The Role of Explainability in Collaborative Human-AI Disinformation Detection

Vera Schmitt
Technische Universität Berlin, German Research Center for Artificial Intelligence
Berlin, Germany
vera.schmitt@tu-berlin.de

Luis-Felipe Villa-Arenas
Deutsche Telekom, Technische Universität Berlin
Berlin, Germany

Nils Feldhus
German Research Center for Artificial Intelligence
Berlin, Germany

Joachim Meyer
Tel Aviv University
Tel Aviv, Israel

Robert P. Spang
Technische Universität Berlin
Berlin, Germany

Sebastian Möller
Technische Universität Berlin, German Research Center for Artificial Intelligence
Berlin, Germany

ABSTRACT

Manual verification has become very challenging based on the increasing volume of information shared online and the role of generative Artificial Intelligence (AI). Thus, AI systems are used to identify disinformation and deep fakes online. Previous research has shown that superior performance can be observed when combining AI and human expertise. Moreover, according to the EU AI Act, human oversight is inevitable when using AI systems in a domain where fundamental human rights, such as the right to free expression, might be affected. Thus, AI systems need to be transparent and offer sufficient explanations to be comprehensible. Much research has been done on integrating eXplainability (XAI) features to increase the transparency of AI systems; however, they lack human-centered evaluation. Additionally, the meaningfulness of explanations varies depending on users' background knowledge and individual factors. Thus, this research implements a human-centered evaluation schema to evaluate different XAI features for the collaborative human-AI disinformation detection task. Hereby, objective and subjective evaluation dimensions, such as performance, perceived usefulness, understandability, and trust in the AI system, are used to evaluate different XAI features. A user study was conducted with an overall total of 433 participants, whereas 406 crowdworkers and 27 journalists participated as experts in detecting disinformation. The results show that free-text explanations contribute to improving non-expert performance but do not influence the performance of experts. The XAI features increase the perceived usefulness, understandability, and trust in the AI system, but they can also lead crowdworkers to blindly trust the AI system when its predictions are wrong.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659031

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **HCI design and evaluation methods**; **User studies**; *Empirical studies in collaborative and social computing*; *Empirical studies in interaction design*; *Interactive systems and tools*; • **Information systems** → **Decision support systems**.

KEYWORDS

Collaborative disinformation detection, XAI, transparent AI systems, expert and lay people evaluation

ACM Reference Format:
Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Möller. 2024. The Role of Explainability in Collaborative Human-AI Disinformation Detection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3630106.3659031

1 INTRODUCTION

According to the World Economic Forum's *Global Risks Report 2024* [21], AI-generated disinformation is categorized as the second most severe risk presenting a material crisis on a global scale. With over 70 elections taking place worldwide during the super-election year of 2024, including pivotal elections like the U.S. presidential race, elections in India, and the European Parliament elections, there are already growing apprehensions about the substantial impact of AI-generated content [27]. AI is becoming a significant part of our everyday lives and significantly shaping the digital landscape also with respect to disinformation generation. This raises the potential for the creation of disinformation, deep fakes, and propagation of hate speech, thereby greatly compromising the reliability of information ecosystems [25, 29, 38, 47, 62, 73]. The *infodemic* during Covid-19 [7] and the war in Ukraine and Israel [15] serve as concrete examples of the severe effect of using AI for generating disinformation and shaping public opinions [48]. However, AI is also increasingly used for content verification to identify disinformation. Automated fake news detection is inherently challenging due to the non-binary nature of news veracity, which exists on a spectrum from clearly true to clearly false. Additionally, the categorization

of news is influenced by viewers' pre-existing beliefs and knowledge. The presence of sarcasm or irony can further complicate the interpretation of their intended meaning [29, 59]. Consequently, detecting disinformation still necessitates human judgment. Recent studies indicate that hybrid systems, combining human and machine intelligence, are capable of achieving tasks that neither could accomplish independently [24, 54]. In these scenarios, the human decision-maker typically monitors the performance of the AI system, aiding in the identification of potentially problematic news items. However, AI systems used in this context often lack sufficient transparency to rely on the predictions and recommendations. Moreover, the European Council and Parliament have tentatively agreed on the proposed AI Act, which sets out harmonized rules for artificial intelligence, including Article 13, titled "Transparency and Provision of Information to Users" in the EU AI Act [1]. Here, obligations are defined for sufficient transparency of AI systems, to enable providers and users to reasonably understand the AI System's functioning and recommendation. Thus, transparency obligations must be followed when using AI systems to detect disinformation within the EU. Additionally, the voluntary *Code of Practice on Disinformation*[2] has been developed by different stakeholders from industry, law and research to define a common ground how to handle disinformation online internationally. The AI Act and the Code of Practice on Disinformation include the requirement for transparent and comprehensive system design when AI systems are used to detect disinformation. Recent studies in the field of eXplainable AI (XAI) have shown its potential to demystify the *black box* nature of AI algorithms, thereby improving human understanding of AI-based classifications or creations [37, 63]. XAI features can foster both reliability and human trust in AI systems for the fact-checking task. Nevertheless, what constitutes a satisfactory explanation for a certain domain and task is challenging, as this heavily depends on the user's background and prior knowledge [2, 44]. This aspect is particularly pertinent, given that the AI Act also mandates the incorporation of *meaningful explanations* into AI systems for users for the fact-checking task. In this context, the Act does not explicitly define the extent of what constitutes a *meaningful explanation*, leaving space for interpretation and implementation in a variety of domains and scenarios involving AI systems.

In this work, we evaluate the *human-meaningfulness* of different types of explanations in the domain of disinformation detection. We conduct a Wizard-of-Oz study [14] with a total of 433 participants, including 406 crowdworkers and 27 journalists recruited through media outlets collaborating with the research institute, to examine the XAI features concerning their meaningfulness to domain experts and laypeople. The evaluation of the XAI features integrates both objective and subjective evaluation dimensions, such as the influence of XAI features on trust, understandability, and perceived usefulness of the explanations. To empirically examine the *meaningfulness* of explanations, we assess the influence of different types of explanations, such as highlights and free-text explanations, on objective and subjective evaluation dimensions, answering the following research questions:

RQ1: How does the use of different AI-system versions with varying explanations influence the accuracy of journalists and crowdworkers in the disinformation detection task?
RQ2: How do the subjective evaluation criteria of understandability, explanation usefulness, and trust differ between journalists and crowdworkers?
RQ3: What is the influence of media literacy on the ability to detect disinformation?
RQ4: Do the expectations towards AI change after using different versions of the AI system?

Overall, this research comprises two key contributions: (1) the assessment of human-centric explanations for identifying disinformation, aiming to gather empirical evidence on the types of explanations that are most effective for incorporation into collaborative human-AI systems designed for disinformation detection; (2) and the empirical analysis by incorporating the individual background, such as their media literacy and expectations towards AI.

2 BACKGROUND AND RELATED WORK

Within the EU, most existing approaches rely on the European Commission High-Level Expert Group (HLEG) definition by differentiating between mis- and disinformation [71]. Disinformation refers to intentionally inaccurate and false information shared by persons to cause harm or deceive the public, whereas misinformation is the unintentional spread of harmful information [51]. In this research, we adopt the definition of disinformation from the HLEG [51] and terms such as fake news and fact-checking refer to the definition of disinformation. With the rise of social media platforms like Instagram, TikTok, WeChat, and X, the spread of disinformation on social media platforms poses a major risk to societal stability and security, amplified by the widespread distribution of multimedia content on these platforms [11]. Furthermore, the proliferation of disinformation has raised concerns over decreasing trust in public institutions, weakening democratic values, and increasing political polarization [3, 65]. Hereby, the role of AI in the realm of disinformation is of a dualistic nature. On the one hand, AI can be employed to create fabricated content to spread disinformation and deep fakes. On the other hand, AI can be used to detect disinformation and identify generated content and false claims. Approaches for (partially-)automated analyses of texts, images, videos, and audio recordings to detect disinformation and aggregate the often complex results are central in the fight against disinformation and are already being used (e.g., InVid[3] and WeVerify[4]). Manual fact-checking of content shared online at a large scale is no longer feasible. Thus, research and industry have focused on developing approaches to include automated or semi-automated fact-checking and include crowdworkers for checking suspicious content, such as Facebook [5]. Current research in this field utilizes various methodologies: Propagation-based approaches focus on the spreading patterns of disinformation [36], source analysis concentrates on the origin of disinformation for early detection [8], and content-based approaches analyze linguistic features, operating under the premise that disinformation often employs deceptive language and styles [53]. In all these approaches, AI and machine

[1] Laying down Harmonised Rules on Artificial Intelligence (AI Act), 15.01.2024.
[2] *Strengthened Code of Practice on Disinformation*, 15.01.2024.
[3] https://www.invid-project.eu/
[4] https://weverify.eu/

learning models play a crucial role in identifying false, misleading, and harmful online content [5]. Within the EU, there is a significant impulse for a human-centric and ethical approach to AI, including the definition of transparency obligations for limited and high-risk categories where AI systems detecting disinformation online would fall in. The obligations include human oversight, comprehensive representation of information regarding the AI systems, and the necessity for clear and meaningful explanations.

In the context of explainable fact-checking, various approaches can be used to provide justifications for the decisions made by a model. These include applying feature importance methods to identify key elements, employing logic-based systems for structured reasoning, or leveraging rationalizing models that generate explanations in natural language [24]. Evaluations of text-based explanations have demonstrated that for specific tasks, explanations generated by AI can enhance human comprehension, and increase trust and confidence in the AI system [33]. The focus on human-centric aspects of explainability has ignited increasing interest among scholars in diverse fields to delve into the realm of XAI [12]. The effectiveness of explanations, specifically in terms of enhancing comprehension and trust in the underlying AI system, is contingent on the receiver's perception and understanding. This suggests that choices made in designing AI systems and respective explanations need to take human perceived understandability and trust into account [34]. While explanations can increase users' comprehension of and confidence in machine learning systems [39], certain researchers argue that the mere availability of explanations, regardless of their content, can induce these effects [18], potentially leading to an illusory sense of understanding and trust. Additionally, there is a natural human inclination towards simpler explanations, which might lead to the preference for systems with more compelling explanatory outputs over those that are more transparent [26]. However, the meaningfulness of different types of explanations depends on the user's prior beliefs, knowledge, and understanding of the underlying AI system [37, 39]. Overall, previous research lacks standardized evaluation approaches covering relevant dimensions of human-centered XAI evaluation by considering individuals' prior knowledge and expectations of the AI system [37]. It is essential to assess the explanations and artifacts concerning a specific AI system before, during, and after the system's deployment to assess the effect of XAI features in different stages of use [39].

