Beyond Transparency: Evaluating Explainability in AI-Supported Fact-Checking

Vera Schmitt
TU Berlin, DFKI, CERTAIN
Berlin, Germany
vera.schmitt@tu-berlin.de

Isabel Bezzaoui
Digital Democracy and Participation,
Karlsruhe Institute of Technology
Karlsruhe, Germany
Bezzaoui@fzi.de

Charlott Jakob
TU Berlin, DFKI
Berlin, Germany
c.jakob@tu-berlin.de

Premtim Sahitaj
TU Berlin, DFKI
Berlin, Germany
sahitaj@tu-berlin.de

Qianli Wang
TU Berlin, DFKI
Berlin, Germany
qianli.wang@tu-berlin.de

Arthur Hilbert
TU Berlin, DFKI
Berlin, Germany
arthur.hilbert@tu-berlin.de

Max Upravitelev
TU Berlin, DFKI
Berlin, Germany
max.upravitelev@tu-berlin.de

Jonas Fegert
Digital Democracy and Participation,
Karlsruhe Institute of Technology
Karlsruhe, Germany
fegert@fzi.de

Sebastian Möller
TU Berlin, DFKI
Berlin, Germany
sebastian.moeller@tu-berlin.de

Veronika Solopova
TU Berlin, DFKI
Berlin, Germany
veronika.solopova@tu-berlin.de

Abstract

The rise of Generative AI has made the creation and spread of disinformation easier than ever. In response, the EU's Digital Services Act now requires social media platforms to implement effective countermeasures. However, the sheer volume of online content renders manual verification increasingly impractical. Recent research shows that combining AI with human expertise can improve fact-checking performance, but human oversight remains crucial, especially in domains involving fundamental rights like free speech. When ground truth is uncertain, AI systems must be both transparent and explainable. While various explainability methods have been applied to disinformation detection, they often lack human-centered evaluation regarding their task-specific usefulness and interpretability. In this study, we evaluate different explainability features in AI systems for fact-checking, focusing on their impact on performance, perceived usefulness, and understandability. Based on a user study (n=406) including crowdworkers and journalists, we find that explanations enhance perceived usefulness and clarity but do not consistently improve human-AI performance, and can even lead to overconfidence. Moreover, whereas XAI features generally help to increase performance, they enabled more individual interpretation among experts and lay-users, resulting in a broader variation of outcomes under. This underscores the need for complementary interventions and training to mitigate overreliance and support effective human-AI collaboration in fact-checking.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in interaction design**; • **Computing methodologies** → *Discourse, dialogue and pragmatics.*

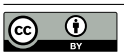
Keywords

Explainable AI, Meaningful Transparency, Fact-Checking, Disinformation Detection, Human-Centered AI, NLP/LLMs, Empirical Evaluation, AI Act, DSA

ACM Reference Format:
Vera Schmitt, Isabel Bezzaoui, Charlott Jakob, Premtim Sahitaj, Qianli Wang, Arthur Hilbert, Max Upravitelev, Jonas Fegert, Sebastian Möller, and Veronika Solopova. 2025. Beyond Transparency: Evaluating Explainability in AI-Supported Fact-Checking. In *4th ACM International Workshop on Multimedia AI against Disinformation (MAD'25), June 30–July 03, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3733567.3735566

1 Introduction

A provisional agreement has been reached by the European Council and the Parliament on the proposal for harmonized rules on artificial intelligence (AI), the so-called AI Act. In Article 13 of the EU AI Act *Transparency and Provision of Information to Users*[1] obligations



This work is licensed under a Creative Commons Attribution 4.0 International License.

MAD'25, Chicago, IL, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1891-5/25/06
https://doi.org/10.1145/3733567.3735566

[1] Laying down Harmonised Rules on Artificial Intelligence (AI Act), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 09.12.2023.

are defined for sufficient transparency to enable providers and users to reasonably understand the AI System's functioning. Especially, in domains where a vast amount of information exists, which is infeasible to screen manually, AI Systems are used for intelligent decision support, to automatically screen and sort information based on certain characteristics. Therefore, advances in Natural Language Processing (NLP) and Machine Learning (ML) are increasingly used for the content verification task to detect disinformation. However, the automated detection of disinformation articles is inherently difficult for various reasons. News items may have no clear, discrete truth value. Rather, the truthfulness of items is on a continuum between clearly true and clearly false [14, 47]. Furthermore, the classification of news items depends on the viewer's prior beliefs and knowledge about relevant domains, and items can contain sarcasm and irony which reverse their meaning [46]. Consequently, the detection of disinformation still requires the involvement of human judgment. Recent research shows that hybrid human-machine systems are able to accomplish tasks that neither can do on their own [14]. In such situations, the human decision-maker usually oversees the AI system's performance, which helps to identify news items that may be problematic. Recent research on explainable AI (XAI) demonstrates that it can be used to make the *black box* of AI algorithms more transparent and enhance human comprehension of AI classifications or generations. This may promote reliability and human trust in such systems and can facilitate the adoption of AI systems for different contexts. However, Miller, 2019 [31] claims that most explainability research relies on researchers' intuitions regarding what qualifies as a satisfactory explanation, rather than a user-centered approach that considers the expectations, worries, and experiences of the user. This is highly relevant, as the AI Act further requires the integration of *meaningful explanations* into AI Systems for users. In this regard, there is no specific clarification regarding the scope of meaningful explanations, leaving room for interpretation and application across various domains and contexts involving AI systems. Thus, in this work, we evaluate human-centric *meaningful* explanations to enhance the Quality of Experience (QoE) for intelligent decision support in the fact-checking context. This study aims to address the following research questions:

RQ1: Can explanations enhance the perceived QoE in terms of understandability, and usefulness in a fact-checking system?

