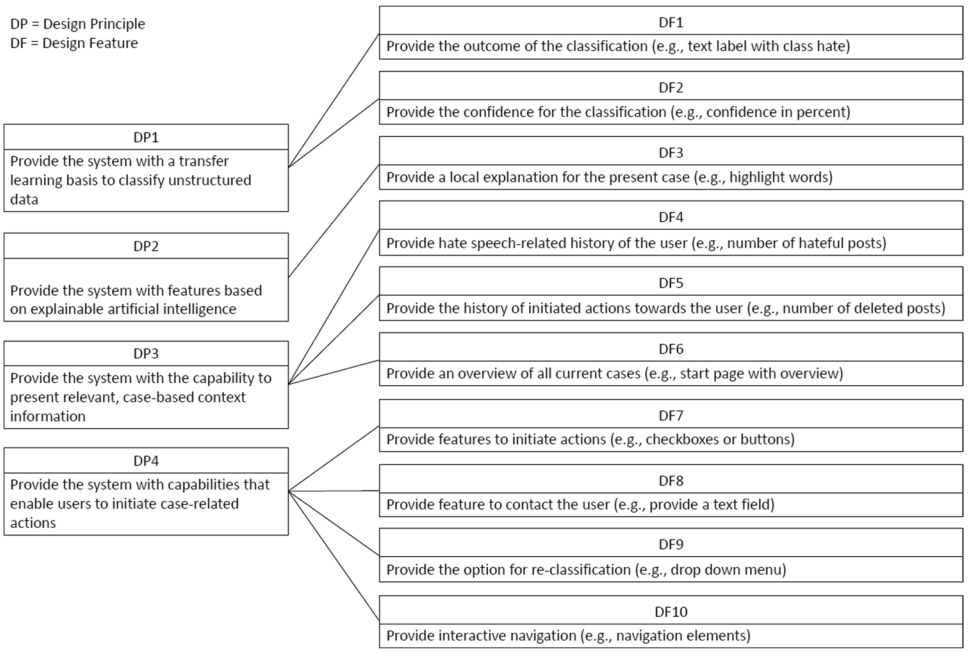
Design principle 3: Provide the system with the capability to present relevant, case-based context information.

The third design principle focuses on the relevant case-based context information for the application. Social media platforms provide a versatile range of data types, ranging from text within a post to the amount of posts a user published. Therefore, it is important to identify the relevant case-based and context information for a decision-maker within the application domain, and to provide this information through the user interface. The relevant case-based context information and its representation depend on different characteristics such as the application domain or data types. In the case of hate speech detection, the knowledge base could be represented through hate speech-related history of a user (has user previously published hateful content?), or the history of actions against a user (was the user temporarily banned?).

Design principle 4: Provide the system with capabilities that enable users to initiate case-related actions.

The last design principle focuses on the decision-making itself (e.g., deleting the text classified as hate speech). The decisions should not be made by the system, only by the human user. Depending on the use case and the policies of the platform, the decisions or actions that can be taken should be integrated within the user interface. This design principle could be instantiated by providing checkboxes or buttons to initiate actions (e.g., delete hateful content) or to re-classify the present example in case of a false prediction.

To summarize the design principles and provide some examples for concrete design features we provide the following overview and afterwards an exemplary interface. As stated before, these are only simple design examples and developers are free to implement or design such features based on their design ideas.



Thank you very much for carefully reading the material. This is the last part of the survey. You are almost finished. Please answer the following questions regarding our four presented design principles to complete the survey:

Table 12 Adapted questionnaire for the evaluation of the design principles (based on Ivari et al., 2018; 2021)