Regarding studies closely related to ours, in Linder et al. [35], participants were tasked to assess news statements and learn to predict the output of the AI system. They found that the level of detail of an explanation improves the utility for understanding the system, but this comes at time and attention expenses needed to make sense of the explanation. While their explanations are example-based and attribute-based, we focus on salient feature explanations including emotionally charged content and free-text explanations. We emphasize the strengths of human-AI collaboration efforts, while Linder et al. [35] are more concerned with the cognitive load of explanations. Mohseni et al. [45] asked if explanations help end users share more credible news. They found that explanations helped users build appropriate mental models of the XAI assistant embedded in a news review platform and adjust trust according to their perceived limitations of the system. They share salient highlight explanations with our setup through attention visualization and come to the same conclusion that explanations are, however, not responsible for improved task performance. While they pointed out a misalignment between explanations and meaningful logic for humans, we find in our study that free-text explanations can close this gap. Nguyen et al. [49] investigated the human-AI collaboration for fake news detection and showed that task performance (assessing claims) did improve when exposed to correct system predictions, which we can also confirm. The exposure to wrong system predictions often led to lower task performance, but if given the option to interact with such incorrect predictions users were able to get better. Lastly, Epstein et al. [19] analyzed the likelihood of users sharing content and found that fake news warnings are made more effective through explanations. However, self-reported trust in the warning labels was not increased by them.

3 METHODOLOGICAL BACKGROUND

In this paper, we focus on the human-centered aspects of XAI evaluation, aiming to first determine which types of explanations are meaningful for human participation in the joint human-AI task of information verification. Therefore, we present an XAI evaluation framework that encompasses important human-centric evaluation aspects, as detailed by Longo et al. [37] and Lopes et al. [39], within the scope of collaborative human-AI disinformation detection. The following explains the different types of explanations and objective and subjective evaluation dimensions to empirically test if the explanations help human users detect disinformation more accurately.

3.1 Types of Explanations

Since most of the current research in the XAI domain lacks human-meaningful explanations [17, 37], we focus on types of explanations that are more accessible to human recipients. Explanations that are not comprehensible and understandable for users can lead to lower trust and overall performance in *human-in-the-loop* settings [39]. In our approach, we adopt the classification system from Wiegreffe and Marasovic [72], differentiating explanations into three categories: *highlights*, *free-text*, and *structured explanations*.

3.1.1 Highlights. Building on the foundations of automated fact-checking [42, 60, 64, 68], our work involves annotating highlights that act as evidence for the credibility (or 'Truthfulness') of news articles. These highlights, added manually by the authors of this paper, play a crucial role in the collaborative human-AI fact-checking task indicating check-worthy assertions. Moreover, highlights are further used to capture emotion, sentiments, and the writing style. Previous work has shown, that manipulative or emotionally loaded language often affect the objectivity and the integrity of content [1, 41, 55][5].

3.1.2 Free-text Explanations. The second type of explanation that previous research has shown to be effective is free-text explanations [31, 37], which are independent natural language explanations

[5]Figure 7 in the Appendix shows an example of the two types of highlights, where the highlights in yellow indicate the truthfulness justification and the emotional content is highlighted in the color cyan, to better differentiate between the two types of highlights.

displayed next to the original text to give a short and comprehensive explanation of the model's reasoning for fact-checking the given news article. Expanding upon existing fact-checking datasets [31, 57], our approach involves gathering justifications for classifying a news article as either true or false. This process may involve applying commonsense reasoning and rephrasing information from external sources that are not referenced in the original article. Expert annotations are needed for these rationales. Relevant data for this purpose can be sourced from introductory summaries on fact-checking websites, like Snopes[6], which is an accredited signatory of the International Fact-Checking Network (IFCN)[7]. In our work, the explanations represent an idealistic version of what a rationale-generating fact-checking system [6, 16, 52] might offer in the future. Our data sources are entirely human-generated, similar to the training data used in language models that drive these fact-checking systems. Our focus is on determining the types of explanations that are both useful and trustworthy in a practical application such as the human-AI collaborative fact-checking task[8].

3.1.3 Readability. In addition to the two types of explanations described above, we consider the readability, which has been examined in previous literature, where the assumption is that news articles featuring longer sentences and scientific terminology are more frequently found in reliable outlets compared to those in untrustworthy and satirical websites [20, 31]. While Allen et al. [4] demonstrated that readability is not always a critical linguistic characteristic in determining how misleading tweets are and may not be an essential component of XAI systems, it remains an important extrinsic feature like style or an article's metadata [55, 69]. Rather than a stand-alone explanation, we use it in combination with highlights to balance out the information load for each AI system version (§4.2). In our work, the *Flesch-Kincaid Grade-Level* score provides a readability assessment of the news articles. The scores were mapped onto human-readable categorical features (< 10 = *Easy*; $10 - 12.5$ = *Medium*; > 12.5 = *Hard*)[9].

3.2 Human-Centered XAI Evaluation

In Lopes et al. [39], XAI features are evaluated based on several human-centered objective and subjective dimensions. The four evaluation dimensions — *performance*, *understandability*, *explanation usefulness*, and *trust* — are described purely in theoretical terms, without being subjected to empirical testing or validation. Our study employs their theoretical framework and we select adequate constructs based on previous interdisciplinary research in the domain of HCI and XAI [39, 46, 67, 74] allowing for an empirical analysis of the four dimensions.

3.2.1 Performance. In verifying the effectiveness of an AI system's support in a human-assigned task, task performance is an important objective metric [56]. Specifically, in the context of information verification, task performance is measured by the proficiency with which humans can identify fake news using the information provided. Therefore, the effectiveness of explanations can be objectively assessed based on whether their integration into the system enhances performance. Additionally, the user's ability to anticipate the model's output, termed simulatability or human forward prediction [30], is assessed before they view the actual system prediction for a specific news item. The performance is then measured a second time after the user has been presented with the AI model's truthfulness rating and the additional information depending on the AI-system versions. Performance is hereby calculated on an individual basis for each participant, considering the order of news items presented, to accurately measure task performance for different types of news items.

3.2.2 Understandability. Understandability is an important subjective evaluation dimension evaluating explanations across various contexts. It refers to the ability to "describe the relationship between a system's input and output in relation to its parameters" [23]. Thus, understandability may pertain to the specific task at hand, the comprehension of the AI system's underlying mechanism, or the clarity of the explanations provided. In our study, understandability specifically relates to understanding the explanations, which in turn aids in understanding the AI system's predictions. Hereby, understandability is measured according to Nourani et al. [50], where the perceived understandability is measured *locally* for each news item, and also *globally* for the whole AI system.

3.2.3 Explanation Usefulness. The third dimension, explanation usefulness, is crucial for determining whether the added explanations contribute to enhancing the performance of a given task [13]. Explanation usefulness can serve as an objective measure by comparing user performance across AI systems with varying explanations and assessing whether the ability to detect fake news improves when explanations are provided. Additionally, the perceived usefulness of explanations can be measured subjectively to gain further insights into the influence of different explanations. Thus, the subjective perception of explanation usefulness is assessed both locally after each interaction with the AI system and globally in the context of the overall evaluation of the AI system.

3.2.4 Trust. Establishing user trust is essential for the effective implementation of any AI system, as it significantly influences user willingness to use the AI system [10]. Trust pertains to the degree of confidence and ease users feel when using an AI system for a practical task. Hereby, trust can be measured objectively through the preference of the AI's prediction over their initial judgment, which is described in Appendix B in more detail. Moreover, participants' perceived trust can be measured with a questionnaire presented locally after each news item and globally to assess the trust in the AI system at the end of the experiment. Wanner et al. [70] provides a questionnaire allowing us to measure trust, which has been slightly adapted to the context of our experiment. As described above, we used two falsely labeled (FP, FN) news items to observe *blind* trust, when AI predictions were wrong. The three subjective dimensions were measured with questionnaires consisting of 7-point Likert scales. Following the recommendations of Mohseni et al. [45], we measured them locally after each news item and in a more extensive

[6] https://www.snopes.com/
[7] https://www.poynter.org/ifcn/
[8] An example of the free-text explanation can be seen in Figure 7 in the Appendix highlighted in pink.
[9] An example of the integration of the readability categories can be seen in Figure 7 in the Appendix.

manner in the end after using the AI system for content verification to assess the subjective evaluation dimensions in more detail.

4 EXPERIMENTAL SETUP

Our experimental setup includes the selection of news items, the description of the AI-system versions, and further constructs that are relevant for evaluating explanations. At the end of this section, we describe our workflow.

4.1 News Items

We selected eight news items on three commonly known topic domains, including environment, crime, and gender-related topics, from accredited debunking websites Snopes[10] and Politifact[11]. The news items were assessed based on a continuous truthfulness rating on a scale between 0 and 100%. The news items were manually labeled by several experts from media and computer science. Three types of news items were defined: (1) The clearly true and clearly fake news items, which were correctly identified by the AI system (two articles), (2) the rather true or fake news items that contain some false and misleading content correctly labeled by the AI system (four articles), and (3) falsely labeled news items to examine, if the users, crowdworkers and journalist alike, followed the AI recommendations even though it was wrong (two articles). Transparency can also have a deceiving effect on the end-user by distorting information about its competence [9]. Therefore, we tested for *blind* trust by incorporating two falsely labeled news items on purpose. One example was a *false positive* (FP) where the news item was truthful, but the AI system presented the news item to the user as fake; and one *false negative* (FN) example, where the AI system presented a news item to the user as true, but it was fake. For these two examples, two clearly fake or true news items were chosen to facilitate the detection of the FN and FP news items. Moreover, the truthfulness scores provided by the AI system were used to form three subcategories for accurate predictions: *clearly fake* (< 25), *rather fake* ($25 \leq$ and < 50), *rather true* ($50 \leq$ and < 75), and *clearly true* (≤ 75).

