Integrating Social Explanations Into Explainable Artificial Intelligence (XAI) for Combating Misinformation: Vision and Challenges

Yeaeun Gong, Lanyu Shang, and Dong Wang, *Senior Member, IEEE*

***Abstract*—This article overviews the state of the art, research challenges, and future directions in our vision: integrating social explanation into explainable artificial intelligence (XAI) to combat misinformation. In our context, “social explanation” is an explanatory approach that reveals the social aspect of misinformation by analyzing sociocontextual cues, such as user attributes, user engagement metrics, diffusion patterns, and user comments. Our vision is motivated by the research gap in the existing XAI that tends to overlook the broader social context in which misinformation spreads. In this article, we first define social explanation, demonstrating it through examples, enabling technologies, and real-world applications. We then outline the unique benefits social explanation brings to the fight against misinformation and discuss the challenges that make our vision complex. The significance of this article lies in introducing the “social explanation” concept in XAI, which has been underexplored in the previous literature. Also, we demonstrate how social explanations can be effectively employed to tackle misinformation and promote collaboration across diverse fields by drawing upon interdisciplinary techniques spanning from computer science, social computing, human–computer interaction, to psychology. We hope that this article will advance progress in the field of XAI and contribute to the ongoing efforts to counter misinformation.**

***Index Terms*—Explainable artificial intelligence (XAI), misinformation, social explanation, sociocontextual cue.**

I. Introduction

THE proliferation of digital communication platforms (e.g., Twitter/X, Meta, and YouTube) has facilitated the swift spread of online misinformation [1]. The consequences of misinformation spread are serious, eroding trust among individuals [2], fostering public anxiety [3], and exacerbating the polarization of opinions [4]. Finding ways to address misinformation is, therefore, crucial at both the individual and societal levels. Various solutions have been proposed, including implementing inoculation measures [5], promoting media literacy [6], and employing fact-checking [7] and content moderation practices (e.g., removing posts with misinformation, using flags, and warning labels) [8]. However, these solutions require a significant amount of time, resources, and long-term investments, such as educational programs and trained personnel. Alternatively, improving the accuracy of misinformation detection models has been suggested as a time- and resource-saving approach [9], [10]. Yet, these models often face user distrust, mainly due to the complex and opaque nature of the underlying algorithms.

Therefore, explainable artificial intelligence (XAI) has emerged as a promising tool in combating misinformation [11], [12]. By presenting explanations alongside AI’s decisions, XAI enables users to grasp the underlying rationale behind the AI’s decision-making process. This improved understanding not only increases trust in AI systems but also discourages misinformation spread by fostering critical analysis [13] and creating a reluctance to disseminate information without proper evidence [14]. For example, techniques like local interpretable model-agnostic explanations (LIMEs) and SHapley additive exPlanations (SHAPs) are two widely explored XAI techniques that are known to enhance interpretability in deep learning models [15]. By providing localized explanations for individual predictions, LIME has proven effective in making AI decisions more transparent and strengthens confidence in the accuracy and reliability of model outcomes [16]. Recent efforts have also been made to integrate XAI with a user-centered perspective within human–computer interaction (HCI). Techniques such as human-understandable explanations through visualization [17], [18] and interfaces for exploring and interrogating the decision-making processes of misinformation detectors [19], [20] aim to enhance users’ understanding and enable them to make informed decisions.

However, a key limitation of current XAI systems is their narrow focus on content-based explanations, which primarily analyze the content, features, and structures within the misinformation itself [18], [19]. While XAI helps users understand the AI model, they overlook the broader *social context* in which misinformation spreads and is accepted. Insights from cognitive science and psychology indicate that when individuals present or evaluate explanations, they incorporate specific cognitive biases and cultural norms that are influenced by the social context [21]. Moreover, misinformation is not always

Manuscript received 28 October 2023; revised 12 February 2024 and 15 April 2024; accepted 19 May 2024. This work was supported by the National Science Foundation under Grant IIS-2202481, Grant CHE-2105032, Grant IIS-2130263, Grant CNS-2131622, and Grant CNS-2140999. *(Corresponding author: Dong Wang.)*

The authors are with the School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL 61820 USA (e-mail: yegong2@illinois.edu; lshang3@illinois.edu; dwang24@illinois.edu).

Digital Object Identifier 10.1109/TCSS.2024.3404236

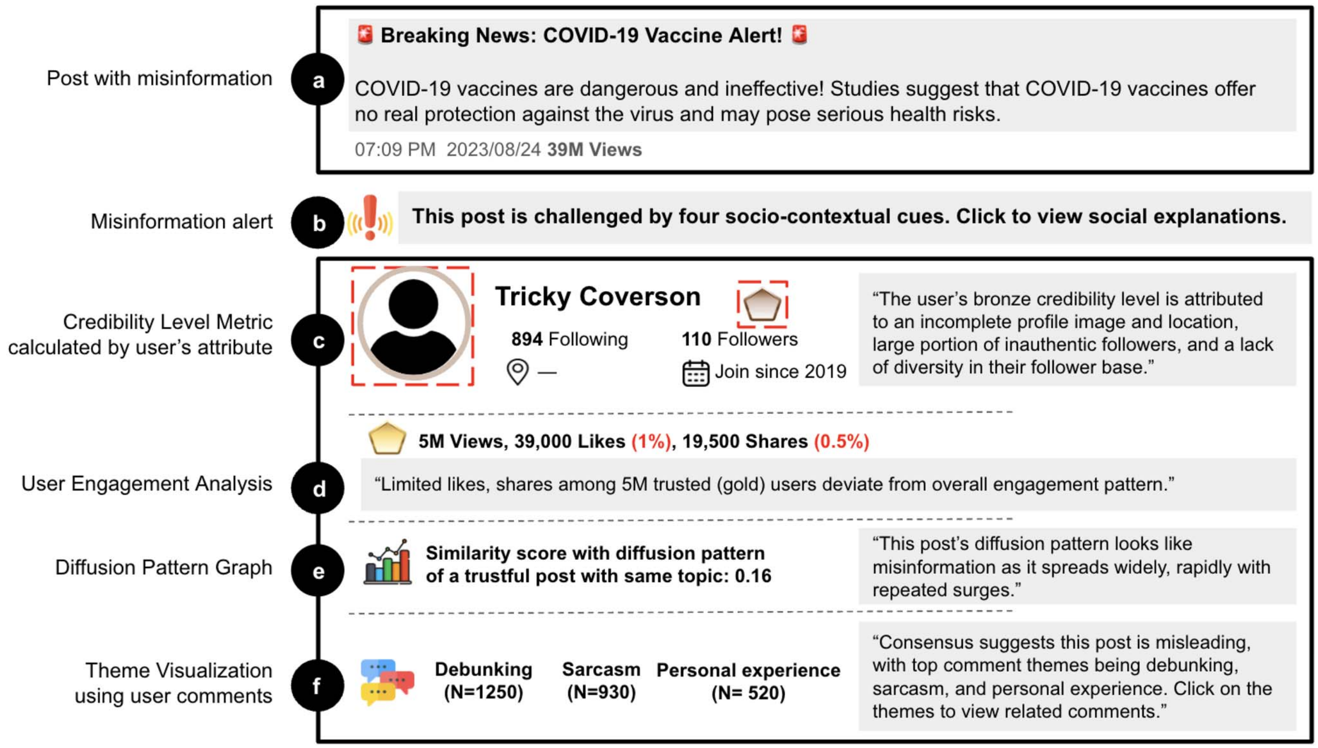
binary; it often includes statements that are both partially true and false [22]. Often, seemingly valid data might be presented without its crucial surrounding context, resulting in misinterpretations. These complexities highlight the importance of explanations that not only focus on content but also probe deeper into the social contexts surrounding misinformation. Building upon these findings, scholars have started advocating for the inclusion of social context in XAI, including the integration of social transparency into AI user interactions [23] and addressing the sociotechnical gap in XAI systems [24]. However, these studies have primarily focused on proposing perspectives or frameworks across broad domains of AI systems, rather than addressing a specific domain of misinformation. Recognizing the substantial role that misinformation plays in influencing how individuals form their opinions and make choices on online platforms [25], along with the influence of cognitive, social, and affective factors that lead to the endorsement of misinformation [14], it becomes evident that an expanded XAI framework that incorporates psychological and social contexts is crucial for more comprehensive mitigation of misinformation.

One promising approach to explaining the social context of misinformation is to utilize sociocontextual cues. In this article, sociocontextual cues refer to observable indicators that provide insight into the social dynamics and factors surrounding misinformation. Broadly, four types of sociocontextual cues have been employed in misinformation detection models: user attributes (e.g., age and gender) [26], user engagement metrics (e.g., number of likes and shares) [27], [28], patterns of misinformation spread (e.g., wide and rapid propagation with repeated surges) [29], [30], and user comments indicating the presence of misinformation [31], [32]. These cues have proven to enhance the accuracy of misinformation detection models by capturing the social dynamics associated with misinformation [31]. User comments, for instance, provide valuable signals for detection models as users often express negative emotions like anger and sarcasm when encountering misinformation. They also present credible sources, scientific evidence, and alternative perspectives, enabling a more comprehensive evaluation of the information's veracity. However, beyond their use in detection models, we envision that sociocontextual cues could also be leveraged in *explanations* to provide a deeper understanding of the social contexts of misinformation. Therefore, we introduce the concept of "social explanation," which utilizes sociocontextual cues during the explanation process to complement the current state of XAI in combating misinformation. Specifically, this article focuses on text-based misinformation, which is the main form of misinformation spread on digital platforms that has been extensively studied in the literature [14].

Fig. 1 shows an example of the utilization of four key sociocontextual cues to offer a social explanation. These cues include user attributes [Fig. 1(c)], engagement metrics [Fig. 1(d)], diffusion patterns [Fig. 1(e)], and user comments indicating misinformation [Fig. 1(f)]. By leveraging these sociocontextual cues, a comprehensive understanding of the social context surrounding misinformation, such as a post about COVID-19 [Fig. 1(a)], can be provided. Fig. 1(c) presents a credibility metric that categorizes users into bronze (low credibility), silver (medium credibility), and gold (high credibility) levels based on user's behavioral attributes (e.g., profile status and social networking activities). Fig. 1(d) demonstrates the use of engagement metrics, particularly the engagement of trusted users (gold-level users), to evaluate the information's trustworthiness. For example, if a post receives minimal engagement from trusted users but significant engagement from others, it suggests potential misinformation. Fig. 1(e) displays the diffusion pattern graph of a misinformation post, explaining why it is likely misinformation based on its diffusion pattern. Additionally, it introduces the similarity score, a metric that measures the degree to which the diffusion pattern of the misinformation post resembles that of a trustworthy post on the same topic. Last, Fig. 1(f) visualizes themes derived from user comments, such as debunking and sarcasm, offering insights into users' attitudes related to misinformation, as well as reliable sources.