Construct	Statement
Accessibility	
ACC1	The design principles are easy for me to understand
ACC2	The design principles are easy for me to comprehend
ACC3	The design principles are intelligible to me
Importance	
IMP1	In my view the design principles address a real problem in developing user interfaces for decision support systems with a focus on hate speech detection in practice
IMP2	In my view the design principles address an important – acute or foreseeable – problem in developing such user interfaces for decision support systems with a focus on hate speech detection in practice
IMP3 OWN	In my view the design principles represent an important source of information for the development of user interfaces for decision support systems with a focus on hate speech detection in practice
Novelty and insightfulness	
NOIN1	I find that the design principles convey new ideas to me
NOIN2	I find the design principles insightful to my own practice
NOIN3 OWN	I find that the design principles communicate novel design opportunities or design combinations to me
Actability and guidance	
ACGU1	I think that the design principles can realistically be carried out in practice
ACGU2	I think that the design principles can easily be carried out in practice
ACGU3	I find that the design principles provide sufficient guidance for developing such user interfaces for decision support systems with a focus on hate speech detection
ACGU4	I find that the design principles provide sufficient direction for such user interfaces for decision support systems with a focus on hate speech detection
ACGU5	I find that the design principles are not restrictive when designing such user interfaces for decision support systems with a focus on hate speech detection
ACGU6	I find that the design principles provide me with sufficient design freedom when designing such user interfaces for decision support systems with a focus on hate speech detection
Effectiveness	
EFF1	I believe that the design principles can help design user interfaces for decision support systems with a focus on hate speech detection
EFF2	I find the design principles useful for designing user interfaces for decision support systems with a focus on hate speech detection
EFF3	Compared to my current situation, I believe that the design principles would improve my performance in developing user interfaces for decision support systems with a focus on hate speech detection
EFF4	Compared to my current situation, I believe that the design principles would increase my productivity in developing user interfaces for decision support systems with a focus on hate speech detection
EFF5	Compared to my current situation, I believe that the design principles would enhance my effectiveness in developing user interfaces for decision support systems with a focus on hate speech detection

Appendix 10

Demographic Data for Participants of the Evaluation within Design Cycle Three with Developers

Table 13 Overview of demographic data for the participants (N = 80)

Characteristic	N	In percent
Self-identified gender		
Female	29	36.25%
Male	51	63.75%
Age		
< 20 years	0	0%
20 – 29 years	38	47.50%
30 – 39 years	32	40.00%
40 – 49 years	10	12.50%
50 – 59 years	0	0%
> 59 years	0	0%
How long have been working in the domain of software development?		
< 1 year	16	20.00%
1 – 4 years	39	48.75%
5 – 9 years	17	21.25%
> 9 years	8	10.00%
Which applications do you develop or design most often?		
Web pages	19	23.75%
Web applications	33	41.25%
Mobile applications	8	10.00%
Desktop applications	12	15.00%
Others	8	10.00%
Have you ever been affected by hateful content on a social media platform?		
Yes	54	67.50%
No	26	32.50%
Would you adapt the design principles for a suitable software development project?		
Yes	66	82.50%
No	14	17.50%
I do not know	0	0%