4.2 AI-System Versions

Our approach involves evaluating the explanations independently, facilitating a comparative analysis across the evaluation of the XAI features. We introduce three distinct versions of a simulated AI system, each demonstrating varied explainability features.[12] This strategy enabled us to effectively evaluate the influence of explanations on the task of detecting fake news. (1) **V1 - Baseline Version:** Displays only the news item, its metadata, and the AI system's truthfulness rating, without including any explanations, (2) **V2 - Salient Explanations Version:** Includes the content from version 1, adds the readability feature and incorporates the two types of highlights, and (3) **V3 - Free-Text Explanation Version:** Encompasses all information from version 1 and integrates free-text explanations. In our study, we evaluate the influence of the integrated explanation types for the content verification task, where the control group received **V1** none of the described XAI features. To maximize resource efficiency, we incorporated explanations into two specific versions of the AI system tested in two separate experimental groups. One experimental group received **V2** containing the two types of highlights and the readability assessment of the news article, and the other experimental group received **V3** only containing the free text explanations. The comparison of the different groups allowed us to empirically observe the influence XAI features have on the evaluation dimensions and the perceptions of crowdworkers and journalists. The three different versions were then evaluated based on the four evaluation dimensions described below. The information provided above was integrated into the **News Verification Dashboard** which we have developed and designed specifically for this experiment. The dashboard included the news article indicating the topic domain, the AI truthfulness rating, the publishing date, and information about the source as basic features included in all three versions. For V2, the text highlights and salient features (displayed in yellow and cyan) were included, and in V3, the free-text explanations (visualized in pink). A screenshot of the News Verification Dashboard with all three versions can be found in Appendix A.

4.3 Media Literacy and Expectations about AI

According to Jones-Jang et al. [28], having higher media literacy significantly contributes to better fake news detection abilities, whereas less literate people have a hard time identifying fake or misleading content. We adopted the questionnaire of Shahzad et al. [61] (Appendix I.3.3) to examine if higher media literacy helps correctly identify false content, especially concerning falsely labeled news items. We also used a questionnaire to evaluate general expectations towards using AI for detecting fake news (Appendix I.3.2). The questionnaire was administered both at the start and the end of the study, enabling a comparison of shifts in expectations toward AI's role in fake news detection among various groups and user demographics. The evaluation of the expectations towards AI has been highlighted by previous research Venkatesh et al. [66], to identify the underlying factors that influence user acceptance and adoption of new technologies. The questions were adopted from previous research [22, 43, 66] to fit the fake news detection context. Participants' expectations were again measured with Likert scales.

4.4 Experimental Workflow

The experiment consisted of three parts: (1) A survey asked about general attitudes towards AI and media literacy questions. (2) The experimental part started by introducing the News Verification Dashboard and the human-AI collaborative task of detecting disinformation. The eight news items were shown in a randomized order, to avoid any order effects. Also, the AI-system versions were randomized for each participant so that equal numbers of participants received the different AI-system versions. After each news item, local evaluations of the four dimensions were given. (3) Another survey was conducted at the end of the experimental part, including questions about the expectations of AI to verify if a change occurred after using the AI system, the global evaluation of the AI system, and demographic questions. Figure 1 displays a visual

[10] https://www.snopes.com/
[11] https://www.politifact.com/
[12] Including every possible combination of explanation types in separate versions would require extensive resources.

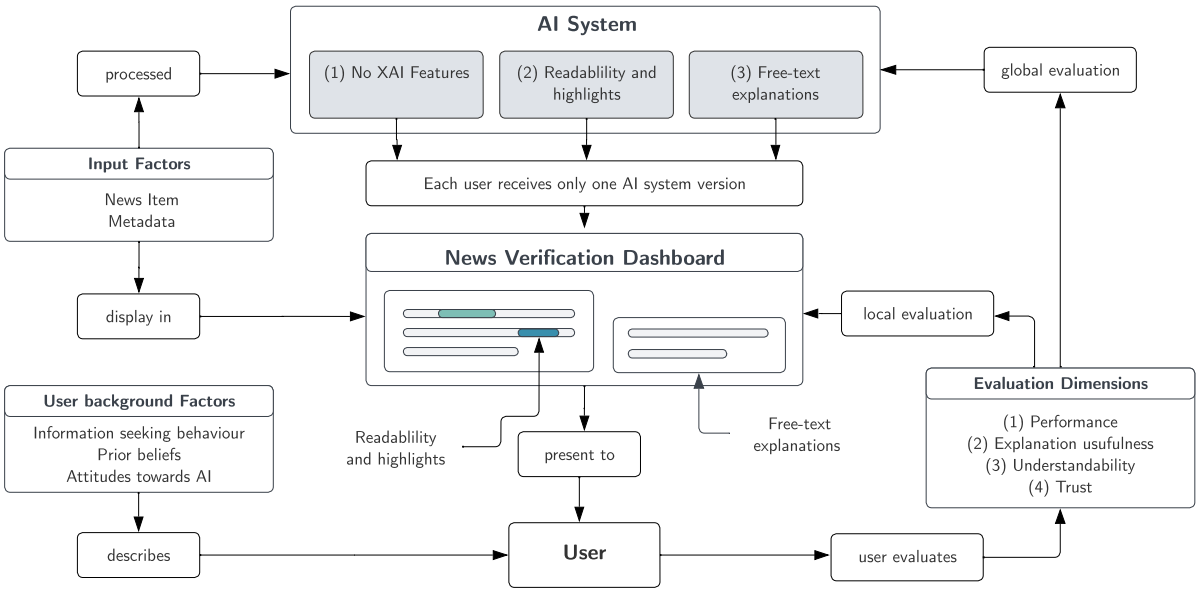


Figure 1: Overview of the different components and constructs used for human-centered XAI evaluation for disinformation detection.

representation of the different aspects of the experiment. The experiment platform is described in detail in Schmitt et al. [58] and is also available open-source[13].

5 EMPIRICAL COMPARISON OF XAI EVALUATION BETWEEN JOURNALISTS AND LAY-PEOPLE

Overall, 600 participants took part in the experiment through the crowdsourcing platform Crowdee[14]. Among them, 167 participants were excluded after failing to correctly answer the screening questions. The remaining 433 participants, who answered test questions correctly, were included in the explanation assessment, including 406 crowdworkers with 27 journalists recruited through media outlets collaborating with the research institute. These participants were randomly divided into three groups, each interacting with a different AI-system version: 134 laypeople with 8 journalist in baseline version 1 without explanations, 140 laypeople with 9 journalist in version 2 with saliency explanations, and 132 laypeople with 10 journalist in version 3 with free-text explanations. All constructs used for the local and global evaluation dimensions for the evaluation of the AI-system versions showed sufficient reliability by having Cronbach's $\alpha > 0.7$ (Appendix D).

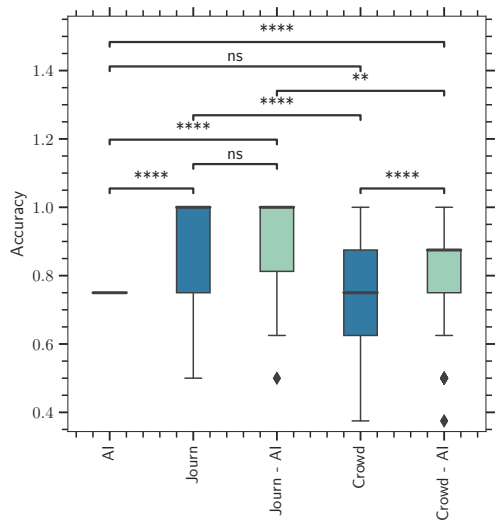
5.1 Evaluation of AI-system Versions based on the Four Evaluation Dimensions

5.1.1 Performance Evaluation. The performance comparison between the journalists, crowdworkers, and the AI system is based on the objective accuracy metric (AI system 0.75, crowdworker alone 0.73, crowdworker-AI 0.82, journalists alone 0.88, journalists-AI 0.89). For the comparison of the accuracy between the AI system, journalists, crowdworkers, and the human-AI collaboration setups, we used a Kruskal-Wallis Test (with Dunn post-hoc comparison) (Figure 2a)[15]. The journalists' accuracy is significantly higher compared to the AI (*H-statistic* 179.6, *p-value* < 0.01), and the crowdworker accuracy (*H-statistic* 19.36, *p-value* < 0.01). Moreover, the journalists' accuracy did not increase when using the AI system (*H-statistic* 0.13, *p-value* 0.72). However, when the crowdworkers used the AI recommendations, the accuracy significantly increased (*H-statistic* 61.43, *p-value* < 0.01), and even outperformed the AI accuracy alone (*H-statistic* 115.9, *p-value* < 0.01). However, the crowdworkers' accuracy without the AI system was slightly (but not significantly) lower than the accuracy of the AI system alone (*H-statistic* 3.36, *p-value* 0.07). Overall, the AI system was outperformed by the journalists and the crowdworker-AI accuracy (*H-statistic* 8.40, *p-value* 0.01).

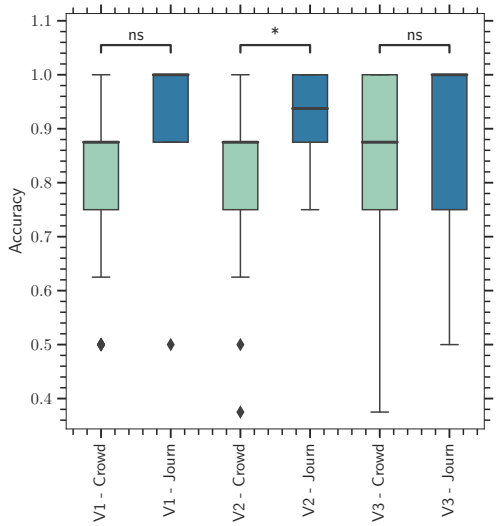
When comparing the accuracy over different AI-system versions, the following differences could be identified after applying a Mann-Whitney U-Test (Figure 2b). When using the basic AI-system

[13]The *XAI Evaluation Framework* is made open-source: https://github.com/news-polygraph/XAI-Evaluation-Framework.git.
[14]https://www.crowdee.com/

[15]Hereby, the significant levels are indicated by: ns - not significant, * 0.01 < p <= 0.05, ** 0.001 < p <= 0.01, *** 0.0001 < p <= 0.001, and **** p <= 0.0001.