Incorporating social explanations into XAI offers several unique benefits in misinformation explanations. First, it provides supplementary information that traditional XAI approaches lack, enabling a deeper comprehension of the social and contextual aspects of misinformation. For example, the theme visualization in Fig. 1(f) illustrates how it facilitates access to user comments flagging misinformation, capturing nuanced contextual interpretations—such as underlying sarcastic tones and personal anecdotes offering real-life perspectives—that content-based methods might miss. Second, social explanation gives users insights into the factors that contribute to the credibility of information. For instance, in Fig. 1(c), a credibility metric reveals how user attributes like profile completion and follower diversity impact their credibility levels. Third, social explanation enhances user trust by shedding light on the AI system's internal processes. Understanding that a post is labeled as misinformation due to its spread pattern, as shown in Fig. 1(e), clarifies the model's decision-making process, leading to increased confidence and trust in the model [33]. Fourth, social explanation empowers users to take action against misinformation. By observing user comments reflecting community sentiments, like debunking in Fig. 1(f), individuals can be encouraged to engage in reporting and fact-checking activities to better align with the social norms of responsible information sharing [34].

This article also presents many research challenges related to social explanations in misinformation explanations. One challenge is the data reliability challenge, which involves handling noisy data (e.g., unrelated discussions and personal anecdotes) that hinder the accurate identification of comments that offer valuable insights for understanding misinformation. The second challenge focuses on the evolving patterns in misinformation and social explanation as both the misinformation and social explanation change over time, across various contexts, and through user interactions. The third challenge pertains to the nonbinary nature of misinformation, which necessitates a more nuanced approach to distinguish between what is factual and what is misleading. The fourth challenge revolves around effectively integrating social explanations into XAI frameworks, such as addressing conflicts that may arise between social explanations and content-based explanations. The fifth challenge



Breaking News: COVID-19 Vaccine Alert!

COVID-19 vaccines are dangerous and ineffective! Studies suggest that COVID-19 vaccines offer no real protection against the virus and may pose serious health risks.

07:09 PM 2023/08/24 **39M Views**

This post is challenged by four socio-contextual cues. Click to view social explanations.

Tricky Coverson

894 Following **110** Followers

— Join since 2019

"The user's bronze credibility level is attributed to an incomplete profile image and location, large portion of inauthentic followers, and a lack of diversity in their follower base."

5M Views, 39,000 Likes (1%), 19,500 Shares (0.5%)

"Limited likes, shares among 5M trusted (gold) users deviate from overall engagement pattern."

Similarity score with diffusion pattern of a trustful post with same topic: 0.16

"This post's diffusion pattern looks like misinformation as it spreads widely, rapidly with repeated surges."

Debunking (N=1250) **Sarcasm (N=930)** **Personal experience (N= 520)**

"Consensus suggests this post is misleading, with top comment themes being debunking, sarcasm, and personal experience. Click on the themes to view related comments."

Fig. 1. Illustration of social explanations for a misinformation post: (a) post containing misinformation; (b) misinformation alert based on sociocontextual cues; and (c)–(f) design examples of social explanations that use different sociocontextual cues.

involves designing user interfaces that can effectively present social explanations while protecting the users' privacy. Last, the ever-changing nature of misinformation and varied user responses to social explanations make establishing effective evaluation criteria for the social explanation of misinformation another challenge.

The rest of the article is organized as follows. In Section II, we provide a comprehensive overview of the application of sociocontextual cues for detecting misinformation and the research on social explanations. In Section III, various examples of social explanation are provided, accompanied by the enabling technologies. Section IV demonstrates how social explanations are applied across various domains. Section V explores the unique research challenges that can inspire researchers to explore future research directions under the vision of this work. Finally, we conclude with our concluding remarks, emphasizing the importance of our vision.

II. Related Work

This section reviews studies that utilize sociocontextual cues (user attributes, user engagement metrics, diffusion patterns, and user comments) to detect misinformation. The primary aim is to understand how these sociocontextual cues can be used and effectively integrated into the explanation process. Additionally, we delve into existing research on social explanations and related concepts, predominantly studied within the recommendation systems. What distinguishes our work from previous studies on social explanations is our focus on leveraging social explanations as a powerful tool to counter misinformation, in contrast to the traditional emphasis on enhancing persuasiveness (i.e., the degree to which explanations enhance user engagement) and informativeness (i.e., user satisfaction with the recommended item) of system recommendations through social explanations. Last, we review papers that serve as the theoretical foundations of social explanations, which are social influence theory, source credibility theory, and information cascade theory.

A. Utilizing Sociocontextual Cues in Misinformation Detection

1) User Attributes: Several studies have identified user attributes that can distinguish individuals who are more likely to share misinformation from those who share reliable information. For example, factors such as verified user status, account age, gender, age, and personality traits have been identified as potential indicators for identifying users with a higher likelihood of sharing misinformation on social media [35]. Building upon the study, additional user attributes such as location, political bias, and profile image were explored, revealing distinct patterns associated with misinformation sharing [36]. They also identified registration time, verified user status, political bias, and personality as the four most influential attributes in predicting misinformation-sharing behavior. The integration of speaker profile cues like party affiliation, speaker title, location, and credit history was also found to significantly improve misinformation detection models, surpassing the state-of-the-art approach by 14.5% [37].

While various user attributes have demonstrated their association with the likelihood of sharing misinformation, they remain untapped in the explanation process. Incorporating attributes like verified user status into the explanation process can provide valuable insights into how these factors influence the credibility of shared information. However, it is crucial to recognize the potential bias introduced by specific user attributes, such as location, personality, and political affiliation, when presented to users. For instance, users belonging to a particular political group might be perceived as more prone to spreading misinformation when a social explanation indicating the group's higher frequency of sharing misinformation is presented to the broader audience, leading to the stigmatization of their views or opinions. To ensure fairness, the explanation process should avoid reinforcing stereotypes or biases while elucidating user attributes.

2) User Engagement Metrics: User engagement metrics, such as the history of message contributions, following–followers ratio, and social network dynamics, play a crucial role in identifying misinformation. For example, low-credible news is often found to be disseminated by users with a limited history of message contributions [38]. This suggests that individuals who have not actively participated in online discussions are more likely to propagate unreliable information. The incorporation of metrics like the following–followers ratio has been demonstrated to significantly enhance the classification accuracy of misinformation detection, outperforming the state of the art by 14% in image misinformation detection [39]. Additionally, user interaction patterns are vital in distinguishing between misinformation and trustful news. For example, behavioral features, including discussion initiation, interaction engagement, influential scope, relational mediation, and informational independence, have been identified as more valuable than linguistic characteristics in the detection of misinformation [40].

By providing explanations that include these engagement metrics, users can develop a deeper understanding of the dynamics of misinformation propagation and the role played by individual users in spreading misinformation. For example, understanding user interaction patterns, such as discussion initiation and engagement, can reveal how the dissemination of misinformation is influenced by specific users' actions on the platform. This enhanced understanding fosters individual users' awareness of their impact on spreading misinformation, empowering them to become more responsible in their interactions with content. Consequently, users are more likely to fact-check information before sharing it and actively avoid unknowingly amplifying misinformation, thus contributing to a more trustworthy online environment.

3) Diffusion Patterns: Understanding the information propagation pattern is crucial for detecting misinformation, as misinformation tends to spread more rapidly, deeply (with dense network connections), and widely compared to trustworthy news [41], [42]. Additionally, misinformation exhibits distinct patterns of intermittent hibernation periods followed by multiple bursts of activity, which is not commonly observed in reliable news [43]. These differences in spread can even be identified in the early stages of propagation [44]. The diffusion pattern of misinformation can be attributed to the influence of social homophily, where individuals tend to connect with like-minded people. Consequently, network-related characteristics like ego density, triad density, and community density have also been used to enhance the detection models' accuracy [45]. Particularly in emerging topics or when training data are scarce, these network information have proven effective by achieving a significant 13.09% increase in F1 score compared to state-of-the-art model, even without content information and with just 10% of the training data [46].

Examining diffusion patterns in the explanation process allows users to acquire valuable insights into the differential spread of misinformation compared to reliable information, enabling a deeper grasp of the factors influencing its dissemination. For example, it can reveal the rapid and widespread nature of misinformation propagation, showing how misleading information rapidly reaches a vast audience using bots and automated accounts operating within densely interconnected networks. By becoming aware of these mechanisms that facilitate the traction of misinformation, users can enhance their ability to identify and critically evaluate potentially misleading claims, for example, by looking at whether real individuals versus automated bots share the information.

4) User Comments: Research has identified specific patterns and language commonly observed in user comments that serve as indicators of misinformation. These indicators include expressions of misinformation awareness, such as skepticism and acknowledgment of the information, the frequent use of emojis and swear words, as well as the use of sarcasm to debunk misinformation [47]. Additionally, instances of misinformation have been associated with the presence of debunking echo chambers, diverse opinions, support for verification, and the use of distinctive vocabulary [31]. Incorporating these indicators into content-based solutions has demonstrated improved accuracy and reduced detection time. For example, a fauxtography detection method that focuses on analyzing user comments over the actual image content has outperformed the existing baselines, achieving a 5.4% higher accuracy than the state-of-the-art methods [31].

Extending prior research that mainly concentrated on the detection of misinformation, user comments were utilized to not only detect but also explain misinformation [32]. They introduced comment-driven explanations by extracting the top-k user comment lists based on the highest attention weights from their model. These explanations outperformed existing methods for explaining AI systems, showing the potential of user comments in offering comprehensive explanations of the misinformation detection results.

B. Social Explanation and Related Concepts

1) Social Explanations in Recommendation Systems: Social explanations have been primarily studied in the context of recommendation systems, which leverage information from social networks to enhance their recommendations [48]. Specifically, the information from social networks encompass users' familiarity (e.g., "a close acquaintance gave this product a high rating"), similarity networks (e.g., "users with similar interests highly rated this product") [49], [50], and domain expertise (e.g., "experienced users in this category frequently give high ratings to this product") [51]. Popularity metrics, such as the number of likes or shares on social media platforms, are commonly used in social explanations [52]. Research has shown that these social explanations have an impact on both persuasiveness and informativeness in various recommendation domains, including music recommendation [48], scholarly recommendation [53], e-commerce recommendation [54], and privacy setting recommendation [55]. In music recommendation, for example, knowing that friends or users with similar taste profiles prefer a particular song or artist can affect an individual's decision to explore or adopt that recommendation [48]. Similarly, in the context of privacy setting recommendations, when users recognize that experts recommend certain privacy settings, they tend to be more willing to follow these recommendations, as they interpret the recommendation as reliable and trustworthy information [55].