Appendix 11

Box Plot for the Measurements of the Third Design Cycle with the Developers

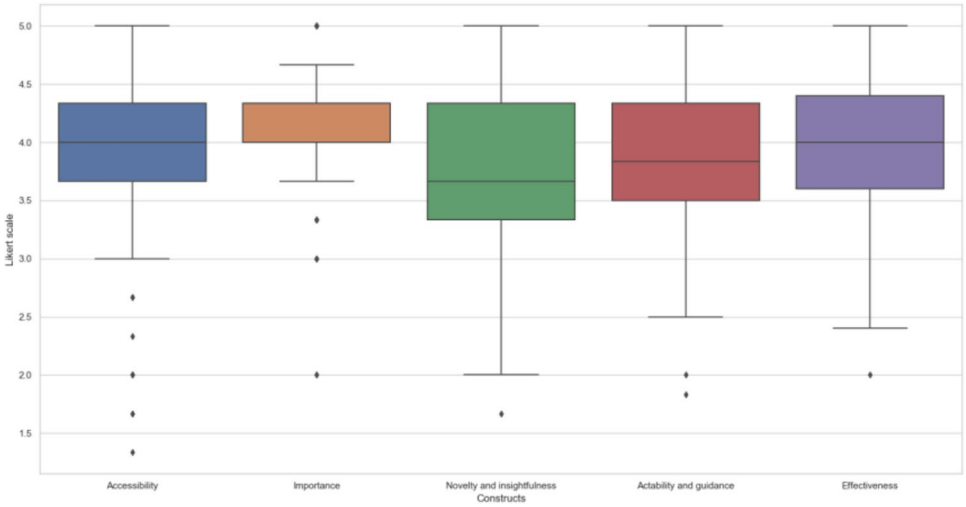


Fig. 14 Graphical summary of the measurements for the third design cycle

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

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

Arapostathis, S. G. (2021). A Methodology for Automatix Acquisition of Flood-event Management Information From Social Media: The Flood in Messinia, South Greece, 2016. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10105-z

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (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

Akata, Z., Balliet, D., Rijke, D., Dignum, F., Dignum, V., Fokkens, G. E., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Jonker, H. H., Jonker, C., Monz, C., Oliehoek, M. N., Oliehoek, F., Pakken, H., Schlbach, S., van der Gaag, L., van Harmelen, F., … Wlling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer, 53*(8), 18–28. https://doi.org/10.1109/MC/.2020.2996587

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review, 38*, 1–34. https://doi.org/10.1016/j.cosrev.2020.100311

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security, 90*, 1–11. https://doi.org/10.1016/j.cose.2019.101710

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design Science Research Contributions: Finding a

Balance between Artifact and Theory. *Journal of the Association for Information Systems, 19*(5), 358–376. https://doi.org/10.17705/1jais.00495

Baskerville, R., & Pries-Heje, J. (2010). Explanatory Design Theory. *Business & Information Systems Engineering, 2*, 271–282. https://doi.org/10.1007/s12599-010-0118-4

Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making, 20*, 1–16. https://doi.org/10.1186/10.1186/s12911-020-01276-x

Bilewicz, M., & Soral, W. (2020). Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology, 41*(1), 3–33. https://doi.org/10.1111/pops.12670

Blaikie, N. (2003). Analyzing Quantitative Data. *Sage Publications Ltd*. https://doi.org/10.4135/9781849208604

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

vom Brocke, J., & Maedche, A. (2019). The DSR grid: Six core dimensions for effectively planning and communicating design science research projects. *Electronic Markets, 29*, 379–385. https://doi.org/10.1007/s12525-019-00358-7

vom Brocke, J., Winter, R., Henver, A., & Maedche, A. (2020). Special Issue Editorial Accumulation and Evolution of Design Knowledge in Design Science Research: A Journey Through Time and Space. *Journal of the Association for Information Systems, 21*(3), 520–544. https://doi.org/10.17705/1jais.00611

Cadavid, J. P. U., Lamouri, S., Grabot, B., Pellerin, R., & Fortin, A. (2020). Machine learning applied in production planning and control: A state-of-the-art in the era of industry 4.0. *Journal of Intelligent Manufacturing, 31*, 1531–1558. https://doi.org/10.1007/s10845-019-01531-7

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Nature, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: Citizen trust, innovation and acceptance factors. *Information Systems Journal, 15*(1), 5–25. https://doi.org/10.1111/j.1365-2575.2005.00183.x

Celik, S. (2019). Experiences of internet users regarding cyberhate. *Information Technology & People, 32*(6), 1446–1471. https://doi.org/10.1108/ITP-01-2018-0009

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 559*, 1–12. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3290605.3300789

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

FastAI (2021, March 7). *Transfer learning in text*. fastai. Retrieved from https://docs.fast.ai/. Accessed 12 Jan 2022.