(a) Accuracy comparison journalists, crowdworkers and, the AI system.



(b) Accuracy comparison between journalists and crowdworkers for V1, V2, and V3.

Figure 2: Accuracy comparison between journalists and crowdworkers.

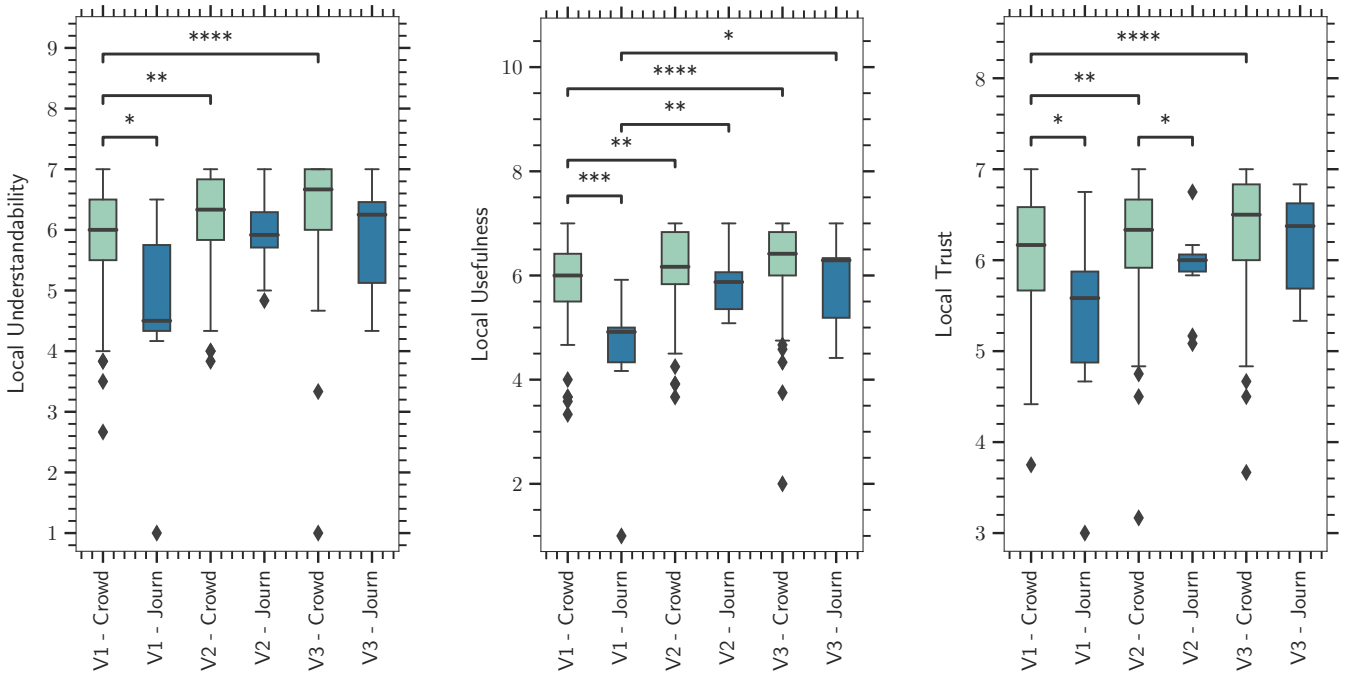
version, journalists had higher accuracy compared to crowdworkers. However, this difference was not statistically significant (*U-statistics* 274.5, *p-value* 0.06). However, journalists had significantly higher accuracy ratings when salient explanation features were present (*U-statistics* 410, *p-value* 0.02). When free-text explanations were given, the crowdworkers' accuracy increased and there was no significant difference between journalists and crowdworkers (*U-statistics* 543, *p-value* 0.32).

Overall, explanations supported the crowdworkers' performance the most. With free-text explanations, the crowdworkers even achieved similar high performance as journalists.

5.1.2 Understandability. The comparison between crowdworkers and journalists for the perceived understandability of the explanations 3a reveals that the only significant differences were in the basic version between journalists and crowdworkers (*U-statistics* 715, *p-value* 0.02), where crowdworkers showed a significantly higher perceived understandability of the presented content in V1 compared to journalists. Moreover, there were significant differences between the three different AI-system versions for the crowdworkers. The perceived understandability significantly increased when salient explanation features were present (*U-statistics* 7470, *p-value* 0.01) and was the highest when a free-text explanation was shown to the crowdworkers (*U-statistics* 5952, *p-value* < 0.01). In comparison to the global evaluation (Appendix E) in Figure 8a, a significant difference was also found for journalists between V1 and V2 and V1 and V3 because of significantly higher understandability ratings on a global level for V2 and V3, respectively. For the crowdworkers, only V1 and V3 show significant differences, with V3 showing significantly higher understandability.

5.1.3 Explanation Usefulness. When comparing the subjective explanation usefulness locally (Figure 3b), the comparison between crowdworkers and journalists reveals significant differences for V1 (*U-statistics* 819.5, *p-value* < 0.01), where crowdworkers reported higher explanation usefulness. Crowdworkers also reported higher explanation usefulness for V2 and V3, but the differences were not significant. Regarding the different AI-system versions for only the crowdworkers, significantly higher explanation usefulness could be observed for V2 in comparison to V1 (*U-statistics* 7525, *p-value* < 0.01), and for V3 in comparison to V1 (*U-statistics* 5921, *p-value* < 0.01). Similar differences could be observed for the journalists: They reported significantly higher explanation usefulness for V2 in comparison to V1 (*U-statistics* 5.5, *p-value* < 0.01), and for V3 in comparison to V1 (*U-statistics* 11.5, *p-value* 0.02). For local vs. global evaluation (Appendix E) in Figure 8b, similar results can be observed as the perceived global explanation usefulness was significantly higher for the crowdworkers with V2 compared to V1, and V3 compared to V1. There were no significant differences between the crowdworkers and the journalists for the different AI-system versions, indicating that the explanation usefulness follows a similar trend for both journalists and crowdworkers.

5.1.4 Trust. For the perceived trust in the AI system (Figure 3c), significant differences exist between journalists and crowdworkers for V1, where journalists reported significantly lower trust compared to the crowdworkers (*U-statistics* 686, *p-value* 0.03). A similar difference existed for V2 (*U-statistics* 973, *p-value* 0.04). For the free-text explanations, the crowdworkers showed a slightly higher, but not significant, trust rating. When comparing the different AI-system versions for the crowdworkers, trust increased significantly when comparing V2 to V1 (*U-statistics* 7576, *p-value* < 0.01) and V3 to V1 (*U-statistics* 6220, *p-value* < 0.01), where the explanations significantly contributed to higher trust. When comparing the different AI-system versions for the journalists, increasing trust could be observed for V2 and V3 in comparison to V1. However, the differences were not significant. When comparing the local and



(a) Local understandability. **(b) Local usefulness.** **(c) Local trust.**

Figure 3: Comparison of local explanation understandability, usefulness and trust between journalists and crowdworkers for different AI-system versions.

global evaluations (Appendix E) in Figure 8c, similar results could be identified, whereas the differences between V2 and V3 in comparison to V1 showed significant differences also for the journalists. This indicates that the explanations significantly increased trust for both journalists and crowdworkers. However, even though free-text explanations showed higher trust ratings locally and globally, the differences were not significant, and thus no further differentiation could be made between salient and free-text explanations.

5.1.5 Blind Trust. To examine *blind* trust in AI recommendations, we compare the alignment of the journalists (Figure 4a) and crowdworkers (Figure 4b) truthfulness rating with the AI rating for the different types of news items. For both groups, the alignment significantly decreased (after conducting a Dunn test with the Holm-Bonferroni method, $p - value :< 0.01$) when presented with wrong AI system predictions for both falsely labeled news items (Appendix F). When considering the AI system alignment for the falsely labeled news items, journalists showed much lower alignment in comparison to crowdworkers. This indicates that journalists were less prone to *blind* trust compared to crowdworkers.

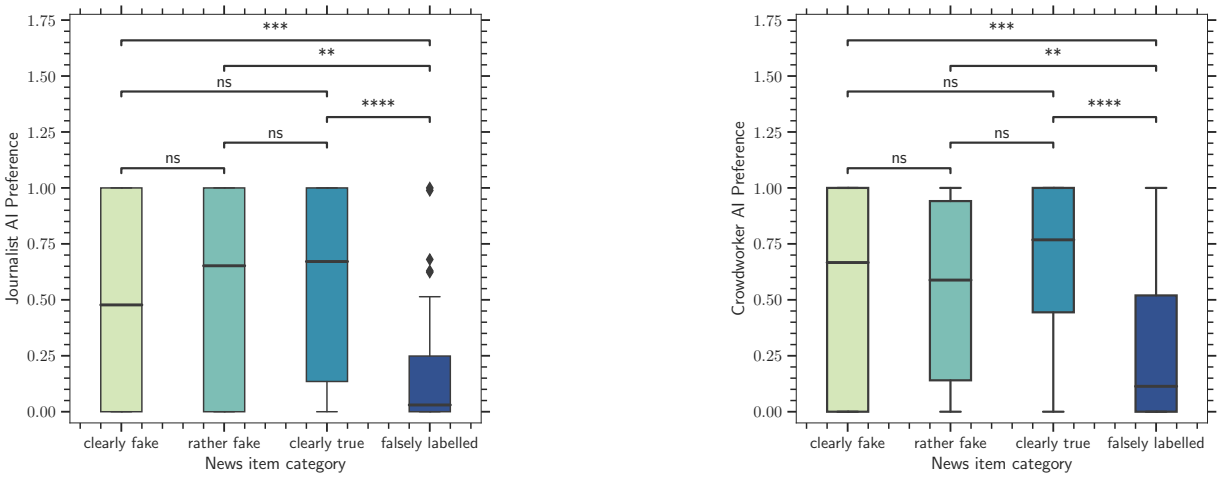
5.2 Media Literacy

The media literacy score was computed by averaging user skills assessed in the survey, encompassing source evaluation, information search, and ethical use of information (Appendix I.3.3). We use literacy score quantiles to establish three categories (*low literacy*, *medium literacy* and *high literacy*)[16] for evaluating the impact of literacy on accuracy. Figure 5a displays the accuracy of both crowdworkers and journalists without AI system support on the media literacy categories. The accuracy of crowdworkers remained similar among all literacy categories, consistently falling behind the accuracy achieved by journalists with similar literacy scores. Likewise, Figure 5b shows their accuracies with the support of the XAI features across various media literacy categories. An observable trend in both groups is an increment in accuracy with an increase in literacy. When comparing the accuracy in relation to media literacy for the falsely labeled and the hard-to-detect news items (Appendix H), the results indicate that high literacy levels are associated with a reduced tendency towards blind trust in the case of journalists.