RQ2: Are the provided information sufficient to determine the truthfulness of the given news items?

Overall, this research evaluates the QoE of human-centric explanations based on their perceived understandability and usefulness for the fact-checking task. This research provides empirical evidence on whether explanations help human users when using Intelligent Decision-Support Systems (IDSS) to increase their ability to detect disinformation. As such, AI tools with XAI explanations would prove to be useful intervention methods for fact-checkers.

2 Related Work

Mis-and disinformation is influenced by a country's constitutional and legal framework, culture, political context, and public awareness and therefore is difficult to define [28]. The European Commission High-Level Expert Group (HLEG) proposed a distinction between mis- and disinformation, whereas misinformation refers to unintentionally false and inaccurate information shared by individuals, and disinformation to verifiable false or misleading information created, presented, and spread for economic gain or to intentionally deceive the public, potentially causing harm[2]. The information overload during the COVID-19 pandemic and the Russia-Ukraine war shows the profound impact that mis- and disinformation can have on shaping public opinion [36]. Since manual fact-checking on a large scale is not feasible, crowdsourcing and automatic detection approaches gained much attention in recent years [17, 35]. Current research on mis- and disinformation detection employed a variety of approaches [4]: Propagation-based research is concerned with the dissemination patterns of mis- and disinformation [24, 53]. Source analysis approaches focus entirely on the source of mis- and disinformation items and allow for early detection [5]. Content-based approaches extract lexical or syntactic linguistic features, assuming that mis- and disinformation items tend to use deceptive language and syntactic styles [16, 40, 44]. Within all these approaches machine learning approaches and Large Language Models (LLMs) are used to facilitate the detection of false, deceptive, and hateful content online [43]. Within the European Union a strong emphasis on ensuring a human-centric and ethical approach to AI is adopted, by proposing a legal framework to regulate AI applications. The so-called AI Act[3] adopts a risk-based approach by categorizing AI applications based on different risk categories. For both categories, limited and high risk, transparency obligations are defined, where human oversight, transparent AI systems, and explanations are essential [1]. Recent research shows that seeing explanations for the behavior of automated mis- and disinformation detection systems can increase the QoE in terms of the perceived usefulness and understandability of AI-generated predictions and the acceptance and credibility of a particular detection model [37]. For explainable fact-checking, one can employ different means to produce justifications for a model decision, e.g., attention weights, feature importance methods, logic-based systems, or rationalizing models that generate natural language explanations [14, 41]. Some evaluations of textual explanations show that automatically generated explanations can increase human understanding, trust, and confidence in the AI system for certain tasks [21]. The human-centered nature of explainability has sparked rising interest among researchers across various domains in exploring XAI [7]. The effectiveness and usefulness of explanations depend on the perception and understanding of the person receiving them. This implies that the design decisions for explaining model results to users need to significantly impact technical decisions, including design and evaluation, to enhance the QoE [22]. With the increasing prevalence and importance of XAI-based solutions, it is crucial to conduct thorough evaluations of their explainability components for the intended use cases, including verifying that the explanations accurately represent the inner workings of the associated machine learning model (faithfulness), and confirming that they are truly effective and meaningful for end-users (plausibility and usefulness). The current focus of XAI research revolves around developing novel methods to enhance explainability without compromising high predictive performance

[2]The 2022 Code of Practice on Disinformation, https://shorturl.at/fEFV2, 09.01.2023

[3]https://artificialintelligenceact.eu/ai-act-explorer/

[49], often showing that explanations can enhance users' understanding and trust in machine learning systems [26]. Yet, some authors contend that the mere presence of explanations can trigger these effects, irrespective of their substance [11], which could potentially create a false sense of understanding. Furthermore, there exists an inherent human predisposition towards simpler explanations, which may result in the adoption of systems with more compelling explanatory outputs instead of more transparent ones [34]. Overall, the evaluation of human-centric explanations lacks empirically tested schemes [25]. Thus, it is important to comprehensively assess the explanatory methods and artefacts generated by XAI systems, both prior to and during their deployment in a production setting [26]. Hence, we focus on the specific context of mis- and disinformation detection to evaluate which types of explanations are relevant to the task to increase the user's understanding of the AI system [12, 23].

3 Methodology

Human-centered explainability evaluations take a *human-in-the-loop* approach, where users actively engage with the system to examine and assess specific attributes. The main goal is to gain deeper insight into how end-users are likely to perceive the system by highlighting the strengths and limitations of the XAI approach and its implementation, along with identifying potential avenues for improvement [25]. Prioritizing evaluation is critical during the design, implementation, and deployment of XAI systems, and it must cover both the technical and human interaction aspects of any machine learning system. This paper focuses on the human-centered aspects of XAI evaluation, aiming to identify which types of explanations enhance human engagement in collaborative human-AI information verification tasks. Thus, we propose an XAI evaluation schema, which covers relevant human-centric evaluation aspects [25, 26] in the context of collaborative human-AI disinformation detection.

Lopes et al. [26] have proposed a taxonomy for XAI evaluation covering subjective and objective dimensions of *post-hoc* evaluation of explanations without specifying any concrete context. The proposed dimensions of human-centered XAI evaluations have not been empirically tested and remain a theoretical taxonomy without specifying the concrete constructs and items to be used for applying the evaluation dimension to a specific context. The dimensions of understandability and explanation usefulness are presented solely on a conceptual level without undergoing empirical investigation or validation. A set of items is comprised of previous literature to measure the two QoE-related dimensions of perceived explanation usefulness and understandability. The items in the questionnaire build on previous multidisciplinary research in the domain of HCI and XAI evaluation to develop item sets for the evaluation of XAI features for the content verification task [26, 33, 49].