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Survey, 51*(4), 1–30. https://doi.org/10.1145/3232676

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly, 37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Gregor, S., Kruse, L. C., & Seidel, S. (2020). Research Perspectives: The Anatomy of a Design Principle. *Journal of the Association for Information Systems, 21*(6), 1622–1652. https://doi.org/10.17705/1jais.00649

Greven, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly, 27*(1), 51–90. https://doi.org/10.2307/30036519

Gunning, D., & Aha, D. W. (2019). DARPA's Explainable Artificial Intelligence (XAI). *AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Gupta, M., Parra, C. M., & Dennehy, D. (2021). Questioning Racial and Gender Bias in AI-based Recommendations: Do Espoused National Cultural Values Matter? *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10156-2

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision Support Systems, 42*, 1481–1493. https://doi.org/10.1016/j.dss.2005.12.003

Hevner, A. R. (2020). The duality of science: Knowledge in information systems research. *Journal of Information Technology*, 1-5. https://doi.org/10.1177/0268396220945714

Hinduja, S., & Patchin, J. W. (2019). *Cyberbullying Identification, Prevention, and Response*. Cyberbullying Research Center. Retrieved from https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2019.pdf. Accessed 12 Jan 2022.

Howard, J., & Gugger, S. (2020). Fastai: A Layered API for Deep Learning. *Information, 11*, 1–26. https://doi.org/10.3390/info11020108

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-Tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, Australia, 328–339. https://doi.org/10.18653/v1/P18-1031

Hu, Y., Xu, A., Hong, Y., Gal, D., Sinha, V., & Akkiraju, R. (2019). Generating Business Intelligence Through Social Media Analytics: Measuring Brand Personality with Consumer-, Employee-, and Firm-Generated Content. *Journal of Management Information Systems, 36*(3), 893–930. https://doi.org/10.1080/07421222.2019.1628908

Huang, H. H. (2003). Effects of multimedia on document browsing and navigation: An exploratory empirical investigation. *Information & Management, 41*(2), 189–198. https://doi.org/10.1016/S0378-7206(03)00047-8

Intel (2021). *Bleep*. Intel Corporations. Retrieved January from https://devmesh.intel.com/projects/bleep#about-section. Accessed 12 Jan 2022.

Ivari, J., Hansen, M. R. P., & Haj-Bolouri, A. (2018). A Framework for Light Resuability Evaluation of Design Principles in Design Science Research. *13th International Conference on Design Science Research and Information Systems and Technology: Designing for a Digital and Globalized World* (DESRIST 2018), India. https://doi.org/10.1007/978-3-319-91800-6

Ivari, J., Hansen, M. R. P., & Haj-Bolouri, A. (2021). A proposal for minimum reusability evaluation of design principles. *European Journal of Information Systems, 30*(3), 286–303. https://doi.org/10.1080/0960085X.2020.1793697

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management, 44*, 1–12. https://doi.org/10.1016/j.ijinfomgt.2018.09.003

Kaggle (2012). *Detecting Insults in Social Commentary*. Retrieved from https://www.kaggle.com/c/detectinginsults-in-social-commentary/overview. Accessed 12 Jan 2022.

Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems, 210*, 1–21. https://doi.org/10.1016/j.knosys.2020.106458

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and

implications of artificial intelligence. *Business Horizons, 62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons, 63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

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*, 1–11. https://doi.org/10.1016/j.dss.2020.113302

Kunst, M., Porten-Chee, P., Emmer, M., & Eilders, C. (2021). Do "Good Citizens" fight hate speech online? Effects of solidarity citizenship norms on user responses to hate comments. *Journal of Technology & Politics, 18*(3), 258–273. https://doi.org/10.1080/19331681.2020.1871149

Kühl, N., Lobana, J., & Meske, C. (2019). Do you comply with AI? – Personalized explanations of learning algorithms and their impact on employees' compliance behavior. *Fortieth International Conference on Information Systems 2019*, 1–6. Retrieved from https://aisel.aisnet.org/icis2019/paperathon/paperathon/1/. Accessed 12 Jan 2022.