2) Social Influence Theory of Social Explanation: The effectiveness of social explanations can be attributed to the social influence theory. According to this theory, individuals are naturally inclined to conform to social norms and seek validation from others [34]. For instance, when users notice that an item is popular among others in their social network, as indicated by metrics like the number of likes or purchases, they are more likely to be influenced by social norms, the tendency of individuals to align their opinions, decisions, and behaviors with those of others [48], [56], [57]. Additionally, social explanations can utilize social proof heuristics, wherein people rely on the actions and opinions of others to guide their behavior in uncertain situations [58]. For instance, users may be more inclined to engage with a post or share it if they observe that the post has received many likes, interpreting it as interesting or relevant based on the others' responses.

3) Social Explanations Beyond Recommendations to Combat Misinformation: An unresolved question remains regarding presenting social explanations within the context of misinformation. While recommendation systems primarily focus on individual opinions and preferences [48], [56], the main challenge in addressing misinformation lies in accurately assessing the veracity of the information itself. This raises the need to explore whether social explanations demonstrate similar effects as seen in recommendation systems, particularly in scenarios where objectivity holds greater importance than subjective preferences [59]. Moreover, building on the misinformation detection literature [38], [41], [47], [60], we can expand the scope of social explanation of misinformation beyond the conventional use of social network information in recommendation systems. Alongside social networks, other sociocontextual cues, such as user attributes, diffusion patterns, and user comments, can be harnessed to construct social explanations for better understanding and combating misinformation, as previously discussed.

C. Theoretical Foundations of Social Explanations of Misinformation

1) Social Influence Theory: Social influence theory explores how an individual's thoughts, beliefs, and behaviors are shaped by their social environment [61]. This theory identifies two key forms of social influence, each driven by distinct motivations: informational social influence, where individuals conform based on others' perceived expertise in seeking accuracy, and normative social influence, where the drive is to align with group norms for social acceptance [61]. These forms of social influence significantly impact perceptions and acceptance of misinformation, providing a framework for understanding why certain content gains traction. For instance, content perceived as popular or credible, often indicated by likes and shares, is more likely to engage users and become widespread due to increased perceived credibility and the desire to align with group norms [62].

Understanding the impact of informational and normative social influences in the spread of misinformation, we incorporate these influences in our social explanation design. We utilize informational social influence by creating mechanisms that evaluate and showcase the credibility of users [63], tapping into the natural human inclination to trust dependable sources. Also, we employ normative social influence to cultivate an environment that prioritizes critical thinking and diligent fact-checking [64], thus encouraging a culture that values accuracy and scrutinizes unverified information. These informational and normative influences are incorporated in our social explanation designs of user credibility metric and thematic visualization of user comments, respectively. In user engagement analysis, we differentiate between these forms of social influence to provide a more detailed analysis of what motivates users to interact with content (e.g., whether interactions are driven by users' desire for accurate information or by their adherence to community norms). More details on these applications are discussed in the corresponding social explanation designs in Section III.

2) Source Credibility Theory: Source credibility theory, emphasizing how users perceive the credibility of a source itself, is crucial in determining the effectiveness of communication and its persuasive power [65]. Central to this theory is the idea that a user's perception of a source's credibility significantly shapes their immediate reaction and opinion change about the content presented [65]. Additionally, the theory posits that sources deemed credible are often viewed as more accurate, thus garnering greater attention and fostering additional information-seeking behavior [66]. Unlike informational social influence in social influence theory, which focuses on individual conformity to the perceived expertise and opinions of others within their social environment [61], source credibility theory emphasizes the intrinsic value of the source itself. In our social explanation design, we integrate these principles by introducing a user credibility level metric. This metric assesses users' credibility based on criteria that can be derived from the theory, such as demonstrated expertise, reliability of information shared, and the authenticity of their online profiles.

3) Information Cascade Theory: Information cascade theory explores the significant influence of others' actions and choices on individual decision-making, resulting in a collective behavior termed "information cascade" [67]. This theory is crucial for understanding misinformation spread, as the theory illustrates the tendency of people to share or accept information based on observed behaviors, often bypassing independent verification or critical thinking [67]. Studies have demonstrated that information cascades can have a profound effect on public opinion and behaviors, swiftly propelling the adoption of specific beliefs or trends across networks [68], [69]. Utilizing information cascade theory in the context of misinformation provides a strategic framework for the development of our social explanation designs. For example, we applied this theory to create a predictive model that identifies potential misinformation cascades in their early stages [70]. This model analyzes patterns of information sharing within a network, flagging instances where the rapid spread of information occurs alongside markers of misinformation, such as the presence of an echo-chamber effect [70]. By understanding the dynamics of these cascades, we can also create interventions that guide the flow of information toward accuracy and reliability. As one example, we designed the "information diffusion graph," a visual representation designed to trace how both accurate and misleading information spread through networks. Further elaboration on the diffusion pattern graph and its application is provided in Section III, specifically under Section III-B3.

D. Incorporating Research Findings into Social Explanation Design

The social explanation designs we introduce in this article builds upon the existing literature in misinformation detection that leverage sociocontextual cues such as user attributes, engagement metrics, diffusion patterns, and user comments. We specifically translate each of these cues into practical social explanations, which are credibility level metric, user engagement analysis, diffusion pattern graph, and theme visualization of user comments. For instance, the development of our credibility level metric is grounded in the literature that elucidates the link between specific user attributes and misinformation spread, guiding our method for calculating credibility scores. In addition, to assess the effectiveness of social explanations in mitigating misinformation, we examine their use in recommender systems, where social explanations have been extensively studied, and explore social influence theory as a potential mechanism underpinning their effectiveness on increasing persuasiveness and informativeness of the system. Additionally, we incorporate insights from other relevant social science theories, such as source credibility theory and information cascade theory, both to enhance our understanding of the efficacy of social explanations in combating misinformation and to further inform our social explanation designs.

III. Vision: Social Explanations of Misinformation

In today's digital age, misinformation is not merely a result of misleading content. Instead, misinformation is also significantly influenced by the social contexts, behaviors, and interactions where information circulates [14]. Central to our vision is the concept of "social explanation," which elucidates the socio-contextual cues surrounding misinformation. In this section, we define social explanation, present its key elements and examples, and explore the enabling technologies that make our vision feasible.

A. Definitions and Key Elements of Social Explanation

This study defines "social explanation" as an explanatory method within XAI that utilizes social contextual cues to explain the reasons why a particular post is classified as misinformation. In this definition, social contextual cues are heterogeneous, ranging from user demographics to interpersonal interactions and broader societal influences. Providing social explanation serves dual purposes. First, it delineates the underlying social and contextual factors that play a role in a post being flagged as misinformation, fostering a more nuanced understanding of the misinformation classification. Second, it encourages individuals to actively combat misinformation by leveraging individuals' natural tendency to follow social norms [48], especially when they observe the social explanations created by others. For example, when prevalent themes in comments flagging misinformation are presented, this observed consensus can prompt users to engage in their own fact-checking and add substantiated viewpoints to provide more evidence, thereby fortifying the community's collective fight against misinformation. Social explanation aims to supplement the existing states of XAI, offering a more comprehensive view of the dynamics surrounding misinformation.

B. Examples of Social Explanation

Below, we present examples of social explanations, including a credibility level metric, user engagement analysis, diffusion pattern graph, and theme visualization of user comments. We also describe how each of these designs is informed by three theories, i.e., social influence theory, source credibility theory, and information cascade theory.

1) Credibility Level Metric: The user's credibility level metric, derived from user attributes, can be used to provide social explanations of misinformation. By leveraging prior research on user attributes, we can compute a credibility metric for each user and provide an explanation for that metric. For instance, studies have shown that users are more likely to share misinformation when their profile image appears inauthentic or when specific characteristics are present in their social network (e.g., a high proportion of followers with strong political biases) [71]. Fig. 2 illustrates the credibility metric calculated based on the user attributes used in the detection model, with trustworthy users depicted at the top of the Fig. 2 and untrustworthy users shown at the bottom of the Fig. 2. Trustworthy users are characterized by authentic profile images, diverse followers from various backgrounds, and active engagement in the social network. Conversely, untrustworthy users may have a nonactual profile image (e.g., manipulated photo, graphical representation, and

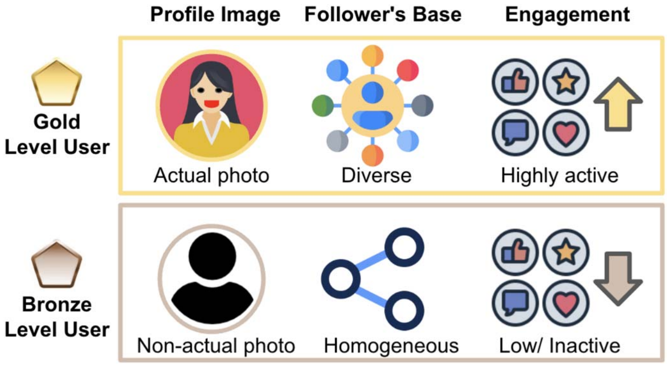


Fig. 2. User credibility level metrics.

avatar), followers with strong political biases, and many inactive or fake accounts.

The credibility level metrics allow users to assess the trustworthiness of a post or comment based on the author's credibility level, helping them determine the authenticity of the information. If they encounter content from a low-credibility user, they can exercise caution in verifying its accuracy. Additionally, the user profile page promotes transparency by offering a comprehensive description of the factors contributing to their credibility level. This transparency fosters trust between users and the misinformation detection system, enabling them to verify the credibility assessment process and gain greater confidence in the system's capability to differentiate reliable content from unreliable one. Moreover, users are empowered to share accurate information in their posts and comments as they become more aware of their credibility level, fostering a sense of responsibility within the community [72]. This collective responsibility often leads to a more informed and trustworthy exchange of information among users [73].

a) Applying Source Credibility Theory and Social Influence Theory: The credibility metric serves as a practical application of source credibility theory, quantifying the trustworthiness of users based on observable attributes such as profile authenticity, network characteristics, and engagement patterns. By assigning credibility levels (e.g., gold, silver, and bronze) to users, the credibility metric enables others to quickly assess the trustworthiness of information based on the credibility of its source, namely the user and the author of the post. Additionally, this metric aligns with the principles of social influence theory, particularly with regard to informational social influence. According to informational social influence, people often look to these high-credibility individuals for guidance on what to share and believe. In this regard, users with high credibility (gold level) may serve as influencers within their networks, setting standards for information quality and reliability.