5.3 Expectations towards AI

The shift in expectations regarding AI before and after using the AI system is illustrated in Figure 6a for crowdworkers and in Figure 6b for journalists. In both instances, the most significant alteration in expectations is associated with the *free-text explanations* (V3). There is a noteworthy increase in expectations, especially notable for crowdworkers, with statistical significance (*U-statistics* 7528, *p-value* 0.03). Additionally, although not achieving statistical significance, journalists also exhibit a rise in the median value.

[16] $Q_{lower} = F_{lit}^{-1}(0.25)$ *and* $Q_{higher} = F_{lit}^{-1}(0.75)$*); Low Literacy* $< Q_{lower}$, $Q_{lower} \leq$ *Medium Literacy* $< Q_{higher}$, and *High Literacy* $\geq Q_{higher}$



(a) Journalist. **(b) Crowdworker.**

Figure 4: AI preference for news item categories for crowdworkers and journalists.

6 DISCUSSION AND CONCLUSION

We present a framework for evaluating human-centric explanations in human-AI collaborative information verification systems. It assesses the impact of different explanation types on performance, understandability, usefulness, and trust. We applied our framework to a practical setting through the News Verification Dashboard and evaluated its effectiveness with a crowdsourcing study involving 433 participants including 406 crowdworkers and 27 journalists. Our analysis provides answers to the research questions outlined in Section 1.

RQ1: Influence of AI systems on the accuracy of journalists and crowdworker: With AI support, the accuracy of crowdworkers significantly improved, rising from 0.73 to 0.82. For journalists, there was a minor, insignificant increase in accuracy from 0.88 to 0.89. Crowdworkers' accuracy significantly increased with free-text explanations, almost matching that of the journalists. Overall, the AI system notably enhanced laypeople's performance in detecting fake news, bringing it close to the expert level with free-text explanations. However, the performance of expert journalists only improved marginally and was unaffected by the different AI-system versions. Therefore, AI assistance significantly improved laypeople's ability in content verification tasks.

RQ2: Difference of subjective evaluation criteria between journalists and crowdworkers: Trust, understandability, and explanation usefulness increased in the order of V1, V2, and V3 for journalists and crowdworkers for the local and global evaluations alike. However, only V3, displaying the free-text explanations, showed significantly higher understandability, explanation usefulness, and trust in comparison to V1. Nonetheless, V3 also led to overreliance and *blind* trust in the AI system, especially when crowdworkers use it.

RQ3: Influence of media literacy on accuracy: Journalists had significantly higher media literacy and showed better ability in detecting fake news when AI system predictions were wrong, and when fake news content is harder to detect. Thus, higher literacy can act as a countermeasure to overreliance and *blind* trust in AI systems.

RQ4: Expectations towards AI: The use of the AI-system versions changed the expectations towards AI, especially with V3 for the crowdworkers, where a significant difference can be observed, and also for journalists with a non-significant increase. This indicates that the free-text explanations support a more positive expectation towards AI.

Overall, the key contribution of our study is the empirical evaluation of various human-centered explanations for human-AI content verification tasks. We consider the previously identified issue of human overreliance on AI systems, especially when the AI is incorrect [19, 35, 45]. We observe similar overtrust in AI among crowdworkers when AI ratings are incorrect, but this effect can be minimized with expert knowledge. Finally, we can confirm the findings of Nguyen et al. [49] that human-AI collaboration outperforms human and AI performance alone for the content verification task, especially when no expert knowledge is available. With free-text explanations, laypeople (crowdworkers) even achieved similar performance comparable to experts (journalists). Additionally, explanations in our News Verification Dashboard improved perceived usefulness, understandability, and trust in the AI system compared to having no explanations. In conclusion, our findings underscore the importance of human-AI collaboration in content verification tasks, particularly in the absence of expert knowledge. Especially in the sensitive area of disinformation detection, which can conflict with free speech rights, human-AI collaboration is essential to address the widespread online disinformation while ensuring human oversight to align with fundamental human rights, as mandated by the AI Act. Further research is needed to balance AI system transparency with avoiding information overload and overreliance on the AI system. Moreover, future research can explore explanations at various levels of abstraction to offer appropriate transparency for different expertise levels and understanding of the tasks.

ETHICAL STATEMENT

The experiment prioritized participant privacy, excluding the collection of personal information. Each participant was assigned a unique ID for identification, ensuring anonymity. Participants received detailed information about data collection and processing and gave explicit consent to participate. They were also informed of their right to withdraw at any point, which would lead to the deletion of their data. The study was approved by the ethics committee of the Technische Universität Berlin, and it faced no further ethical objections or requirements.

ACKNOWLEDGMENTS

We thank Balázs Patrik Csomor and Anna Mockenhaupt for their support in preparing and conducting the experiment. This research is funded by the Federal Ministry of Education and Research (BMBF, reference: 03RU2U151C) in the scope of the research project newspolygraph.

REFERENCES

[1] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Association for Computational Linguistics, Brussels, Belgium, 85–90. https://doi.org/10.18653/v1/W18-5513

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805.

[3] Jennifer Allen, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* 6, 14 (2020), eaay3539. https://doi.org/10.1126/sciadv.aay3539

[4] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 245, 19 pages. https://doi.org/10.1145/3491102.3502040

[5] Wissam Antoun, Fady Baly, Rim Achour, Amir Hussein, and Hazem Hajj. 2020. State of the Art Models for Fake News Detection Tasks. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. 519–524. https://doi.org/10.1109/ICIoT48696.2020.9089487

[6] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7352–7364. https://doi.org/10.18653/v1/2020.acl-main.656

[7] Vimala Balakrishnan, Wei Zhen Ng, Mun Chong Soo, Gan Joo Han, and Choon Jiat Lee. 2022. Infodemic and fake news – A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction* 78 (2022), 103144. https://doi.org/10.1016/j.ijdrr.2022.103144

[8] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting Factuality of Reporting and Bias of News Media Sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3528–3539. https://doi.org/10.18653/v1/D18-1389

[9] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (apr 2023), 17 pages. https://doi.org/10.1145/3579460

[10] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[11] Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences* 497 (2019), 38–55. https://doi.org/10.1016/j.ins.2019.05.035

[12] Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-Centered Evaluation of Explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*. Association for Computational Linguistics, Seattle, United States, 26–32. https://doi.org/10.18653/v1/2022.naacl-tutorials.4

[13] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174098

[14] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies: why and how. In *Proceedings of the 1st international conference on Intelligent user interfaces*. 193–200. https://doi.org/10.1145/169891.169968

[15] Omar Darwish, Yahya Tashtoush, Majdi Maabreh, Rana Al-essa, Ruba Aln'uman, Ammar Alqublan, Munther Abualkibash, and Mahmoud Elkhodr. 2023. Identifying Fake News in the Russian-Ukrainian Conflict Using Machine Learning. In *Advanced Information Networking and Applications: Proceedings of the 37th International Conference on Advanced Information Networking and Applications (AINA-2023), Volume 3*. Springer, 546–557. https://doi.org/10.1007/978-3-031-28694-0_51

[16] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information processing & management* 60, 2 (2023), 103219. https://doi.org/10.1016/j.ipm.2022.103219

[17] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[18] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312787

[19] Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings? *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (May 2022), 183–193. https://doi.org/10.1609/icwsm.v16i1.19283

[20] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A Comparison of Features for Automatic Readability Assessment. In *Coling 2010: Posters*. Coling 2010 Organizing Committee, Beijing, China, 276–284. https://aclanthology.org/C10-2032

[21] World Economic Forum. 2024. Global Risks 2024: At a Turning Point. https://www.weforum.org/publications/global-risks-report-2024/in-full/global-risks-2024-at-a-turning-point/#global-risks-2024-at-a-turning-point

[22] Nicole Gillespie, Steve Lockey, and Caitlin Curtis. 2021. Trust in artificial intelligence: A five country study. (2021). https://doi.org/10.14264/e34bfa3

[23] Michael Gleicher. 2016. A framework for considering comprehensibility in modeling. *Big data* 4, 2 (2016), 75–88. https://doi.org/10.1089/big.2016.0007

[24] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (02 2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[25] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and other Large Generative AI Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1112–1123. https://doi.org/10.1145/3593013.3594067

[26] Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *arXiv* abs/1711.07414 (2017). https://arxiv.org/abs/1711.07414

[27] Dudi ISKANDAR, Indah SURYAWATI, Geri SURATNO, Liliyana LILIYANA, Muhtadi MUHTADI, and Ngimadudin NGIMADUDIN. 2023. Public Communication Model In Combating Hoaxes And Fake News In Ahead Of The 2024 General Election. *International Journal of Environmental, Sustainability, and Social Science* 4, 5 (2023), 1505–1518.