Understandability is a relevant subjective dimension when evaluating explanations in any context. It can be defined as the capability to "characterize the relation between the input and output of a system with respect to its parameters" [13]. Hereby, understandability can refer either to the given task, the understanding of the underlying function of the AI system, or the explanations given. In this schema, it refers only to the understanding of the explanations, which support the understanding of the AI system prediction. The understandability is measured by the user's perceived understandability [38] of the explanations locally for each interaction with the AI system, and globally when evaluating the overall AI system.

Explanation usefulness is relevant to assess if added explanations are relevant for improving a given task or not [10]. For the content verification domain, it is important to create different versions of the AI system differing in the information and explanations provided. The usefulness of the explanation is measured objectively by comparing the user's performance of the AI systems with differing explanations given and verifying if the human performance in detecting mis- and disinformation increases when explanations are added. Furthermore, the subjectively perceived explanations usefulness [10] is measured locally after each interaction with the AI system and also globally when evaluating the overall AI system.

By employing these subjective and objective evaluation metrics, it becomes possible to assess the impact of explanations on trust and user preferences, offering valuable insights into the alignment of human evaluators with the AI system for the mis- and disinformation detection task.

3.1 Specification of Explanations

It is not only important to provide explanations for a given AI system output but also to ensure that they are *human-meaningful* explanations, since *weak* explanations that are not comprehensible for the user might lower their trust, perceived understandability, and overall performance [26]. Regarding different types of explanations, we distinguish between highlights, free-text, and structured explanations [52].

Highlights: Highlights mark text segments most influential to the model's prediction [52]. We define two types: (1) *Truthfulness justifications*, identifying check-worthy claims based on prior fact-checking work [30], and (2) *Emotional content*, capturing sentiment, emotion, and manipulative style [29]. In our interface, these are shown in yellow and cyan, respectively.

Free-text Explanations: We use expert-annotated, natural language rationales—building on fact-checking datasets [18]—to justify why a news article is classified as true or false, incorporating commonsense reasoning and paraphrased external sources. Commonsense reasoining and expert annotations are collected from Snopes[4]. Following a Wizard-of-Oz setup, these explanations simulate ideal outputs of rationale-generating systems [39], using fully human-derived content to assess which explanation types foster trust and utility in our news dashboard.

Readability: We use the Flesch-Kincaid grade-level score [18] to estimate text readability, expecting that complex language appears more often in trustworthy sources. While not critical for detecting misleading tweets [3], readability remains a useful extrinsic feature for explainable fact-checking [42, 51]. In our *News Verification Dashboard* [45], scores are categorized as *Easy* (< 10), *Medium* ($10–12.5$), and *Hard* (> 12.5).

The explanations are integrated into a *News Verification Dashboard*, which displays the news article alongside the respective metadata (domain, title, source, and publishing date), and the truthfulness rating of an AI system fact-checking the respective news

[4]Fact-Checker: Snopes, https://www.snopes.com/, 09.01.2024

item. The explanations are displayed next to the AI system rating and the news item's metadata. A full description of the technical framework and *News Verification Dashboard* can be found here [45].

3.2 AI System Versions

We evaluate the explanations separately to allow for a comparison of the four evaluation dimensions respectively. We propose three different versions of an AI system with each of them showcasing different explainability features.

Version 1 (V1) - Baseline: Does not contain any explanations and only displays the news item, its metadata, and the automatic truthfulness rating from the AI system,

Version 2 (V2) - Salient Features: Encloses the information of version 1, the readability feature, and the two types of highlights,

Version 3 (V3) - Free-Text Explanation: Contains all information from version 1 and free-text explanations.

To assess the effectiveness of suggested explanation types for content verification, user studies must include a control group (V1) that receives no additional explanations regarding the AI system's output on truthfulness. The explanations can then be tested individually in experimental groups (V2, and V3) to determine their usefulness for collaborative content verification. The explanations should be assessed both globally for the overall system and locally by evaluating each interaction with the AI system individually. The local and global evaluation of the evaluation dimensions allows for the analysis of the evaluation dimensions for each news article separately. Hereby, the effect of the topic on the evaluation dimensions, the AI performance, and the explanations can be analyzed more in-depth [33].

3.3 Ethical Statement

The experiment emphasized the privacy of participants by not collecting any personal information. A unique ID was used for each participant to maintain anonymity. Detailed information regarding data collection and processing was provided to participants, who then gave their explicit consent to participate. Participants were also made aware that they had the right to withdraw from the study at any time, which would result in the removal of their data. The ethics committee of faculty IV of the Technische Universität Berlin approved the study and with no additional ethical concerns or requirements.

4 Results

In total, 600 individuals participated in the experiment. 167 participants have been excluded from the screening task, as they did not fill out the trapping questions correctly. The responses from 406 participants showed correctly answered test questions and are therefore included in the analysis and assessment of explanations. The experiment was split into three groups, each group receiving only one of the following versions: baseline version 1 without any explanations, which involved 133 participants; version 2 with saliency explanations, which included 140 participants; and version 3 with free-text explanations, which had 133 participants. The reliability of the QoE evaluation in terms of understandability and usefulness showed good internal consistency for both the global and local evaluation. The Cronbach's α for understandability (local=0.86, global = 0.91) and usefulness (local = 0.91, global=0.82) is above 0.7, indicating a good reliability of the used items loading on the same construct. Among the 433 participants, 63% identified as female, 36% identified as male, and 1% identified as belonging to the diverse gender category. Most of the participants have a university degree 54%, or high-school degree 40%, are employed 36%, self-employed 32%, or student 10%, and 45% earn more than 50k € per year. Most of the participants were between 30 and 50 years old 61, 3%, and 29, 8% were between 18-29 years old. 8, 8% reported to be older than 50 years.