2) User Engagement Analysis: Engagement metrics on platforms such as Meta and Twitter/X, represented by likes, shares, and comments, not only reflect content's popularity but also serve as tools to elucidate the social contexts related to misinformation. During a COVID-19 outbreak, for example, a post claiming "The alcohol in vodka, acting as a sterilizer, can neutralize COVID-19 viruses in the throat" gained traction on social media platforms [74]. When such a post receives numerous likes and shares but lacks in-depth discussion or expert input to prevent the spread of misinformation [14], it suggests a societal preference toward simple health advice. This preference, possibly due to the public's desire for easy health remedies and a collective optimism about combating COVID-19, can be explained by a pervasive need for reassurance and control during uncertain times [74]. Furthermore, by analyzing the user engagement patterns during the initial spread of the post about vodka-as-a-sterilizer, we can gain deeper insights into the motivations that fuel misinformation spread. For instance, the analysis can reveal how the rapid amplification within the alcohol community was facilitated by the endorsement of a well-known influencer, whose perspectives strongly resonated with their followers, potentially due to shared interests or affiliations. This demonstrates how niche communities can sometimes serve as catalysts, amplifying misinformation swiftly and extensively due to their inherent interests and biases, before misinformation permeates the general populace [75]. Thus, by explaining engagement patterns, including elucidating which groups are first to amplify a message and their underlying motivations, we can understand the societal biases behind the spread of misinformation.

a) Applying Social Influence Theory: Our design of user engagement analysis is directly influenced by social influence theory, which provides a detailed framework for understanding the dynamics behind social media post popularity. This theory distinguishes between informational social influence, driven by the content's perceived expertise or factual accuracy, and normative social influence, driven by peer pressure or the desire to conform to group norms. Inspired by this distinction, our user engagement analysis is tailored to help people critically distinguish between engagement that denotes reliability and that might indicate susceptibility to misinformation influenced by social conformity. For example, to assist people in discerning these motivations and understanding whether engagement reflects genuine credibility or mere social conformity, we can analyze engagement quality, focusing on discussion depth, expert input, and community sharing patterns.

3) Diffusion Pattern Graph: Diffusion pattern graphs can be viewed as a form of a visual social explanation that depicts how trustworthy and misleading posts spread. By charting the trajectory of information spread, diffusion pattern graphs elucidate the underlying social contexts and mechanisms driving the spread of misinformation. For instance, Fig. 3 highlights that misinformation spreads widely and rapidly through repeated bursts, capturing the attention of numerous users within a short period due to factors such as bot activity and celebrity endorsement. In contrast, credible information typically exhibits a more gradual and sustained dissemination, with its impact fading over time and fewer bursts of attention being observed. Building upon this analysis, we can leverage these diffusion patterns to develop a predictive model. This model would proactively identify potential misinformation cascades at their onset by analyzing early sharing patterns within networks, particularly looking for rapid spread combined with known markers of

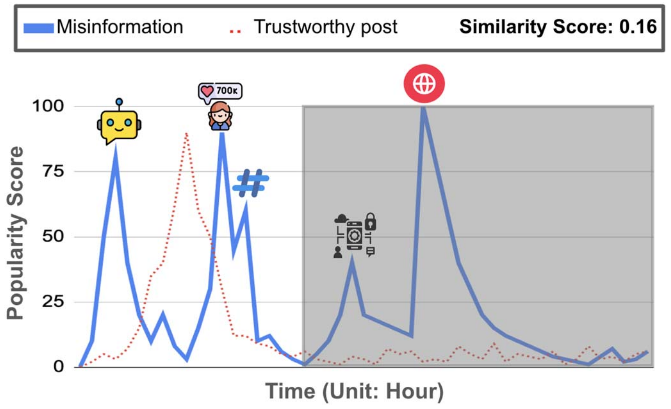
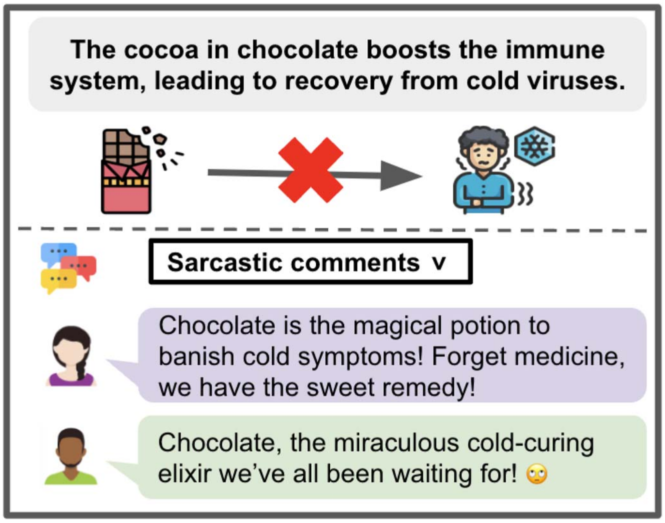


Fig. 3. Diffusion pattern graph. Here, the misinformation graph experiences rapid bursts caused by bot activity, celebrity endorsement, viral hashtags, algorithmic amplification, and fake news websites. The gray area highlights the significant contrast in the extent to which misinformation spreads compared to trustworthy posts.


The cocoa in chocolate boosts the immune system, leading to recovery from cold viruses.

Chocolate is the magical potion to banish cold symptoms! Forget medicine, we have the sweet remedy!

Chocolate, the miraculous cold-curing elixir we've all been waiting for! 🙄

Fig. 4. Theme visualization of user comments. Exemplary comments for sarcastic debunking.

misinformation like echo-chamber effects. In addition, a similarity score can be utilized, which is derived from comparing dissemination patterns between a misleading post and a trustworthy one on the same topic, to help users quickly identify the post's veracity [76].

Understanding and engaging with the social explanations derived from diffusion patterns offers benefits to users in terms of media literacy. Media literacy involves the skills of accessing, analyzing, evaluating, and creating media content [77]. By grasping diffusion patterns, users gain insights into how information spreads and becomes popular. This knowledge empowers users to differentiate between credible and misleading information, identify common tactics used in misinformation campaigns, and critically assess the sources and content they come across. For instance, when users learn that repeated popularity spikes can result from unverified bots, they can be more cautious when encountering new information by checking the source of information before accepting it as accurate or sharing it further, ultimately reducing their vulnerability to misinformation [14].

a) Applying Information Cascade Theory: The diffusion pattern graph, drawing on insights from information cascade theory, emphasizes the impact of early shares on individual behavior, directing our attention to initial sharing activities and swift surges in engagement as key indicators of misinformation cascades. This graph methodically tracks the speed and trajectory of information dissemination, particularly through influential users or closely knit communities. Its design allows users to discern between typical information flow and potential misinformation outbreaks. By harnessing the theory's understanding of the social dynamics behind information spread, the diffusion pattern graph provides a lucid visualization of cascading patterns. This not only improves user awareness but also helps them recognize and understand the pathways through which misinformation can proliferate.

4) Theme Visualization of User Comments: Visualizing user comments, categorized by themes, provides users with a deeper insight into the contextual aspects of misinformation through flagged comments. Fig. 4 shows themes aligned with specific flags for misinformation, such as a sarcastic rebuttal. Clicking on a particular theme allows users to access related comments falling under that category. For instance, a sarcastic comment mocking a dubious claim about chocolate curing colds highlights users' skepticism toward implausible health claims, encouraging critical evaluation of health-related information in a lighthearted manner. Themes can also be organized into categories like "expert opinion" and "personal experience," each offering a distinct lens into the social explanation of misinformation. While comments in the "expert opinion" category may anchor discussions based on authoritative sources, those in "personal experience" may expose the contexts, origins, or stories that made the individual skeptical about the information's truthfulness.

Visualizing user comments based on thematic categories is a vital tool in the fight against misinformation, facilitating the swift identification of diverse opinions and themes that indicate possibly deceptive content. By highlighting themes often associated with misinformation, such as debunking and skepticism, users are collectively encouraged to challenge misinformation. This collective effort can be further augmented when integrated into a community-driven approach like Twitter's Birdwatch [78]. Observing themes and accompanying opinions from others not only prompts users to explore further evidence but also to contribute from their own knowledge to validate the content's veracity [72]. In this way, the thematic visualization serves as an intuitive window into the societal underpinnings of misinformation, fostering a more public discourse and enhancing the community's ability to combat misinformation.

a) Applying Social Influence Theory: Our theme visualization for user comments draws directly from social influence theory, particularly leveraging normative social influence. By accentuating comments that scrutinize or refute dubious claims, this theme visualization applies normative pressures to promote critical engagement and guide community norms toward a greater appreciation of evidence and diverse viewpoints. Furthermore, the deliberate emphasis on a broad spectrum of perspectives not only makes these views more prominent but also actively encourages users to engage in meaningful examination and dialogue. Such a strategy bolsters the community's collective prowess in critically evaluating information, thereby fortifying defenses against misinformation.

C. Enabling Technologies

Actualizing our vision of integrating social explanations into XAI for combating misinformation requires a fusion of cutting-edge technologies. This section delves into the enabling technologies that power social explanations, encompassing aspects of content, context, and presentation.

1) Natural Language Processing (NLP): NLP technologies play a crucial role in analyzing misinformation content by offering advanced tools to dissect user comments and posts. Sentiment analysis, for example, enables the extraction of individuals' viewpoints, ideas, and emotions [79], offering insights for developing social explanations that reflect the prevailing public sentiments surrounding misinformation. Topic modeling assists in categorizing comments into misinformation-related themes by detecting recurring phrases and context clues to reveal shared themes surrounding the content [80]. Named entity recognition enhances credibility assessment by identifying and cross validating the credibility of named entities such as individuals, organizations, and places in a text against trusted sources [81]. Event detection's ability to identify and categorize underlying events from textual data is pivotal in generating social explanations that elucidate the genesis and spread of misinformation, particularly when misinformation is linked to specific events such as natural disasters [82], criminal activities [83], and significant traffic accidents [84]. Moreover, using a Q&A system to analyze repetitive question patterns over time can reveal the underlying causes of shared uncertainties and concerns surrounding specific subjects, events, or issues [85]. This process aids in developing a comprehensive social explanation by shedding light on what misconceptions exist and what aspects of a topic are causing confusion or concern among the general public.