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., & Seroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine, 94*, 42–53. https://doi.org/10.1016/j.artmed.2019.01.001

Li, M., & Gregor, S. (2011). Outcomes of effective explanations: Empowering citizens through online advice. *Decision Support Systems, 52*(1), 119–132. https://doi.org/10.1016/j.dss.2011.06.001

Li, Y., & Kettinger, W. J. (2021). Testing the Relationship Between Information and Knowledge in Computer-Aided Decision-Making. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10205-w

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and Solutions. *PLoS ONE, 14*(8), 1–16. https://doi.org/10.1371/journal.pone.0221152

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics, 18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Martens, D., & Provost, F. (2014). Explaining Data-Driven Document Classifications. *MIS Quarterly, 38*(1), 73–99. https://doi.org/10.25300/MISQ/2014/38.1.04

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society, 20*(6), 930–946. https://doi.org/10.1080/1369118X.2017.1293130

Meske, C., & Bunde, E. (2020). Transparency and Trust in Human-AI-Interaction: The Role of Model-Agnostic Explanations in Computer Vision-Based Decision Support. In Degen H., & Reinerman-Jones L., (Eds.), Artificial Intelligence in HCI. HCII 2020. *Lecture Notes in Computer Science, 12217*, 54–69. Springer, Cham. https://doi.org/10.1007/978-3-030-50334-5_4

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management,* 1-11. https://doi.org/10.1080/10580530.2020.1849465

Meske, C., & Amojo, I. J. (2020). Enterprise Social Bots as Perception-Benefactors of Social Network Affordances. *Forty-First International Conference on Information Systems 2020*, 1–17. Retrieved from https://aisel.aisnet.org/icis2020/social_media/social_media/5/. Accessed 12 Jan 2022.

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the Association for Information Systems, 16*(9), 779–837. https://doi.org/10.17705/1jais.00408

Modha, S., Majumder, P., Mandl, T., & Mandalia, C. (2020). Detecting and visualizing hate speech in social media: A cyber Watchdog for surveillance. *Expert Systems with Applications, 161*, 1–11. https://doi.org/10.1016/j.eswa.2020.113725

Motorny, S., Sarnikar, S., & Noteboom, C. (2021). Design of an Intelligent Patient Decision aid Based on Individual Decision-Making Styles and Information Need Preferences. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10125-9

Nienierza, A., Reinemann, C., Fawzi, N., Riesmeyer, C., & Neumann, K. (2019). Too dark to see? Explaining adolescents' contact with online extremism and their ability to recognize it. *Information, Communication & Society, 24*(9), 1229–1246. https://doi.org/10.1080/1369118X.2019.1697339

Park, H., Bellamy, M. A., & Basole, R. C. (2016). Visual analytics for supply network management: System design and evaluation. *Decision Support Systems, 91*, 89–102. https://doi.org/10.1016/j.dss.2016.08.003

Paschalides, D., Stephanidis, D., Andreou, A., Orphanou, K., Pallis, G., Dikaiakos, M. D., & Markatos, E. (2020). MANDOLA: A Big-Data Processing and Visualization Platform for Monitoring and Detecting Hate Speech. *ACM Transactions on Internet Technology, 20*(2), 1–21. https://doi.org/10.1145/3371276

Patton, M. Q. (2014). *Qualitative Research & Evaluation Methods* (4th ed.). Sage Publications Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

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

Peng, S., Wang, Y., Liu, C., & Chen, Z. (2020). TL-NER: S Transfer Learning Model for Chinese Named Entity Recognition. *Information Systems Frontiers, 22*, 1291–1304. https://doi.org/10.10007/s10796-019-09932-y

Pereira-Kohatsu, J. C., Quijano-Sanchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and Monitoring Hate Speech in Twitter. *Sensors, 19*(21), 1–37. https://doi.org/10.3390/s19214654

Plaza-del-Arco, F., Molina-Gonzalez, M., Urena-Lopez, L., & Martin-Valdivia, M. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications, 166*, 1–10. https://doi.org/10.1016/j.eswa.2020.114120

Ramos, G., Meek, C., Simard, P., Suh, J., & Ghorashi, S. (2020). Interactive machine teaching: A human-centered approach to building machine-learned models. *Human-Computer Interaction, 35*(5–6), 413–451. https://doi.org/10.1080/07370024.2020.1734931

Schneider, J., Handali, J., Vlachos, M., & Meske, C. (2020). Deceptive AI Explanations: Creation and Detection. *arXiv*, 1–9. Retrieved from https://arxiv.org/abs/2001.07641. Accessed 12 Jan 2022.