4.1 Performance

Figure 2 compares the accuracy of the three system versions (V1–V3) evaluated by crowd workers and professional journalists. Although accuracy levels are generally high across groups, Version 3 (V3), which includes free-text rationales, shows that journalists and crowdworkers achieve higher median accuracy in this condition, but also exhibited notably greater variance, as indicated by a wider interquartile range and extended whiskers. This suggests that the introduction of free-text explanations enabled more individual interpretation among experts and lay-users, resulting in a broader spread of outcomes. The finding highlights a trade-off between interpretability and consistency in expert and lay-people assessments, where increased transparency may foster deeper engagement but also amplify variability in decision quality.

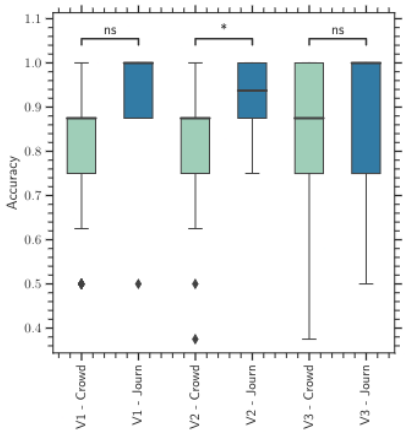
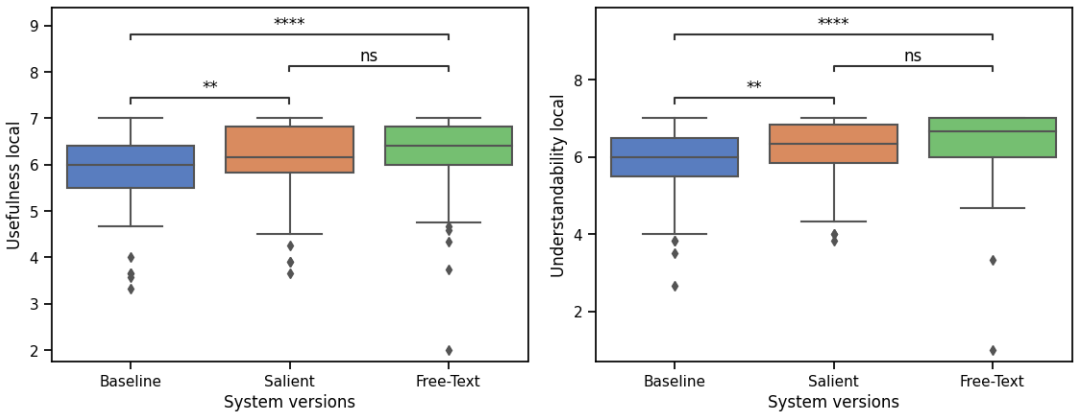


Figure 2: Performance.

4.2 Perceived Understandability

The subjective measure of understandability of explanations has been examined locally after each news item and also globally, in a more extensive survey at the end of the experiment. After conducting a Mann-Whitney-U test to examine the statistically significant differences between the three versions concerning understandability (see Figure 1b), increased understandability can be observed for the salient feature version V2 in comparison to the baseline version V2 (*U-statistics* 7470, *p-value* 0.01)[5]. The perceived understandability was the highest when free-text explanations are shown to the participants (*U-statistics* 5952, *p-value* < 0.01). When comparing the local and the global evaluation of understandability for each AI system version significant differences are present, whereas the local evaluation results in significantly higher usefulness ratings compared to the global evaluation. No significant difference can be found between the salient and free-text explanations, providing no

[5]Bonferroni correction is applied to *p-values* avoiding α accumulation error



(a) Perceived usefulness. **(b) Perceived understandability.**

Figure 1: Comparison of AI system versions and their influence on the perceived usefulness and understandability (**: $p \leq .0001$, ns: $p > .05$).**

indication of which explanation results in increased understandability.

4.3 Explanation Usefulness

For the subjective evaluation dimension explanation usefulness, a Mann-Whitney-U test showed significant differences between the AI system V1 one and V2 (*U-statistics* 7525, *p-value* < 0.01), and V1 and V3 (*U-statistics* 5921, *p-value* < 0.01). This indicates that the salient features and free-text explanations positively contribute to the explanation's usefulness compared to the baseline version. However, we find no significant differences between the free-text explanation and the salient features version, whereas the free-text explanation version achieves a slightly higher usefulness rating (mean=6.3) than the salient features version (mean=6.1). When comparing the local with the global evaluations, 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. The results show that the tested XAI features in V2 and V3 significantly increase the QoE in terms of significantly higher perceived understandability and usefulness. Even though the understandability and usefulness ratings for V3 in comparison to V2 are higher, no significant difference can be observed. This finding shows that XAI features enhance the QoE, but no further indication can be made which XAI features contribute more to a higher QoE based on the results of the crowdsourcing experiment.