The emergence of large language models (LLMs) like ChatGPT has demonstrated substantial potential capabilities such as sentiment analysis and question-answering, driven by extensive training data and the integration of reinforcement learning from human feedback [86]. However, challenges remain, particularly in deciphering complex human communications like sarcasm and irony, which can lead to potential misinterpretations or irrelevant analyses [87]. Recent advancements have been made in this area, exemplified by the development of models integrating sentence-based embeddings and autoencoder techniques for sarcasm detection [88]. However, these models are resource-intensive and may not scale as easily to the vast amounts of data that LLMs can handle. Furthermore, generating accurate or relevant social explanations for misinformation regarding novel events that LLMs have not been trained on is another challenge [89]. Future advancements in this field should focus on developing adaptability and a deeper contextual understanding within NLP to increase the quality and relevance of the analysis applied to misinformation detection and its accompanying social explanations.

2) User Profile and Social Network Analysis: User profile and social network analysis are crucial technologies that enable a contextual understanding of the dissemination of misinformation. User profile analysis helps understand individual behaviors, preferences, and biases [36], which gives insights into the individual's unique context that might motivate them to spread misinformation. For instance, users frequently engaging with conspiracy theory websites might be more prone to sharing unverified claims due to their numerous exposure to such content. On a broader scale, social network analysis elucidates how individuals influence each other within a network of misinformation spread [90], thereby assisting in identifying the sources and routes of misinformation. By employing centrality measures in social network analysis, we can identify key influencers or the primary nodes within a network by analyzing the positions, connections, or roles of each node in the flow of information within the network [91], [92]; thus, if an individual is frequently retweeted or shared within a group that spreads misinformation, they can be identified as a main propagator of such misinformation. Another method involves community detection, which groups individuals based on their interactions and the information they share [93], [94]. Analyzing these groups can help identify clusters of individuals that may be highly prone to or actively involved in disseminating misinformation. However, due to the rapidly evolving landscape of misinformation, where misinformation propagators persistently modify their strategies to bypass detection measures, it's crucial that user profile analysis and social network analysis remain up-to-date [95]. Moreover, the complexity and fluidity of social relationships present challenges for social network analysis, necessitating sophisticated data structures that accurately reflect complex social dynamics [96], as well as the development of dynamics models and algorithms capable of real-time adaptation to these dynamics [97], [98].

3) Visualization: Various visualization tools are reshaping our ability to understand misinformation in a more intuitive way. By combining sentiment analysis from NLP with visualization methods like heat maps and word clouds, individuals can more effectively grasp emotions and sentiments within textual content [99]. Tools such as D3.js portray data as nodes and edges, offering a comprehensive understanding of how misinformation spreads through online communities [100], [101]. TimeLineCurator charts the chronological trajectory of misinformation narratives, aiding in the identification of misinformation origins and its progressive evolution [102]. PowerBI's capability to integrate with diverse data types can help illuminate the societal forces driving misinformation [103]. For example, by combining geographical indicators with regional political affiliations, PowerBI can pinpoint areas prone to misinformation, revealing factors like political orientations as potential contributors to misinformation spread.

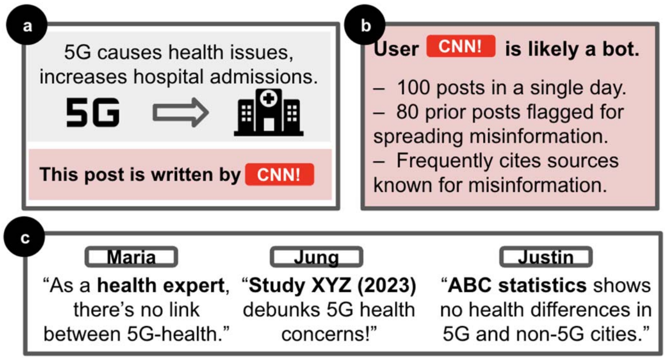
Advancements in social computing and HCI research, including adaptive personalization algorithms, are tailoring visual tools to better match user preferences and needs [104]. These algorithms, which have shown significant progress, include traditional constraint-based optimization, where interfaces are designed according to specific layout rules, and data-driven methods that leverage machine learning to analyze user behavior for interface generation [105]. Notably, in the context of explaining misinformation, implementing such adaptive algorithms into a system can be especially beneficial. For example, a system could analyze a user's past data to create a tailored dashboard that traces a user's encounters with misinformation over a certain period. This dashboard would explain how a user's social network interactions lead to exposure to misinformation, emphasizing the roles played by various contacts and the patterns of misinformation spread within their social circles. Nonetheless, challenges persist, such as accommodating diverse cognitive processes, developing thorough multimodal systems, and upholding user confidentiality [105]. As efforts continue to tackle these obstacles, integrating data visualization with social computing and HCI is poised to offer more human-centered insights into social explanations.

IV. Real-World Applications

Moving from theoretical foundation of social explanations to their practical applications, the integration of social explanations into XAI reveals vast potential in domains such as robust journalism and fact-checking, critical thinking and media literacy in educational programs, timely interventions in health emergency management, and increasing transparency in social media algorithms.

A. *Leveraging Social Explanations for Robust Journalism and Fact-Checking*

Social explanations offer benefits not only to individual social media users but also to online news outlets engaged in both journalism and fact-checking. They can use social explanations to support their journalistic investigations and verification of facts. A notable example of this is the 2012 incident where renowned media outlets, CNN and Fox News initially reported that the Supreme Court had struck down the Affordable Care Act's individual mandate, when in reality, the court had upheld the mandate as a tax.[1] Given the credibility of these outlets, this misinformation was propagated quickly and widely online. While traditional journalistic practices such as internal fact-checking and peer review have allowed misinformation to be corrected within minutes to an hour, one complementary method of rapid verification would be the strategic use of user credibility metrics. By filtering real-time feedback based on highly credible users (e.g., legal experts or users who have a proven track record of accurate fact-checking), outlets can potentially identify errors more quickly. Similarly, analyzing the patterns of information dissemination can aid fact-checkers in pinpointing and prioritizing potentially misleading content that reaches a wide audience or spreads rapidly. This integration of social explanations with journalistic practices and fact-verification processes empowers news portals and fact-checkers to strengthen their reporting and combat misinformation more efficiently, ultimately fostering a more trustworthy information landscape.



5G causes health issues, increases hospital admissions.

5G

This post is written by CNN!

User CNN! is likely a bot.

– 100 posts in a single day.
– 80 prior posts flagged for spreading misinformation.
– Frequently cites sources known for misinformation.

Maria
"As a **health expert**, there's no link between 5G-health."

Jung
"**Study XYZ (2023)** debunks 5G health concerns!"

Justin
"**ABC statistics** shows no health differences in 5G and non-5G cities."

Fig. 5. Social explanation on online news and fact-checking. (a) Controversial post. (b) Analysis of the post's author "CNN!" (not the official account of "CNN," but possibly a mimic or fake account). (c) Comments from highly credible level users, "Maria," "Jung," and "Justin."

Fig. 5 shows an example of an online news outlet conducting a comprehensive investigation into a controversial social media post concerning the impact of 5G technology on health [Fig. 5(a)]. The post in question is authored by CNN!, a username that is similar but not affiliated with the official CNN news network. To address the potential impact of this post from CNN!, including the generation of groundless health concerns within the public and erosion of trust in technological advancements, the outlet deploys an AI system with advanced user attribute analysis capabilities for a rigorous examination of the post's veracity. The AI scrutinizes not only the content but also the credibility level of the post author, CNN!, by evaluating factors such as the user's past engagement history (e.g., excessive daily posting activity) [Fig. 5(b)]. Following this, the AI system examines alternative sources cited in comments from users with high-credibility levels, such as expert opinions, scientific study, and statistics [Fig. 5(c)]. Through this thorough process, the AI conclusively confirms the misleading nature of the post, enabling the online news outlet to proactively take measures to prevent the further spread of the misinformation. Moreover, the AI system can go beyond mere misinformation detection by actively engaging with the audience in generating posts that counteract misinformation. For example, the system could develop a feature that auto-generates articles highlighting comments made by users with high-credibility levels. With this integration of user comments and credibility assessments, news consumers are not merely passive recipients of information but active contributors in the fight against misinformation.

B. *Empowering Critical Thinking and Media Literacy in Educational Programs*

The application of social explanations in XAI also extends to educational settings. By providing users with profound insights

[1] https://www.washingtonpost.com/blogs/erik-wemple/post/cnn-correction-on-health-care-ruling-insane/2012/06/28/gJQAg6w78V_blog.html

into the social contexts of misinformation, this approach empowers individuals with the necessary skills and knowledge to assess the information they encounter online more effectively. This application area is especially critical for the younger generation who often source their information from social media platforms,[2] where distinguishing between credible and misleading news becomes challenging. An example of this is the pervasive vaping myth circulating among teenagers that vaping is a safe alternative to smoking. This misconception is often fueled by peer pressure [106], where friends frequently share and endorse vaping content on social media platforms. Younger individuals, who are more susceptible to these social influences [107], [108], might confuse the popularity of vaping with its safety. To counteract this susceptibility to misinformation, it is crucial to foster a critical mindset in teenagers, enabling them to discern the credibility of the information they encounter online. As a result, developing educational programs that integrate social explanations, such as an explanation of user engagement like an analysis of how social influence affects misinformation spread, becomes vital to enhance media literacy and critical thinking for a more resilient and informed society. [109].

Let us imagine a scenario where a high school implements an XAI system with social explanation features as an integral part of its educational program. This XAI system incorporates two complementary components: misleading sociocontextual cues identification training [Fig. 6(a)] and immersive gameplay [Fig. 6(b) and 6(c)]. At first, students participate in interactive Q&A training sessions guided by the system, where they are encouraged to analyze and pinpoint misleading sociocontextual cues present in built-in data, such as misinformation instances found in school event articles and social platforms [Fig. 6(a)]. Subsequently, the system provides answers, along with detailed social explanations, to aid students in more effectively identifying and comprehending sociocontextual cues. After this foundational training, students engage in an interactive game, “Misinformation Detective,” where they are provided with real-world misinformation instances [Fig. 6(b)] and act as detectives who are tasked with finding and deciphering misinformation within scenarios simulated from real-world using vetted data for education purpose. Throughout the game, students encounter and face off against fictional “monsters” notorious for spreading misinformation, scoring points each time they identify misleading sociocontextual cues tied to a monster. For example, there could be one game activity within this game named “Finding motivation phase,” where students are challenged to discern the monster’s intent using engagement metrics. Students can develop their own social explanations about the underlying motives of misinformation monsters, whether those motives revolve around gaining attention and popularity, securing financial gains, or merely instigating chaos through the spread of misinformation. Further, to validate their social explanations, students have the opportunity to participate in virtual debates where they can present and discuss their social explanations [Fig. 6(c)]. Through the debate, students develop critical thinking skills and better understand the misinformation dynamics. Finally, this immersive game concludes with a detailed debrief, a structured session where students revisit and analyze instances of misinformation they encountered during the game. This session ensures that students do not believe in the misinformation from their in-game experience, as each misinformation instance is clearly labeled as false and the corresponding accurate information is presented alongside for clarification. LLMs such as ChatGPT, which generate responses based on historical data patterns, struggle to create contextually relevant questions or understand real-time context [110], partially due to lack of transparency surrounding the dataset sources used to train these models [111]. On the other hand, this XAI system implemented for educational purpose is designed to ask timely and pertinent questions, accompanied by comprehensive social explanations of misinformation. By providing students with timely queries and in-depth social explanations, this targeted approach not only enhances student engagement but also provides students with a deeper understanding of how to detect misinformation using the subtle sociocontextual cues behind deceptive content.