[28] S Mo Jones-Jang, Tara Mortensen, and Jingjing Liu. 2021. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American behavioral scientist* 65, 2 (2021), 371–388. https://doi.org/10.1177/0002764219869406

[29] Razieh Khamsehashari, Vera Schmitt, Tim Polzehl, Salar Mohtaj, and Sebastian Moeller. 2023. How Risky is Multimodal Fake News Detection? A Review of Cross-Modal Learning Approaches under EU AI Act Constrains. In *Proc. 2023 ISCA Symposium on Security and Privacy in Speech Communication*. 47–51. https://doi.org/10.21437/SPSC.2023-1

[30] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and

R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

[31] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7740–7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

[32] Philipp Kulms and Stefan Kopp. 2018. A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making With Computers. *Front. Digit. Humanit.* 5 (June 2018), 352444. https://doi.org/10.3389/fdigh.2018.00014

[33] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5195–5205. https://doi.org/10.18653/v1/D19-1523

[34] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv* abs/2110.10790 (2021). https://arxiv.org/abs/2110.10790

[35] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K. Pentyala, Eric D. Ragan, and Xia Ben Hu. 2021. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters* 2, 4 (2021), e49. https://doi.org/10.1002/ail2.49

[36] Yang Liu and Yi-Fang Wu. 2018. Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1. https://doi.org/10.1609/aaai.v32i1.11268

[37] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301. https://doi.org/10.1016/j.inffus.2024.102301

[38] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from Generative Artificial Intelligence Is Believed Less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (, Seoul, Republic of Korea,) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/3531146.3533077

[39] Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Appl. Sci.* 12, 19 (Sept. 2022), 9423. https://doi.org/10.3390/app12199423

[40] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b8eda9593fbcb63b7ced1866853d9622737533a2

[41] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A Survey on Computational Propaganda Detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4826–4832. https://doi.org/10.24963/ijcai.2020/672 Survey track.

[42] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14867–14875. https://doi.org/10.1609/aaai.v35i17.17745

[43] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on management information systems (TMIS)* 2, 2 (2011), 1–25. https://doi.org/10.1145/1985347.1985353

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

[45] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. 2021. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 421–431. https://doi.org/10.1609/icwsm.v15i1.18072

[46] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3–4 (Sept. 2021), 1–45. https://doi.org/10.1145/3387166

[47] Salar Mohtaj, Ata Nizamoglu, Charlott Sahitaj, Premtim Jakob, Sebastian Möller, and Vera Schmitt. 2024. NewsPolyML: Multi-lingual European News Fake Assessment Dataset *(MAD '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3643491.3660290

[48] Linda Monsees. 2023. Information disorder, fake news and the future of democracy. *Globalizations* 20, 1 (2023), 153–168. https://doi.org/10.1080/14747731.2021.1927470

[49] An T. Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C. Wallace, and Matthew Lease. 2018. Believe It or Not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. Association for Computing Machinery, New York, NY, USA, 189–199. https://doi.org/10.1145/3242587.3242666

[50] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105. https://doi.org/10.1609/hcomp.v7i1.5284

[51] High Level Expert Group on Fake News and Online Disinformation. 2018. Report to the European Commission on A Multi-Dimensional Approach to Disinformation. (2018). https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation

[52] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 6981–7004. https://doi.org/10.18653/v1/2023.acl-long.386

[53] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3391–3401. https://aclanthology.org/C18-1287

[54] Tim Polzehl, Vera Schmitt, Nils Feldhus, Joachim Meyer, and Sebastian Möller. 2023. Fighting Disinformation: Overview of Recent AI-Based Collaborative Human-Computer Interaction for Intelligent Decision Support Systems. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP,*. INSTICC, SciTePress, 267–278. https://doi.org/10.5220/0011788900003417

[55] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2931–2937. https://doi.org/10.18653/v1/D17-1317

[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32. https://doi.org/10.1609/aaai.v32i1.11491

[57] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2116–2129. https://doi.org/10.18653/v1/2021.acl-long.165

[58] Vera Schmitt, Balazs Csomor, Joachim Meyer, Luis-Felipe Villa-Arenas, Charlott Jakob, Tim Polzehl, and Sebastian Möller. 2024. Evaluating Human-Centered AI Explanations: Introduction of an XAI Evaluation Framework for Fact-Checking *(MAD '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3643491.3660283

[59] Vera Schmitt, Veronika Solopova, Vinicius Woloszyn, and Jessica de Jesus de Pinho Pinhal. 2021. Implications of the New Regulation Proposed by the European Commission on Automatic Content Moderation. In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*. 47–51. https://doi.org/10.21437/SPSC.2021-10

[60] Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 624–643. https://doi.org/10.18653/v1/2021.naacl-main.52

[61] Khurram Shahzad et al. 2021. Measuring Information Literacy (IL) Skills among University Research Scholars: A Case Study of GC University Lahore. (2021). https://digitalcommons.unl.edu/libphilprac/6418/

[62] Felix M Simon, Sacha Altay, and Hugo Mercier. 2023. Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review* 4, 5 (2023). https://doi.org/10.37016/mr-2020-127

[63] Timo Speith and Markus Langer. 2023. A new perspective on evaluation methods for explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE, 325–331. https://doi.org/10.1109/REW57809.2023.00061

[64] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[65] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018). https://doi.org/10.2139/ssrn.3144139
[66] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478. https://doi.org/10.2307/30036540
[67] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106. https://doi.org/10.1016/j.inffus.2021.05.009
[68] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7534–7550. https://doi.org/10.18653/v1/2020.emnlp-main.609
[69] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067
[70] Jonas Wanner, Lukas-Valentin Herm, Kai Heinrich, and Christian Janiesch. 2022. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets* (2022), 1–24. https://doi.org/10.1007/s12525-022-00593-5
[71] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report* 27 (2017), 1–107. https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html
[72] Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/698d51a19d8a121ce581499d7b701668-Paper-round1.pdf
[73] Vinicius Woloszyn, Eduardo G Cortes, Rafael Amantea, Vera Schmitt, Dante AC Barone, and Sebastian Möller. 2021. Towards a novel benchmark for automatic generation of claimreview markup. In *Proceedings of the 13th ACM Web Science Conference 2021*. 29–35. https://doi.org/10.1145/3447535.3462640
[74] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* 10, 5 (2021), 593.

A NEWS VERIFICATION DASHBOARD

News Verification Dashboard is shown in figure 7.

B AI PREFERENCE SCORE

$$\text{AI Preference} = \begin{cases} 1 & \text{if } R_0 = R_1 = R_{AI} \\ 0 & \text{if } R_0 = R_{AI}, R_1 \neq R_{AI} \\ \min\left(1, \max\left(0, \frac{R_1 - R_0}{R_{AI} - R_0}\right)\right) & \text{otherwise} \end{cases} \tag{1}$$

In this context, R_0 represents the user's initial truthfulness rating before viewing the AI's output, R_1 is the subsequent truthfulness rating after seeing the AI's output, and R_{AI} is the truthfulness rating from the AI system. The AI Preference value falls within the range of $[0, 1]$. When analyzing participant AI Preference, it refers to the average of their AI Preference values across various news items.

C INFORMATION ABOUT PARTICIPANTS

Demographically, of the 433 participants, 63% were female, 36% male, and 1% identified as diverse. Educational backgrounds included 54% of participants with a university degree and 40% with a high school degree. Employment status showed 36% employed, 32% self-employed, and 10% students, with 45% earning over 50k € annually. Age-wise, 61.3% were between 30 and 50 years, 29.8% between 18 and 29 years, and 8.8% over 50 years old. An information summary about participants is presented in table 1.

D RELIABILITY ANALYSIS OF CONSTRUCTS

The reliability of constructs and descriptive statistics of crowdworkers and journalists are shown in table 2.

E GLOBAL COMPARISON OF SUBJECTIVE EVALUATION DIMENSIONS

In the following, the analysis of journalists and crowdworkers for the global evaluation of the AI system with respect to the three subjective evaluation dimensions can be found. In Figure 8a, the global understandability is displayed where the journalists show a significant increase of V1 to V2 (*U-statistics* 12, *p-value* 0.03), and V1 to V3 (*U-statistics* 12, *p-value* 0.03), indicating that the explanations increase the overall perceived usefulness of the AI system[17]. For the crowdworkers significant differences can be detected between V1 and V3 (*U-statistics* 6355, *p-value* < 0.01), whereas V2 is also higher than V1 but not significant. This indicates, that for both, journalists and crowdworkers, the existences of explanations increase the perceived understandability. For the perceived usefulness, a significant increase can be observed when comparing V1 to V2 (*U-statistics* 10, *p-value* 0.02) and V1 and V3 (*U-statistics* 6.5, *p-value* < 0.01) for journalists. Similarly for crowdworkers where there is a significant increase between V1 and V2 (*U-statistics* 7611, *p-value* < 0.01) and V1 and V3 (*U-statistics* 6776, *p-value* < 0.01). This indicates, that the usefulness increases when XAI features are present. Similar to global understandability, trust shows significant differences for V1 to V2 and for V1 and V3 for both, journalists and crowdworkers alike. This indicates, that overall, XAI features increase global understandability, usefulness, and trust. However, no differentiation can be made between V2 and V3, providing no further inferences if the explanations provided in V2 or V3 are of more help for the task at hand.

F SIGNIFICANT DIFFERENCES OF AI SYSTEM ALIGNMENT

Pairwise comparison of news item categories for AI Preference using Dunn's tests (p-values corrected with the Holm-Bonferroni method) is shown in table 3.

G BLIND TRUST

To evaluate if *blind* trust can be observed and human users adhere to AI system recommendations even when the AI system is incorrect, the accuracy ratings of participants who only viewed the news item is compared with those who also considered the AI system's truthfulness rating and additional information from the News Verification Dashboard. The difference in accuracy before and after presenting XAI features is measured to assess the extent of agreement with the AI system by journalists and crowdworkers. In Figure 9, the comparison of accuracy ratings before and after the exposure to XAI features reveals an increase in truthfulness ratings

[17](ns - not significant, * 0.01 < p <= 0.05, ** 0.001 < p <= 0.01, *** 0.0001 < p <= 0.001, **** p <= 0.0001).