4.4 Qualitative Evaluation

Asking 1) "What criteria do you usually use to judge whether a news/ article is reliable?" and (2) "What other information would you like to obtain to better assess the truthfulness of an article?"; "What functionality would be a good addition?". Figures 3 and 4 show the most frequent terms for the respective question. In addition to the evaluation of selected salient features and free-text explanations, we incorporated two open-ended questions aimed at understanding participants' desires for further XAI features. Given the similarity in the responses to these questions, we combined the analysis of the results. For the analysis of the open-ended questions, the answers are grouped and ranked according to their occurrence. In Figure 3 the responses are shown for the question concerning the first question. The main criterion mentioned within approximately 69% of the answers was the source's reputation and credibility. Key terms associated with this criterion include *source*, *credibility*, *reputation*, *publish*, *publisher*, *reliability*, *reputable*, and *credible*.About 16% of the responses mentioned fact-checking, using terms like *fact*, *check*, *research*, or *google*. Additional criteria included the use of language (grammar or emotional tone), *common sense*, the *author* of the articles, *bias*, *objectivity* of articles, and the use of *citations*.

Figure 3: 50 most frequent terms in trustworthiness assessments.

The findings from the qualitative analysis of the responses are in line with the recommendations given by the Federal Government of Germany how to identify disinformation[6]. The responses of the second question concerning the additional features and information required (see Figure 4), the most requested additional information was about the article's *source*, including the *name* of the *source*, a *link* to the original article, the *name* of the *publisher*, the *type* of *source* (e.g. blog, online, print), and information about the source's *trustworthiness*. Furthermore, the publication date was also frequently requested. Participants wanted to see citations and references used by the article for claims, or a second article or alternative source/AI rating. Moreover, information about biases or political views of the source and author are requested and highlighting grammatical and spelling mistakes were noted. Users expressed interest in more details about the AI system, including who programmed it, the sources it uses,

[6] *How to recognize mis- and disinformation*, https://shorturl.at/eqR46, 07.02.2025.

and explanations for its ratings. Additionally, there was a demand for an AI capable of detecting manipulated images or videos and identifying AI-generated content. Participants desired the ability to provide feedback on the AI system's ratings, mentioning instances where they believed the system made errors. Also, a feature allowing users to listen to audio versions of articles was suggested, aligning with digital media trends toward providing AI-generated audio content[7].

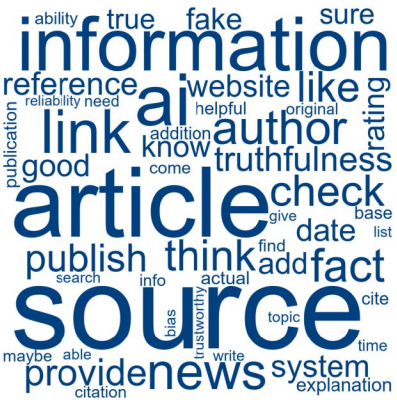


Figure 4: 50 most frequent terms in user feature and information requests.

The study hypothesized that participants using the baseline version of an AI system would more frequently request explanations on how the AI determined its truthfulness score compared to those using versions with integrated explanations. To investigate this, responses were specifically marked as "requests explanation" whenever participants wanted more detailed information on the AI's scoring process or the criteria it employed. Following the categorization, a chi-square test was conducted to assess the relationship between using the three different XAI versions and the frequency of explanation requests. The findings revealed that 20.3% of participants in the baseline version sought explanations, significantly more than those in Version 2 (10.0%) and Version 3 (8.3%), with χ^2=10.13, $p\text{-}value< 0.01$. This outcome suggests a clear interest among users for explanations within AI-assisted decision-making systems.

Overall, the results show, that XAI features increase the usefulness and understandability of AI-based fact-checking systems significantly and contributing to a higher performance in detecting false information. However, there are still several features and requirements which need to be incorporated, to fulfill all users needs, as the qualitative evaluation has shown. In the next section, the requirements are described in more detail by also analyzing the legal obligations arising from the European AI Act and Digital Services Act (DSA).

5 Recommendations for Designing Explanations for Enhanced Fact-Checking

With the increasing prevalence of disinformation and the growing reliance on AI systems for fact-checking, it is imperative to design AI systems that not only detect disinformation effectively but also provide users with meaningful transparency utilizing explanations of their processes. Based on the results of our study, this section outlines empirically grounded recommendations for designing XAI features to enhance disinformation detection systems. These recommendations focus on improving user trust, engagement, and understanding while addressing concerns of over-reliance and usability. Additionally, it integrates the legal obligations defined by the AI Act and DSA, which highlight the need for human oversight, transparency, and reliability in AI-based fact-checking systems.

5.1 Prioritizing Source Credibility and Transparency

Our study revealed that the primary criterion participants used to evaluate the reliability of news articles was the reputation and credibility of the source, underscoring the critical role of source transparency in determining trustworthiness. This finding suggests that explanations should provide clear and detailed information about the origin of the article, including the publisher's name, the reputation of the source, and a link to the original article. Additionally, systems should clearly indicate the type of source and disclose potential political biases or affiliations associated with the publisher or author. Such transparency aligns with user expectations and enhances the usability of AI-based fact-checking systems by allowing users to critically evaluate the source's reliability [27]. Moreover, it is important to conduct evaluation of transparency measures included into AI-based fact-checking systems. This evaluation should not only consider performance metrics but also user-specific perspectives such as usefulness, trust, and understandability [27]. Previous research has shown, that xAI features increase trust in AI system outputs even when they are wrong. This indicates that transparency increases *blind trust* in AI systems [47], especially for lay users in comparison to expert users. Thus, reliability and faithfulness of model outputs are required to provide meaningful transparency. This aligns with the AI Act's requirement for transparency and human oversight, ensuring that users can critically evaluate the trustworthiness of a source. Systems should also offer explanations of the decision-making process, including the datasets and sources used to inform the AI's conclusions, further fulfilling legal transparency obligations [15].