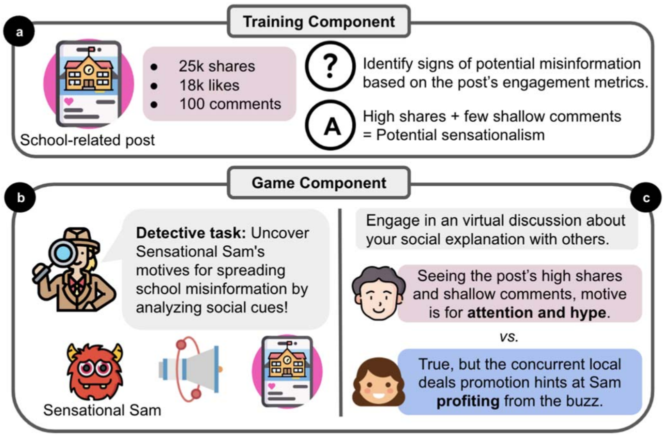

Training Component
a
25k shares
18k likes
100 comments
School-related post
?
Identify signs of potential misinformation based on the post's engagement metrics.
A
High shares + few shallow comments = Potential sensationalism
Game Component
b
Detective task: Uncover Sensational Sam's motives for spreading school misinformation by analyzing social cues!
Sensational Sam
c
Engage in an virtual discussion about your social explanation with others.
Seeing the post's high shares and shallow comments, motive is for attention and hype.
vs.
True, but the concurrent local deals promotion hints at Sam profiting from the buzz.

Fig. 6. Social explanation on educational programs. Here, (a) user engagement analysis is used to train students in identifying misinformation and (b) and (c) uncover the motive behind it.

C. Timely Interventions via Social Explanations in Health Emergency Situations

In times of crisis, like the COVID-19 pandemic, misinformation can yield immediate and severe real-world consequences [112]. An example of this is the spread of the mask-wearing myth that masks were ineffective or even harmful [113]. The dissemination of such misinformation not only erodes public trust in health authorities but also fosters confusion, preventing people from following the preventive measures. For example, 2,400 incidents of passengers resisting face masks were reported by the Federal Aviation Administration,[3] potentially aggravating the transmission of the pandemic. Therefore, it is important for health organizations to act quickly

[2] https://www2.deloitte.com/se/sv/pages/technology-media-and-telecommunications/topics/digital-consumer-trends/are-younger-generations-moving-away-from-traditional-news-sources.html

[3] https://www.cnbc.com/2021/07/06/disputes-over-mask-mandates-comprise-75percent-of-faas-unruly-passenger-complaints-on-planes-.html

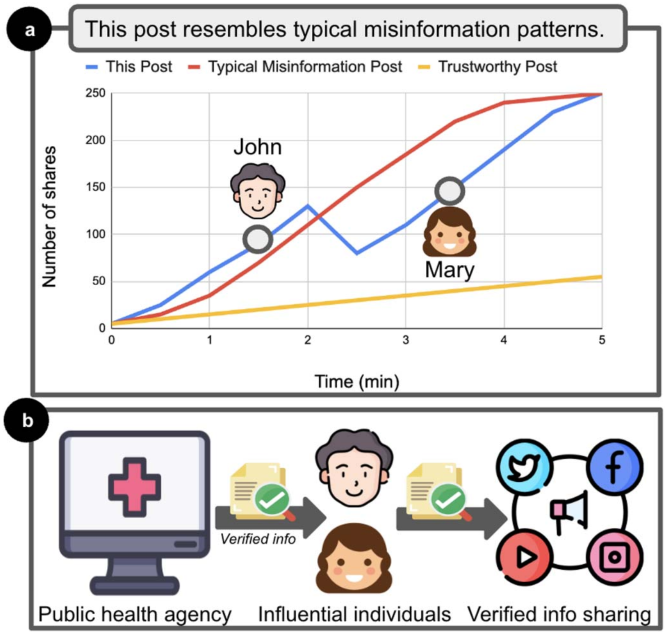


Fig. 7. Social explanation on public health and crisis management. (a) Diffusion pattern graph highlighting misinformation spread and influential nodes. (b) Public health agency's direct outreach to influential nodes for sharing verified information.

to communicate accurate information, especially during a crisis like the virus outbreak, when the impact of misinformation is significant.

In the face of such a situation where misinformation is spreading rapidly and causing public confusion, the public health agency can harness the power of social explanations for timely intervention, as illustrated in Fig. 7. Using a sophisticated AI system enhanced with social explanation capabilities, the agency rapidly examines the information spread on the diffusion graph in real time. Within just 5 min of the information's dissemination, the agency spots that the post's diffusion pattern resembles that of typical misinformation [29] and displays a marked difference from patterns of the trustworthy post [Fig. 7(a)]. To identify individuals as influential nodes during the rapid information dissemination, the system analyzes the diffusion pattern graph. Notably, the agency observes a distinct spike in the graph, indicating a surge in shares originating from influential nodes [e.g., John and Mary in Fig. 7(a)]. By calculating betweenness and eigenvector centrality metrics [114], the agency further confirms the significant influence of these influential nodes within their social network. Armed with these timely insights, the agency takes multifaceted approaches: they engage directly with individuals identified as influential nodes of misinformation dissemination, providing the individuals with verified information and encouraging the sharing of this correct information through various social media platforms [Fig. 7(b)]. This strategic initiative ensures that corrective messages intersect precisely with the audiences initially exposed to the misinformation, thereby enhancing the chances of correcting misconceptions swiftly. Simultaneously, for heightened credibility and wider reach, the agency collaborates with "gold" level users, leveraging their expertise and trustworthiness to amplify the dissemination of the counter-message across various social media platforms. Through coordinated efforts, a rapid response mechanism is established, effectively mitigating the adverse effects of misinformation during crucial moments.

D. Enhancing Transparency in Social Media Algorithms

Finally, utilizing social explanation particularly on social media platforms has the potential to improve the transparency of social media algorithms. At present, these platforms are not fully transparent in elucidating why specific posts are identified as misinformation [115]. Rather than providing detailed explanations, they either provide generic statements like "checked by an independent fact-checker" or "learn what experts say," leaving users uncertain about the specific reasons behind the misinformation detection results. An example of this is a controversy surrounding Twitter's "shadowbanning," a content moderation practice that allegedly reduces the visibility of certain users' tweets and comments without notifying them of such actions.[4] This shadowbanning practice came under scrutiny as individuals began to notice and showed concerns about their content seemingly being suppressed, suspecting they were being shadowbanned. Later on, Twitter attempted to justify the practice by stating that they are using algorithms to promote healthy conversations, which may have reduced the visibility of content considered harmful. Despite this explanation, many users found the initial lack of transparency regarding the content visibility concern, emphasizing a need for enhancing transparency in social media algorithms to maintain user trust in content moderation processes.

However, incorporating social explanations within the social media platforms could address the challenges associated with the current lack of transparency in their algorithms. In the scenario depicted in Fig. 8, when people encounter misinformation on their social media feed, they are given the option to click on the "Challenged post! Click to see social explanations" [Fig. 8(a)]. By clicking on this option, the user can access comprehensive and detailed explanations of why the content has been flagged as misinformation. Each social explanation provided offers insights into how the AI system assesses and identifies misinformation. For instance, when the user selects the "Expert Opinion" category within the theme visualization of user comments [dotted line box in Fig. 8(b)], they can discover that the content was flagged because it received a substantial number of comments from domain experts debunking the content [Fig. 8(c)]. Additionally, the user can delve into another social explanation, such as the diffusion pattern graph, to observe how the post of interest mirrors the dissemination pattern of misinformation in comparison to credible posts. This process enables the user to develop a deeper understanding of the reasons behind the flagging of content as misinformation through exploring these social explanations, ultimately fostering trust in the AI system.

[4] https://www.theatlantic.com/technology/archive/2023/01/twitter-shadow-ban-transparency-algorithm-suppression/672736/

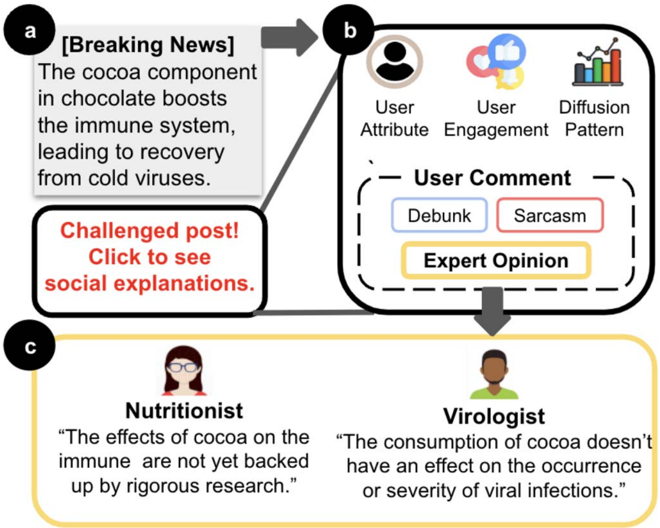

a
[Breaking News]
The cocoa component in chocolate boosts the immune system, leading to recovery from cold viruses.
Challenged post! Click to see social explanations.
b
User Attribute
User Engagement
Diffusion Pattern
User Comment
Debunk
Sarcasm
Expert Opinion
c
Nutritionist
"The effects of cocoa on the immune are not yet backed up by rigorous research."
Virologist
"The consumption of cocoa doesn't have an effect on the occurrence or severity of viral infections."

Fig. 8. Social explanation on social media platforms. (a) Alert notifies the user of a flagged post, prompting exploration of social explanations. (b) User selects "Expert Opinion" from the thematic visualization of comments among four types of social explanations. (c) Comments from domain experts refuting the misleading post are displayed.