Table 1: Information about participants.

Group	Overall	System Versions			Gender			Age (mean)
		basic	salient	free-text	male	female	diverse	
Crowdworker	406	133	133	140	145	258	3	35.14
Journalist	27	10	7	10	12	15	0	37.2

Table 2: Reliability of constructs and descriptive statistics of crowdworkers and journalists.

Construct	Cronbach's α	Crowdworker		Journalist	
		mean	std	mean	std
Understandability global	0.71	5.43	1.16	5.12	1.14
Understandability local	0.86	6.12	0.82	5.58	1.24
Trust global	0.95	5.10	1.13	5.21	1.10
Trust local	0.83	6.17	0.82	5.84	0.81
Usefulness global	0.82	5.45	0.97	5.40	0.90
Usefulness local	0.91	6.10	0.77	5.84	0.81
Information Literacy	0.90	5.73	0.88	5.67	0.63
Exp. of AI before	0.82	4.59	1.20	4.34	0.78
Exp. of AI after	0.75	4.54	1.10	4.48	0.95

Table 3: Pairwise comparison of news item categories for AI Preference using Dunn's tests (p-values corrected with the Holm-Bonferroni method)

User Type	Category A	Category B	p	Cohen's d
Crowd	clearly true	falsely labelled	$< .01$	1.05
	rather fake	falsely labelled	$< .01$	0.65
	clearly fake	falsely labelled	$< .01$	0.66
Journalist	clearly true	falsely labelled	$< .01$	1.26
	rather fake	falsely labelled	$< .01$	1.22
	clearly fake	falsely labelled	$< .01$	0.92

for both journalists and crowdworkers when exposed to the AI system ratings and additional information for all news item types, except falsely labeled items. The decline in truthfulness ratings for both journalists and crowdworkers concerning falsely labeled items suggests that users might have been persuaded by the AI system to adopt inaccurate predictions.

H MEDIA LITERACY

Furthermore, another experiment was performed to assess the impact of literacy on accuracy when individuals use XAI features under three scenarios defined by truthfulness scores (S_t) provided by the AI system: (1) news that is *easy* to classify ($S_t < 25$ or $75 \leq S_t$), (2) news that is *hard* to classify ($25 \geq S_t < 75$), and (3) news that is *falsely labeled*. In the initial scenario, where only news that is *easy* to classify is considered (Figure 10a), both journalists and crowd workers exhibit relatively good accuracy across various literacy categories. However, when the news is *hard* to classify (Figure 10b) or *falsely labeled* (Figure 10c), there is a significant decline in accuracy observed among crowdworkers. In contrast, for journalists, the drop in accuracy is inversely correlated with the literacy score, indicating that high literacy levels are associated with a reduced tendency towards blind trust in the case of journalists.

I SURVEY

I.1 Global System Evaluation

7 Point Likert scale from strongly agree to strongly disagree Questions adapted from Wanner et al. [70]

I.1.1 General Trust in AI-Systems.

(1) The AI-System supports decision-making in fake news detection very well
(2) The AI-System is able to classify the news articles competently
(3) The AI-System can correctly classify the presented news articles
(4) The AI-System can efficiently classify the presented news articles
(5) In general, the AI-System is able to detect fake news
(6) The AI-System decides neutral
(7) The AI-System decides consistently according to the same criteria
(8) The criteria according to which the AI-System evaluates are acceptable
(9) I think I understand why this AI-System provided the decision it did
(10) I think I understand what this AI-System bases its provided decision on
(11) The classification of the AI-System is comprehensible for me
(12) It is easy to follow what the AI-System does
(13) How satisfied are you with the explanatory quality of the News Dashboard?
(14) I know what will happen the next time I use the AI-System because I understand how it behaves

(15) I tend to trust this AI-System, even though I have little or no knowledge of it
(16) My tendency to trust this AI-System is high

I.1.2 Explainability Usefulness.

(1) Using the News Dashboard would increase my effectiveness in detecting fake news
(2) I think the News Dashboard is useful for assessing news articles
(3) Using the News Dashboard will help me detect fake news faster in the future
(4) Overall, I understand how this News Dashboard assists me with decisions I have to make
(5) Overall, I think the explanations given by the AI-System in the News Dashboard for the news article are useful
(6) It takes too long to learn how to use the News Dashboard to make it worth the effort

I.1.3 Understandability.

(1) Overall, the presented explanations in the News Dashboard are comprehensible and help me with assessing the news articles
(2) The presented explainability features in the News Dashboard seem too complicated
(3) The metadata (source of the article and publishing date) are presented as comprehensible and useful for the task

I.2 Local System Evaluation

7 Point Likert scale from strongly agree to strongly disagree

Trust

(1) The AI-system classified the news items correctly - adapted from Kulms and Kopp [32]
(2) I understand what the AI-system does - adapted from Madsen and Gregor [40]

Explainability Usefulness

(1) The explainability features presented are useful to assess the truthfulness of the news item.
(2) The indications given by the AI-system are useful to assess the credibility of the news item.

Understandability

(1) The presented explanations are comprehensible and help me with assessing the news articles

I.3 Further Constructs

I.3.1 Expectation of AI before experimental part. 7 Point Likert scale from strongly agree to strongly disagree Questions adapted from Gillespie et al. [22], Mcknight et al. [43]

(1) My typical approach is to trust new technologies until they prove to me that I shouldn't trust them [43]
(2) I generally give technology the benefit of the doubt when I first use it [43]
(3) How willing are you to rely on information provided by an AI system in the context of fake news detection?
(4) In general, are you sceptical about AI
(5) In general, do you trust AI?
(6) My typical approach is to trust new technologies until they prove to me that I shouldn't trust them

I.3.2 Expectation of AI after the experimental part. 7 Point Likert scale from strongly agree to strongly disagree

(1) My typical approach is to trust new technologies until they prove to me that I shouldn't trust them [43]
(2) I generally give technology the benefit of the doubt when I first use it [43]
(3) How willing are you to rely on information provided by the AI-System in the context of fake news detection? [22]
(4) Overall, are you skeptical about the AI-System? [22]
(5) My typical approach is to trust new technologies until they prove to me that I shouldn't trust them [22]

I.3.3 Information Literacy. 7 Point Likert scale from always true to never true

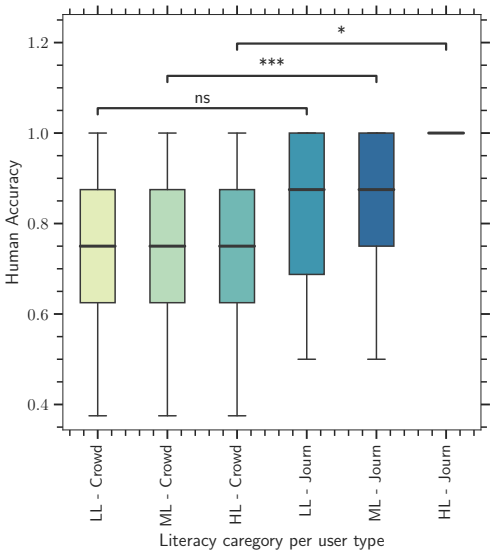
Questions adapted from from Shahzad et al. [61]

I think...

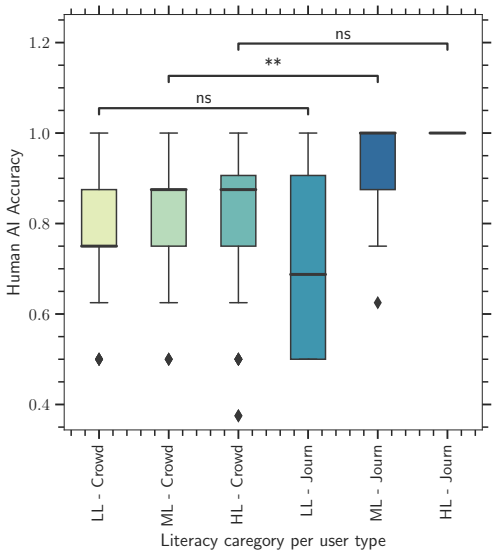
(1) ...I have the skill to access information resources for finding relevant literature, about topics I want to verify (e.g. looking up information about climate change and its effect on our everyday life)
(2) ...I can check the reliability of the searched information
(3) ...I can differentiate between right and fake information when I am reading news, also online in my social media channels
(4) ...I can efficiently use subscribed and open-access information resources when I am looking for information
(5) ...I have skills in Information Technology (IT)
(6) ...I can identify the best authentic sources of information
(7) ...I can identify misinformation in news
(8) ...I know about plagiarism and publication ethics
(9) ...I have research-level skills

I.3.4 Demographics.

(1) To which age category do you belong?
(2) In which country do you currently live?
(3) What is your highest school-leaving qualification?
(4) Which of the following categories best describes your employment status?
(5) What is your gender?
(6) What is your annual household income (gross)?
(7) What is your nationality?

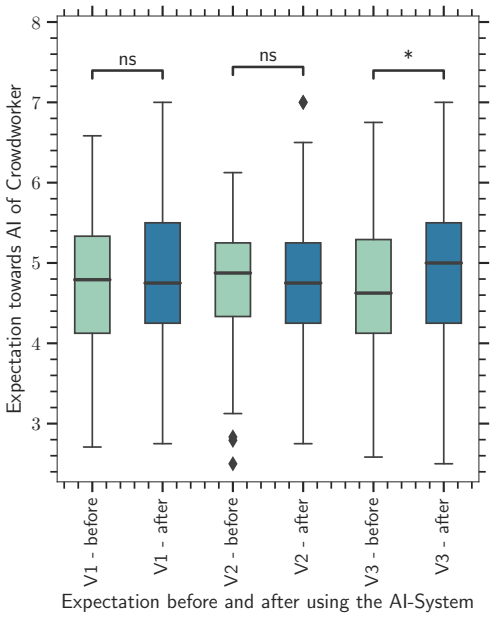


(a) Human accuracy.