5.2 Integrating Fact-Checking Tools

Participants noted that fact-checking tools are essential for evaluating claims. This highlights the importance of providing users with tools that allow for the independent verification of content on various platforms online, like social media and news outlets. To further strengthen the transparency of fact-checking tools, further evidence can be retrieved from existing knowledge bases consisting of factual information. This additional evidence can be shown to the users as short summaries by also referencing existing fact-checks and trustworthy sources to linking and cross-verifying existing knowledge effectively [43, 44]. These features may provide an additional layer of validation, improving the overall usefulness and reliability of the explanations provided by the system. In line with the DSA's requirements for content moderation, these systems should offer transparency regarding how decisions are made and provide references that substantiate claims. Fact-checking tools must ensure reliability and robustness, achieved through high-quality, diverse data and advanced model architectures that generalize well across various contexts [9].

[7] Journalism, media, and technology trends and predictions 2023, 07.01.2025.

5.3 Addressing Bias and Objectivity

Participants expressed a desire to discern not only the factual accuracy of news articles but also whether the content is presented in a balanced and neutral manner. The third requirement journalists stated to use LLMs for fact-checking, and is also demanded by the AI Act is the handling of biases. Different biases need to be considered: (i) Data biases which can contain historical biases when trained on large datasets reflecting and reproducing historical biases, and also representation bias where subgroups might be under- or overrepresented, which can skew outputs [32]. Thus, the biases in training data need to be considered and explored how the model outputs are affected by skewed data distributions towards certain classes and labels. (ii) Algorithmic or model bias concerns the architecture and training process of LLMs [6]. For instance, models might favor more frequently occurring patterns in the data, ignoring minority viewpoints. Additionally, LLMs might perpetuate stereotypes present in the training data, reinforcing harmful biases in their outputs [54]. (iii) Selection biases, where the source used for training data might have inherent biases. LLMs might prioritize information from sources with higher visibility or perceived credibility, which might not always be neutral or balanced. Additionally, the confirmation bias, where the model might favor information that confirms existing beliefs or widely held views, potentially ignoring contradictory evidence needs to be considered when using LLMs for fact-checking. (iv) Contextual biases pose further challenges where LLMs might struggle with understanding context and nuance, leading to misinterpretation of ambiguous statements or satire as factual information [20]. (v) Similarly, cognitive biases and framing effects where LLMs might give undue weight to the initial information they process, leading to biased fact-checking results [19]. This can be addressed by using only diverse and representative training data, deploying bias mitigation approaches before and during the training process, and human oversight and collaboration by using human knowledge to ensure the correctness of the results. For achieving correctness of the results, AI-based fact-checking systems should provide explanations that explicitly address potential biases in the articles or their sources. Publishing political affiliations of the editor or author, identifying emotionally charged or biased language within the article, and offering a neutrality or objectivity rating based on content analysis could help to overcome this issue [16]. Providing users with such insights empowers them to critically assess the balance of the information presented, potentially enhancing the overall effectiveness of the AI system in detecting disinformation. The AI Act also emphasizes the need to address bias, ensuring that AI systems are fair and unbiased, which might be achieved through careful selection of training data, algorithmic transparency, and human oversight[19, 20].

5.4 Enhancing System Transparency and Robustness

A key finding from our study is the strong user interest in understanding the internal workings of the AI system. This was especially evident in the baseline system (V1), where a substantial part of the participants requested more detailed explanations about how the AI determined its truthfulness scores, significantly more so than in the systems with integrated explanations. To address the need for system transparency, AI tools should offer detailed explanations of their decision-making processes, including information about the datasets and sources used to inform the system. Moreover, the AI system should clearly articulate the reasoning behind the truthfulness score, allowing users to understand how and why specific judgments were made [47, 48, 50]. To ensure accurate outputs, disinformation detection systems must prioritize robustness and reliability, as required by both the AI Act and DSA. One of the key requirements is the robustness and reliability of results. Robustness in this context refers to the model's ability to maintain performance and produce accurate, reliable outputs under varying conditions and potential adversities. This can be achieved by training or fine-tuning the models on highly qualitative and diverse data, containing a wide range of topics, and contexts to improve generalization performance and minimize potential biases. Moreover, robust training processes, including data augmentation, regularization techniques, and fine-tuning strategies, help enhance the model's ability to generalize well across different scenarios. But also the design and complexity of the model architecture play a crucial role in its robustness. Advanced architectures like transformers with attention mechanisms have shown improved robustness in handling diverse and complex inputs. Moreover, by adding contrasting layers the factuality of generalization outputs (e.g. for summarization tasks) can be significantly improved, and the risk of *hallucinations* reduced [9]. For robust and reliable outputs, contextual understanding, resilience to noise, adaptability, error handling, and scalability to handle large volumes of data need to be considered and monitored when using LLMs for AI-based fact-checking systems. This also includes fine-tuning models on diverse datasets and using advanced architectures like transformers to handle complex inputs. Reliable outputs must be able to withstand noise and variations in data, ensuring that the AI system performs well across different scenarios. Additionally, error-handling mechanisms should be integrated to ensure the reliability of AI-based fact-checking systems [2, 6].