V. Research Challenges and Opportunities

This section discusses the challenges and opportunities related to integrating social explanations of misinformation into XAI. Examples of the key challenges include data reliability, the dynamic and nonbinary characteristics of misinformation, effective integration of social explanations into XAI frameworks, user interface design that maintains privacy, and establishing robust evaluation for social explanations. Also, we examine the practical challenges that social media platforms face in effectively managing misinformation using sociocontextual cues, given that these platforms are often recognized as primary proliferators of misinformation [116].

A. Data Reliability Challenge

The process of gathering data from unverified crowds to provide social explanations presents challenges due to the potential noise and unreliability of the collected data. For example, irrelevant discussions and personal anecdotes shared by users in comments can lead to inaccurate social explanations and pose challenges in identifying misinformation [117]. Consider a scenario where a deep fake video about a celebrity admitting to illegal activities circulates on social media platforms. In such a case, users might discuss unrelated controversies or past scandals involving the celebrity, backing these claims as evidence for why this deep fake video is convincing. These discussions could be erroneously regarded as contextual cues to verify the authenticity of the post, leading to the generation of a wrong social explanation, such as "Based on prevailing comments surrounding the celebrity's history of controversies, the video content is likely to be authentic."

This issue of low data quality collected from various crowds, referred to as data reliability in crowdsourcing and HCI literature [118], has prompted the development of solutions to filter out unreliable data. For instance, quality control techniques such as prescreening participants and implementing quality control questions have been proposed [119]. Additionally, ML algorithms for identifying patterns of low-quality responses have been utilized [118]. However, these solutions are impractical for generating comprehensive social explanations, especially in emerging or rapidly evolving fields that lack sufficient sociocontextual cues. In the case of a deepfake video example, when this technology is new to the public, differentiating between authentic content from manipulated can be challenging for most people [120]. This unfamiliarity with the technology is likely to produce a variety of comments that contain misunderstandings and speculations, which are considered "noise" by conventional data filtering methods. In such a situation, an advanced machine-learning algorithm might filter out most comments, leaving only those from verified accounts or that cite credible sources [121]. While these filtered comments may be high quality, their limited number of comments can result in a social explanation that might be overly narrow, focusing solely on manipulation techniques used in the deep fake video. That is, by filtering out low-quality comments that reflect the public's sentiments, the broader social dynamics surrounding the post (e.g., the public believes the video simply because it confirms their preexisting biases against the celebrity) could not be captured.

Additionally, truth discovery techniques have been developed to enhance data reliability, evaluating both the authenticity of input data and the credibility of online users. Addressing the challenges of data reliability when utilizing human-generated data, estimation theory has been applied to identify potential biases and inconsistencies inherent in human-provided data [122]. Similarly, truth discovery techniques in social media sensing were refined, with a focus on validating data authenticity and the assessment of user credibility [123]. Despite these truth discovery techniques effectively evaluating data truthfulness and origin trustworthiness, they often face challenges when it comes to capturing the nuanced subtleties of human interactions due to limited context-awareness [124]. For instance, consider a comment like "Sure, chocolate's a vegetable 🙄." The eye-roll emoji serves as a sarcastic cue in this statement, indicating the author's intent to jestingly debunk or mock misinformation. However, conventional truth discovery techniques might overlook this nuance, interpreting the statement at face value. This misinterpretation can lead to the incorrect conclusion that the statement is misinformation, thereby leading to a wrong social explanation that the author endorses in the inaccurate dietary classifications, probably due to the author's limited knowledge. Therefore, it remains an open challenge to develop practical and context-aware approaches to ensure that reliable data are acquired for generating social explanations. One promising research direction would be the integration of human-in-the-loop systems [125], such as utilizing machine-learning algorithms for initial data prefiltering, followed by human verification to ensure a nuanced understanding and contextual relevance. Research in this area could focus on finding ways to enhance the synergy between automated and human insight, aiming to develop a more effective and robust data collection framework.

B. Evolving Patterns in Misinformation and Social Explanation

Misinformation, by its nature, is not static; rather, it evolves over time, context, and through user engagement [126]. Information may be considered true at a specific moment but can transform into misinformation through paraphrasing or shifts in context. For instance, a valid scientific finding may be distorted or oversimplified on social media, resulting in misconceptions [127]. The narrative regarding Gordie Howe's stem cell treatments exemplifies how information can mutate on social media [128]. Specifically, in December 2014, renowned ice hockey player Gordie Howe underwent an unapproved stem cell treatment in Mexico following a severe stroke. While the story began with the factual event of Howe's treatment in Mexico, it soon morphed into a narrative of a near-miraculous stroke recovery on Twitter. Meanwhile, much of the crucial scientific information, including the treatment's experimental nature and its absence of FDA approval, was either simplified or completely left out of the conversation on social media platforms. Conversely, information initially labeled as misinformation might later be accepted as truth if thoroughly debunked or societal views change, as evidenced by the example of Galileo's heliocentrism. Initially, the Catholic Church rejected Galileo's heliocentrism, labeling it "heretical" as it contradicted the prevalent geocentric church doctrine. However, over time, growing scientific evidence and a societal shift toward empirical reasoning during the Scientific Revolution led to the eventual acceptance and validation of heliocentrism as a fundamental principle of astronomy.

The dynamic nature also applies to the social explanations that can be utilized to explain misinformation. Suppose a user disseminates content initially deemed accurate, and later shares information identified as misleading. In this case, the user's credibility level might diminish, leading other viewers or followers to experience confusion and mistrust upon noticing this decline in trustworthiness. Particularly, if the social explanation relied heavily on the user's high credibility level to validate the information, a subsequent drop in the user's credibility could lead other viewers or followers to distrust the social explanation they previously saw and ultimately to distrust the judgment of the XAI system. Beyond this, the inherent unpredictability of human behavior can also play a part in a user's fluctuating credibility level [129]. Someone who is seen as reliable today might display unreliable behavior tomorrow due to reasons ranging from personal life changes to evolving beliefs [130]. For instance, if someone loses their job in the medical field and begins to post dubious health advice stemming from personal frustration or limited access to updated information, their credibility as a previously trusted source on health topics could suffer a significant decline. Furthermore, user engagement metrics such as the number of likes and shares, influenced by factors such as platform algorithms [131], keep changing, highlighting how a user's influence and perceived trustworthiness can vary over time. This dynamic interplay between misinformation and social explanations poses a critical question: how can computational methodologies be tailored to create systems that can adapt to these dynamic shifts?

The varied human factors shaping interpretations of social explanations present an additional challenge. For example, an individual with confirmation bias might easily accept a social explanation that aligns with their existing beliefs, even when such an explanation is poorly supported by evidence [132]. Cultural background also influences the sources individuals consider trustworthy. In some cultures, social explanations that echo collective experiences might be prioritized over those from domain experts, regardless of the information's validity or neutrality [133]. Given the intricate nature of how different users interpret social explanations of misinformation, another vital question emerges: How can we design these systems to meet diverse user needs arising from varied cognitive processes and cultural backgrounds? Also, considering these cognitive and cultural challenges, how can we ensure that the system remains user-friendly for all users, offering a seamless, intuitive, and responsive interface? Addressing these questions demands a deep understanding of the quickly transforming misinformation landscape and the human factors influencing how people interpret and respond to social explanations.

To address the evolving nature of misinformation effectively, developing adaptive methodologies within XAI systems is critical. One such approach involves enhancing the capabilities of generative pretrained transformers (GPT) for not only detecting misinformation but also for generating real-time, accurate explanations for the detection. Recent advancements in GPT technology have shown promise in real-time misinformation detection through its robust ability to recall factual knowledge without the need for fine-tuning [134]. Specifically, this capability allows GPT to quickly compare new information against a vast database of established facts, highlighting its potential to contribute to the development of an automated real-time misinformation detection model. However, the exploration of GPT's potential in generating explanations of misinformation remains relatively understudied. This research gap is significant given the current limitations of GPT, such as tendencies toward hallucinations and producing unreliable outputs [135]. In the rapidly changing misinformation landscape, these inaccuracies and unreliability of explanations generated by GPT models are further amplified, as misinformation tactics are becoming increasingly subtle and complex [14]. Therefore, further research focused on improving the accuracy of GPT's explanations is needed to develop reliable real-time explanatory outputs.

Another important research avenue for developing adaptive XAI systems involves integrating continual learning with user feedback loops. Continual learning enables AI systems to update and refine their knowledge bases continuously, a feature that is particularly beneficial in the development of adaptive XAI systems, where the ability to integrate new information while retaining existing knowledge is crucial [136]. However, the challenge of the integration of continual learning with XAI systems lies in ensuring that the integration of new data does not deteriorate the quality of explanations provided by XAI systems [136]. For example, the incorporation of new data can potentially impact the consistency and clarity of explanations negatively, as continually integrating new information might complicate the AI's ability to provide clear and stable explanations over time. One solution to mitigate this potential degradation of explanation is to incorporate a user feedback mechanism into the XAI system [137]. Users could highlight specific aspects of explanations where the system's explanations may not be keeping pace with the evolving nature of misinformation through their feedback, which can then be used to adjust and enhance the AI's learning algorithms and explanatory methods in real time [137]. However, this approach introduces two critical research questions for future exploration: first, the timely acquisition, analysis, and integration of user feedback—how can we develop methods for immediate feedback processing that can keep pace with the rapidly changing dynamics of misinformation? Second, another question to explore is: how can XAI systems effectively use limited initial user feedback without delaying necessary updates in fast-evolving misinformation landscape? This challenge, known as "cold start" problem [138], involves the difficulty of collecting sufficient feedback quickly enough to ensure the feedback's usefulness and confidence for the XAI system. For instance, with few user feedback at the start of an event, the system may face a delay in gathering enough reliable and accurate feedback. Such a delay, while aimed at validating and enhancing the quality of the input, risks rendering the feedback outdated or less relevant as the context of misinformation quickly evolves. Therefore, future research efforts could focus on developing methodologies that enable XAI systems to circumvent the "cold start" problem. This could involve developing algorithms that can assess the credibility of the initial feedback, or implementing confidence scoring mechanisms that allow for immediate yet reliable adaptation based on limited feedback (e.g., by assigning confidence levels to new information through comparison with existing data using pattern recognition techniques [139]).

C. Nonbinary Nature of Misinformation

The complexity of social explanation increases with the nonbinary nature of misinformation, which often includes statements that are partially true or false [22]. For instance, consider the claim that "While COVID-19 vaccines are effective in preventing infection and transmission of the virus, those who already had COVID-19 do not require vaccination." This statement is partially true because the COVID-19 vaccines have been shown to significantly reduce the risk of transmission [140] but also partially false because individuals who have had COVID-19 but remain unvaccinated face a risk of reinfection more than double that of those who were infected and subsequently vaccinated [141].