Seidel, S., Kruse, L. C., Székely, N., Gau, M., & Stieger, D. (2018). Design principles for sensemaking support systems in environmental sustainability transformations. *European Journal of Information Systems, 27*(2), 221–247. https://doi.org/10.1057/s41303-017-0039-0

Shin, D., He, S., Lee, G. M., Whinston, A. B., Cetintas, S., & Lee, K.-C. (2020). Enhancing Social Media Analysis with Visual Data

Analytics: A Deep Learning Approach. *MIS Quarterly, 44*(4), 1459–1492. https://doi.org/10.25300/MISQ/2020/14870

Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russel, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. *Journal of Child Psychology and Psychiatry, 49*(4), 376–385. https://doi.org/10.1111/j.1469-7610.2001.01846.x

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., … Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology, 20*(7), 938–947. https://doi.org/10.1016/S1470-2045(19)30333-X

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human-computer collaboration for skin cancer recognition. *Nature Medicine, 26*, 1229–1234. https://doi.org/10.1038/s41591/s41591-020-0942-0

Ullmann, S., & Tomalin, M. (2020). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology, 22*, 59–80. https://doi.org/10.1007/s10676-019-09516-z

United Nations (2019). *United Nations Strategy and Plan of Action on Hate Speech*. Retrieved from https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml. Accessed 12 Jan 2022.

Vallejos, S., Alonso, D. G., Caimmi, B., Berdun, L., Armentano, M. G., & Soria, A. (2021). Mining Social Networks to Detect Traffict Incidents. *Information Systems Frontiers, 23*(1), 115–134. https://doi.org/10.1007/s10796-020-09994-3

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A Framework for Evaluation in Design Science Research. *European Journal of Information Systems, 25*, 77–89. https://doi.org/10.1057/ejis.2014.36

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Vitharana, P., Zahedi, F. M., & Hemant, J. K. (2016). Enhancing Analysts´ Mental Model for Improving Requirements Elicitation: A Two-stage Theoretical Framework and Empirical Results. *Journal of the Association for Information Systems, 17*(12), 804–840. https://doi.org/10.17705/1jais.00444

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 1–19. https://doi.org/10.1016/j.artint.2020.103404

Wang, W., & Benbasat, I. (2009). Interactive Decision Aids for Consumer Decision Making in E-Commerce: The Influence of Perceived Strategy Restrictiveness. *MIS Quarterly, 33*(2), 293–320. https://doi.org/10.2307/20650293

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users´ Moral Orientation. *Communication Research, 47*(6), 921–944. https://doi.org/10.1177/0093650219855330

Zack, M. H. (2007). The role of decision support systems in an indeterminate world. *Decision Support Systems, 43*, 1664–1674. https://doi.org/10.1016/j.dss.2006.09.003

Zhang, K. Z. K., Zhao, S. J., Cheung, C. M. K., & Lee, M. K. O. (2014). Examining the influence of online reviews on consumers´ decision-making: A heuristic-systematic model. *Decision Support Systems, 67*, 78–89. https://doi.org/10.1016/j.dss.2014.08.005

Christian Meske is is a Full Professor of Socio-technical System Design and Artificial Intelligence at the Institute of Work Science and Faculty of Mechanical Engineering at Ruhr-Universität Bochum, Germany. His research has been published in journals such as Business and Information Systems Engineering, Business Process Management Journal, Communications of the Association for Information Systems, Information Systems Frontiers, Information Systems Management, Journal of Enterprise Information Management, Journal of the Association for Information Science and Technology, or MISQ Executive. Amongst others, he has been recognized with the AIS Best Information Systems Publication of the Year Award and ICIS Paper-a-Thon Award.

Enrico Bunde is a research assistant and PhD at the Chair of Socio-technical System Design and Artificial Intelligence at Ruhr-Universität Bochum (Germany). His research with a focus on design science research and explainable artificial intelligence was published in journals such as the Information Systems Management as well as conferences such as the International Conference on Information Systems, Hawaii International Conference on Systems Sciences, or International Conference on Artificial Intelligence in Human–Computer Interaction.