5.5 Implementing User Feedback Mechanisms

Participants expressed a desire to provide feedback on the AI system's decisions, particularly in cases where they believed the AI made an error. This feedback may also help improve the system's accuracy and responsiveness over time. The requirement of user feedback is especially apparent in detecting generated or manipulated content, such as text, images, or videos, while this collaborative framing also conforms with the AI Act's (Chapter III. Article 14) requirement for human oversight and engagement. Moreover, this contributes to the required "risk management" efforts, considering automated fact-checking tools as "high-risk applications" with their impacts on fundamental rights, such as the right to accurate information and freedom of expression and possible legal ramifications for the fact-checkers in accordance with the local media regulations. Additionally, XAI offers an educational benefit by enabling fact-checkers to identify emerging patterns associated with manipulative content—patterns that may not yet be formalized in fact-checking guidelines. This is especially evident in e.g. detecting generation and manipulation artefacts that the human eye has difficulty identifying, or manipulative narrative-based formulations which are not obvious without longer inspection, which is usually

not feasible in the fast-paced fact-checking environment. Therefore, disinformation detection systems should incorporate mechanisms that allow users to flag instances where they believe the AI has made an incorrect decision and to offer feedback on specific truthfulness scores. Involving users in the system's ongoing development not only improves the accuracy of AI classifications but may also foster a sense of collaboration, where users feel more empowered and engaged in the fact-checking process. The AI Act mandates human oversight in high-risk applications like automated fact-checking, making transparency critical for accountability. Incorporating feedback mechanisms, where users can flag errors or provide input, not only enhances system transparency but also contributes to ongoing system improvements, as required by the DSA [46].

5.6 Providing Article Summarization and Audio Versions

Several participants suggested incorporating features such as article summarization and the ability to listen to audio versions of articles. These suggestions align with broader digital media trends, where users increasingly prefer to consume content in more flexible formats. AI systems could enhance usability by offering automatically generated summaries of articles, allowing users to quickly understand the key points. Additionally, providing a text-to-speech option, which enables users to listen to articles, would enhance accessibility and cater to a wider range of user preferences. Ensuring accessibility and inclusion in these features is crucial, as it allows people with diverse abilities and needs to engage with information on equal terms and promotes a more inclusive digital environment [8]. The AI Act supports inclusivity and accessibility, encouraging AI systems to be designed with features that cater to people with disabilities. Ensuring non-discriminatory access, including through text-to-speech options, aligns with the Act's goal of making AI systems more usable for all individuals, including those with disabilities.

5.7 Educating Users on AI System Limitations

Although detailed explanations reduced the need for further clarification, there is a risk of users becoming overly reliant on the system. Our findings suggest that explanations, while improving the perceived usefulness and understandability of the AI, do not necessarily lead to better human-AI performance and may even contribute to overdependence. Therefore, it is essential to educate users on the limitations of AI systems by including educational prompts or materials that explicitly highlight the potential limitations in the system's analysis, such as the possibility of false positives or contextual misunderstandings. Additionally, explanations may involve the encouragement of users to cross-check AI decisions with their own judgment. Providing this education could help promote critical thinking and mitigate the risk of over-reliance on algorithmically generated outputs, ensuring that critical human oversight remains central to the fact-checking process. This supports critical thinking and ensures that human oversight remains central to the fact-checking process, as mandated by the AI Act [47]. The recommendations and requirements outlined above provide a comprehensive framework for designing XAI features that enhance disinformation detection. By prioritizing transparency, reliability, and bias mitigation, XAI systems can improve user trust and engagement while addressing legal requirements set forth by the AI Act and DSA. The empirical validation of these recommendations will ensure that AI systems and human oversight work together to effectively combat mis- and disinformation.

6 Discussion and Conclusion

Overall, our results indicate that XAI features significantly enhance the (QoE) in fact-checking decision support systems, particularly in terms of perceived understandability and usefulness. Consequently, we can answer our research questions as follows:

RQ1: Can explanations enhance the perceived QoE in terms of understandability, and usefulness in a fact-checking system? V2 and V3 showed significantly higher ratings for both, understandability and usefulness in comparison with the baseline version V1. Thus, we can conclude that explanations help to increase the QoE. However, based on the findings from the crowdsourcing experiment, no conclusion can be drawn which XAI feature is more suitable. For V3 containing free-text explanations higher ratings have been observed, but they are not statistically significant.

RQ2: Are the provided information sufficient to determine the truthfulness of the given news items? The qualitative analysis of the open-ended questions shows that further information is required to make final judgments of whether the presented information is fake or not. Hereby, the source, original article, publication date, further references and citations and more information about the underlying functioning of the AI system are required to rely on the AI-system *truthfulness* rating.

Finally, we can conclude that XAI features increase the QoE in terms of understandability and usefulness, but further information is required to rely on AI-systems for the human-AI collaborative fact-checking task. However, while explanations improve user experience, they do not necessarily lead to better human-AI performance, and in some cases, they can foster over-reliance. To mitigate this, our recommendations emphasize the importance of transparency, reliability, and bias mitigation in designing XAI-enhanced systems, aligning with the legal obligations of the AI Act and the DSA. In particular, XAI systems should prioritize source credibility, integrate fact-checking tools, and address bias to improve user trust and engagement. Moreover, transparency regarding AI decision-making processes and robust feedback mechanisms are critical for system reliability and user collaboration. Enhancing usability through features such as article summarization and text-to-speech options further aligns with broader inclusivity goals, ensuring accessibility for all users, including those with disabilities. It is also essential to research the design and usability of XAI in fact-checking with a strong focus on Human-Computer Interaction; through this research insights into the successful implementation of these systems could be gained. Lastly, educating users about the limitations of AI systems is essential to prevent overdependence and to encourage critical human oversight, as mandated by the AI Act.

Overall, the integration of these recommendations will improve the effectiveness of disinformation detection systems and contribute to a better synergy between AI systems and human oversight in the fight against disinformation.