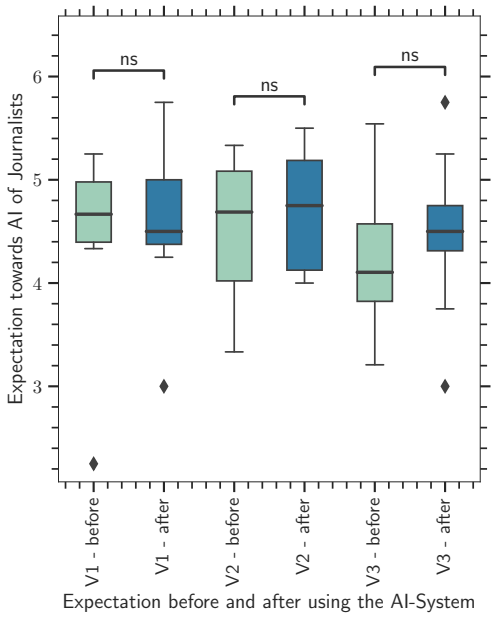


(b) Human-AI accuracy.

Figure 5: Accuracy comparison for different levels of literacy: low (LL), medium (ML), and high literacy (HL).



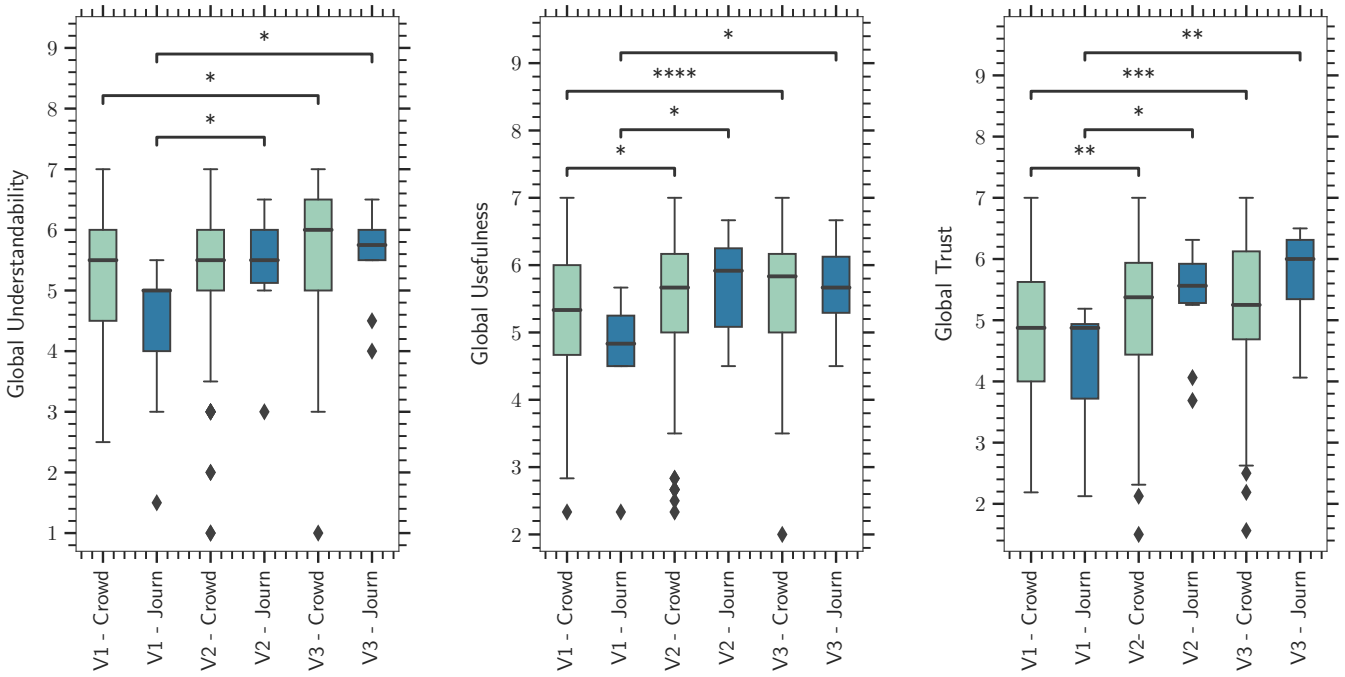
(a) Crowdworkers.



(b) Journalist.

Figure 6: Expectation comparison before and after using the AI-System for crowdworkers and journalists.

Figure 7: News Verification Dashboard, where the second version containing the text highlights is visualized combined with the free-text explanations.



(a) Global understandability. (b) Global usefulness. (c) Global trust.

Figure 8: Comparison of global explanation understandability, usefulness, and trust between journalists and crowdworker for the different AI-system versions.

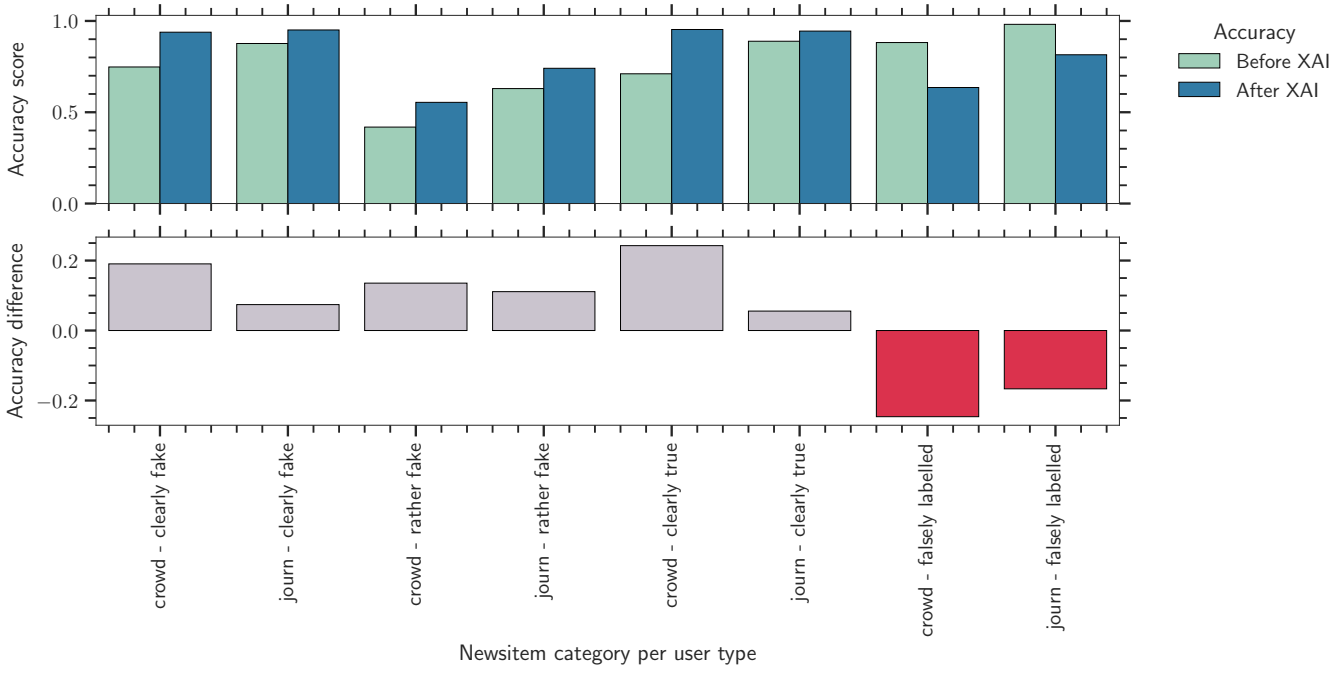
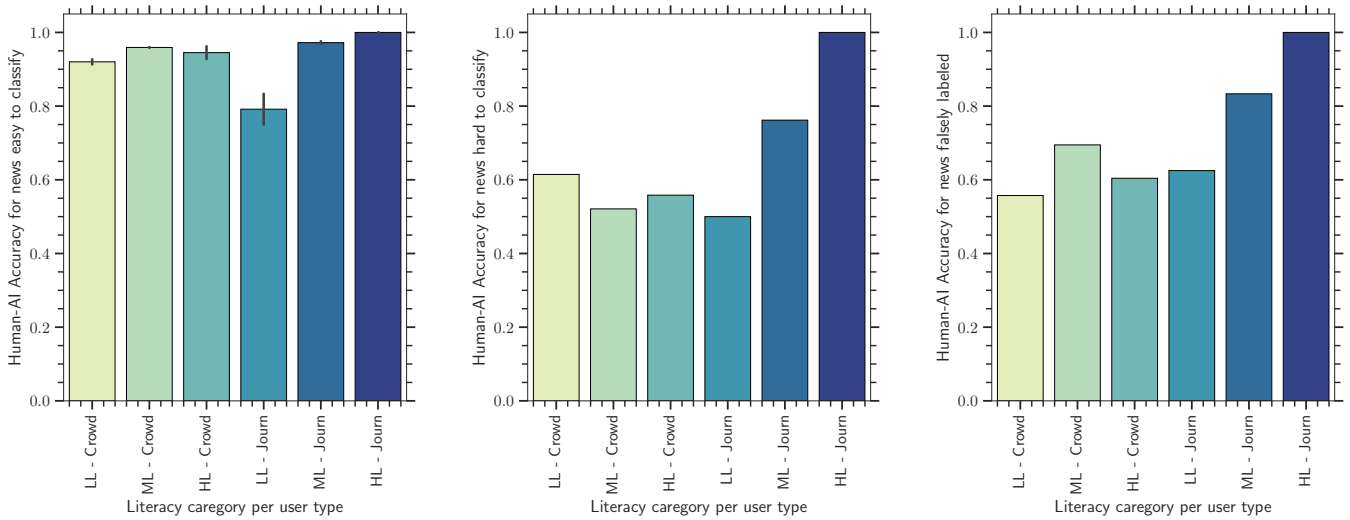


Figure 9: Comparison by user type of performance over all item categories, before and after the explanations have been presented.



(a) News easy to classify. **(b) News hard to classify.** **(c) News falsely labeled.**

Figure 10: Accuracy comparison per classification difficulty (easy and hard to classify, and falsely labeled) by the system for different levels of literacy: low literacy (LL), medium literacy (ML), and high literacy (HL).