Acknowledgments

This research is funded by the Federal Ministry of Education and Research (BMBF) and is a collaborative effort of several research projects tackling mis- and disinformation detection and XAI. These projects are news-polygraph (03RU2U151C), VeraXtract (01IS24066), DeFaktS (16KIS1524K), VERANDA (16KIS2047), and the BIFOLD Agility Project FakeXplain.

References

[1] 2021. Proposal regulation: laying down harmonised rules artificial intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[2] Athira A.B., S.D. Madhu Kumar, and Anu Mary Chacko. 2023. A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence* 122 (June 2023), 106087. doi:10.1016/j.engappai.2023.106087

[3] 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, 19 pages.

[4] 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)*. IEEE, 519–524.

[5] Ramy et al. Baly. 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.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and machine learning: Limitations and opportunities. fairmlbook. org.

[7] Jordan et al. Boyd-Graber. 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.

[8] Oliver Hinz Wil M. P. van der Aalst Christof Weinhardt, Jonas Fegert. 2024. Digital Democracy: A Wake-Up Call. *Business & Information Systems Engineering* (2024), 127–134.

[9] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883* (2023).

[10] 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*. 1–13.

[11] 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*.

[12] Ziv et al. Epstein. 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 (May 2022), 183–193.

[13] Michael Gleicher. 2016. A framework for considering comprehensibility in modeling. *Big data* 4, 2 (2016), 75–88.

[14] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206.

[15] Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batarseh. 2023. Rationalization for explainable NLP: a survey. *Frontiers in Artificial Intelligence* 6 (2023), 1225093.

[16] Charlott Jakob, Pia Wenzel, Salar Mohtaj, and Vera Schmitt. 2024. Augmented Political Leaning Detection: Leveraging Parliamentary Speeches for Classifying News Articles. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*. 126–133.

[17] 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.

[18] 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.

[19] Himabindu Lakkaraju and Cynthia Rudin. 2017. Learning cost-effective and interpretable treatment regimes. In *Artificial intelligence and statistics*. PMLR, 166–175.

[20] Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems* 34 (2021), 29348–29363.

[21] 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.

[22] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[23] Rhema et al. Linder. 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.

[24] Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[25] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. 2023. Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *arXiv preprint arXiv:2310.19775* (2023).

[26] Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences* 12, 19 (2022), 9423.

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

[28] Chris Marsden and Trisha Meyer. 2019. *Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism.* European Parliament.

[29] Giovanni Da San et al. Martino. 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.

[30] Binny et al. Mathew. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 14867–14875.

[31] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[32] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[33] 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 (TiiS)* 11 (2021), 1–45.

[34] Sina et al. Mohseni. 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.

[35] Salar Mohtaj, Ata Nizamoglu, Premtim Sahitaj, Vera Schmitt, Charlott Jakob, and Sebastian Möller. 2024. NewsPolyML: Multi-lingual European News Fake Assessment Dataset. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*. 82–90.

[36] Linda Monsees. 2023. Information disorder, fake news and the future of democracy. *Globalizations* 20, 1 (2023), 153–168.

[37] Hao et al. Nie. 2019. Deep sequence-to-sequence entity matching for heterogeneous entity resolution. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 629–638.

[38] 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.

[39] Liangming et al. Pan. 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.

[40] Verónica et al. Pérez-Rosas. 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.

[41] 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. doi:10.5220/0011788900003417

[42] Hannah et al. Rashkin. 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.

[43] Ariana Sahitaj, Premtim Sahitaj, Salar Mohtaj, Sebastian Möller, and Vera Schmitt. 2024. Towards a computational framework for distinguishing critical and conspiratorial texts by elaborating on the context and argumentation with LLMs. *Working Notes of CLEF* (2024).

[44] Premtim Sahitaj, Ramon Ruiz-Dolz, Ariana Sahitaj, Ata Nizamoglu, Vera Schmitt, Salar Mohtaj, and Sebastian Möller. 2024. From Construction to Application: Advancing Argument Mining with the Large-Scale KIALOPRIME Dataset. In *Computational Models of Argument*. IOS Press, 229–240.

[45] Vera Schmitt, Balázs Patrik Csomor, Joachim Meyer, Luis-Felipe Villa-Areas, Charlott Jakob, Tim Polzehl, and Sebastian Möller. 2024. Evaluating Human-Centered AI Explanations: Introduction of an XAI Evaluation Framework for Fact-Checking. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation* (Phuket, Thailand) *(MAD '24)*. Association for Computing Machinery, New York, NY, USA, 91–100. doi:10.1145/3643491.3660283

[46] 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.

[47] 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 *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 2157–2174. doi:10.1145/3630106.3659031

[48] Luis Felipe Villa-Arenas, Ata Nizamoglu, Qianli Wang, Sebastian Möller, and Vera Schmitt. 2024. Anchored Alignment for Self-Explanations Enhancement. *arXiv preprint arXiv:2410.13216* (2024).

[49] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106.

[50] Qianli Wang, Tatiana Anikina, Nils Feldhus, Simon Ostermann, Sebastian Möller, and Vera Schmitt. 2024. Cross-Refine: Improving Natural Language Explanation Generation by Learning in Tandem. *arXiv preprint arXiv:2409.07123* (2024).

[51] 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.

[52] 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.

[53] 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.

[54] Xiang Zhou, Heba Elfardy, Christos Christodoulopoulos, Thomas Butler, and Mohit Bansal. 2021. Hidden Biases in Unreliable News Detection Datasets. http://arxiv.org/abs/2104.10130 arXiv:2104.10130 [cs].