Alternatively, there can be the case where the claim itself is factually correct but can lead to misleading implications. Consider a social media post that presents statistics showing a significant increase in gun violence rates within a city over the previous year. This seemingly accurate information can mislead users by omitting a crucial detail: the apparent increase is primarily due to changes in crime reporting. For example, the shift from the FBI's summary reporting system (SRS) to the national incident-based reporting system (NIBRS) has improved gun violence reporting through comprehensive data collection [142]. While this statistic may suggest a rise in reported gun violence owing to the change in the reporting method, it does not necessarily correspond to an actual increase in gun violence occurrences. These complexities highlight the importance of explanations that delve into the underlying motivations and broader contexts beyond mere content of misinformation. Such gray areas require nuanced social explanations that go beyond validating the statistics and probe the broader context and motives behind sharing such information. In the aforementioned gun violence rate scenario, for example, the social explanations can reveal insights from domain experts about the real reasons behind the data, as well as the sources of information spread that may aim to influence public demands for stricter law enforcement or political support for specific policies. However, developing effective social explanations for this nonbinary nature of misinformation poses several challenging questions for researchers. For instance, what methodologies can integrate sociocontextual cues, human behavior, and psychological factors into understanding nonbinary misinformation? How can these nuanced insights be communicated to users without causing confusion or mistrust, given the complex relationship between truth and falsehood in nonbinary claims?

These questions call for a multidisciplinary approach toward achieving a comprehensive understanding of nonbinary nature of misinformation. Recent studies have shown how insights from behavioral economics, such as the limited-attention utility model [143] and the concepts of cognitive effort and appeal to emotions [144], can inform the design of algorithms to detect misinformation. To further enhance these algorithms, it is imperative to shift our focus toward the nonbinary aspects of misinformation. For example, incorporating research on decision-making process under uncertainty can be beneficial for designing algorithms that can effectively convey the complexities of nonbinary misinformation [145]. Another promising approach for addressing nonbinary misinformation complexities involves leveraging crowdsourcing techniques, which harnesses the collective judgment and diverse interpretations of a broad audience to grasp the nuanced aspects of misinformation [146]. Recent study has shown the success of this approach in identifying both binary and nonbinary misinformation, including partially misleading or unverifiable content [146]. Given this promising result, future research should further explore the capabilities and methodologies of crowdsourcing to improve nonbinary misinformation detection. For example, developing advanced participant profiling systems that provide in-depth, ongoing assessments of contributors' specific skills and knowledge is crucial, especially since nonbinary misinformation frequently requires domain-specific expertise to discern misleading or partially false content.

D. Integrating Social Explanations With an XAI Framework

Social explanations complement a traditional XAI framework that centers on content-based explanations. However, effectively integrating them is a complex task. One challenge arises from the different granularity of data inherent in content-based and social explanations. Content-based explanations often analyze granular elements like specific keywords or image pixels to understand the unique characteristics and attributes underlying misinformation [18], [19]. In contrast, social explanations present a varied granularity: at the micro level, they encompass individual user interactions, like comments or likes, which can reveal personal misconceptions or individual's reactions to false narratives in community dialogues. On a macro scale, social explanations analyze patterns that emerge from clusters of users within the broad social network, capturing collective sentiment or shared perspectives of a community [147], [148]. Given such varying levels of details between the two types of explanation, a question arises: how can we seamlessly integrate these varying levels of details into a unified and coherent XAI framework?

Another challenge arises when the social explanation and the content-based explanation conflict with each other. For instance, what if a post is deemed to be misinformative, but content-based and social-based explanations offer different reasons for this judgment? Imagine a post stating new health benefits of a specific food. Here, a content-based explanation can reveal the use of accurate scientific terminology, devoid of the sensationalism or exaggeration often seen in misinformation. In contrast, the social explanation can highlight that the information's dissemination on social networks parallels that of earlier debunked health myths. In this case, should both explanations be presented to the user, allowing them to make their final decision, or should the system determine the primary explanation to present to users with an option for users to access a justification of the system's decision? These novel questions present interesting research avenues for future work. For example, researchers could work on developing mechanisms within XAI systems that are adept at reconciling differences between content-based and social explanations. One potential method would be developing decision-making algorithms that not only evaluate the credibility and relevance of different explanations but also integrate these explanations to deliver a more unified and comprehensive understanding to users [149]. Additionally, the need for XAI systems to operate effectively across varying levels of data granularity calls for the design of frameworks capable of fluidly transitioning between detailed analysis of individual interactions and broader examination of social patterns. As empirical studies in XAI suggests that users would like to receive explanations at various levels and stages, depending on what they know and would like to know [150], research on conflict resolution and data granularity adaptation promise to not only facilitate the seamless incorporation of social explanations into XAI frameworks but also advance the XAI field toward more nuanced and user-centric explanations.

E. Privacy-Preserving User Interfaces

Designing a user interface that effectively communicates the utilization of user data without compromising privacy presents a significant challenge. Misinformation detection models often rely on various user data categories, such as geographical location and political affiliation, to assess the credibility of information [36], [151]. However, incorporating such details into explanations to enhance the contextual understanding of misinformation risks exposing sensitive personal data or perpetuating stereotypes [152]. For example, consider a scenario where the social explanation demonstrates a higher prevalence of misinformation in areas characterized by significant political polarization. Combining this information with other publicly available data, like social media activity, might enable the identification of individual users more susceptible to misinformation within those areas. Specifically, by analyzing the social media engagement pattern and online news consumption (e.g., types of news sources individuals follow) among residents of these identified regions, a more detailed understanding of which individuals are more likely to be exposed to and interact with misinformation can be gained. Such identification not only poses a privacy risk by exposing personal beliefs to the public but also opens the door for malicious actors to create deceptive campaigns tailored to the unique beliefs and tendencies of the misinformation-vulnerable individuals.

Also, in a case when an AI system does not explicitly disclose private information, it can indirectly imply user attributes. For example, if the AI system identifies misinformation based on multiple reports from highly credible users, it implies that those users are considered more trustworthy by the AI. This, in turn, may lead users to form conclusions about their own or others' credibility based on the behavior of the AI. Such indirect implications can be a breach of privacy because, while the AI may not overtly label users as "trustworthy" or "untrustworthy," it's actions imply the system inadvertently creates a hierarchy of credibility among its users. Potential solutions may involve removing sensitive sociocontextual cues in the explanation process or providing more abstract explanations to prevent the explicit or inferred identification of personal user characteristics or behaviors. However, defining what is considered as a sensitive identifier can be subjective; factors such as user preference [153] and cultural affiliations [154] vary among users, posing a challenge in establishing universally accepted standards. While generalized explanations can prevent exploiting or manipulating the AI system, they may not meet a user's demand for transparency or understanding of the AI system's actions. Given these challenges, it is crucial to investigate how AI systems can effectively present social explanations while maintaining the privacy of users. For example, researchers could focus on developing algorithms that can adeptly abstract sociocontextual cues, ensuring social explanations are informative yet privacy-preserving. Additionally, investigating user-specific preferences and cultural nuances in privacy perception will be crucial in tailoring social explanations to diverse user groups [155]. For example, this could involve conducting cross-cultural studies to understand how different societies perceive misinformation and privacy, and using these insights to develop adaptive algorithms that can modify their approaches based on the user's cultural background identified (e.g., personalized message framing).

F. Evaluation Challenge

Rather than merely explaining the reasons behind misinformation, social explanations aim to guide individuals in better identifying and combating misinformation. Therefore, the assessment of social explanations requires an understanding of their impact on users' beliefs and behaviors. However, various social variables influence human beliefs, including peer pressure and social norms [156], adding complexity to isolating the effectiveness of social explanations in users' belief change. In addition, behavioral changes from social explanations can be subtle and evolve gradually, much like the unconscious influences of misinformation [14], complicating efforts to gauge the impact of social explanations accurately. Furthermore, the user experience is inherently subjective; while some explanations might resonate with specific demographics, cultures, or situational contexts, others might not [157]. For example, a social explanation that emphasizes community values might strongly resonate with one cultural group but may be perceived as irrelevant or even offensive by another culture that prioritizes individual autonomy.

Acknowledging the inherent challenges in quantitatively evaluating social explanations, we conducted a preliminary experiment as an initial step to measure the explanations' impact on users' misinformation detection accuracy. Specifically, we designed social explanations in natural language format using the generative pretrained transformer 3.5 (GPT-3.5). Our focus was on elaborating user attributes and information diffusion contexts with claims sourced from the LIAR dataset—a renowned political misinformation detection dataset [158]. For example, a claim from the LIAR dataset says "Barack Obama's plan calls for mandates and fines for small businesses (Information diffusion context: Presidential debate in Nashville, Tenn; Speaker: John McCain)." The corresponding social explanation is as follows: "The context the claim was made, a political debate, is known for its strategic rhetoric and partisan viewpoints, which can sometimes lead to exaggerated or distorted statements. Also, McCain's varied credibility history (as provided in the dataset) suggests a tendency toward statements that are not consistently factual, further casting doubt on the veracity of his claim." All social explanations used in our experiment were manually fact-checked to ensure their correctness. Using such a set of social explanations, we recruited 20 participants via Amazon Mechanical Turk (MTurk) to evaluate the effectiveness of these explanations in detecting misinformation. Informed consent for our experiment was obtained by following the corresponding Institutional Review Board (IRB) protocol. We compared users' accuracy in identifying false information with and without the use of social explanations. The results showed that detection accuracy increased to 63.33% with social explanations, compared to 39.74% without, highlighting the effectiveness of social explanations in misinformation detection. Moving forward, we aim to broaden our research by exploring more diverse explanation designs beyond natural language, incorporating various types of social explanations as outlined in Section III, and extending our studies to nonpolitical settings.

Despite our initial finding that shows the potential of social explanations in misinformation detection, the complexity of social explanations raises an important research question: How can we establish comprehensive metrics and criteria to accurately assess the impact of social explanations on user experience and their effectiveness in debunking misinformation across varied user perspectives? For example, how can we develop metrics that accurately capture both the immediate and long-term effects of social explanations on changing user beliefs and behaviors concerning misinformation? Also, how can we ensure our evaluation method accurately captures the nuances of diverse user groups (e.g., developing metrics to assess how social explanations impact users of varied ages, educations, and cultures, for a comprehensive evaluation)? These questions call for a collaborative research effort aimed at developing sophisticated, multidimensional frameworks for evaluating the effectiveness of social explanations in debunking misinformation. For example, researchers can consider leveraging methods in behavioral psychology (e.g., controlled experiments and longitudinal studies) and computational modeling techniques (e.g., machine-learning algorithms and statistical network analysis). These techniques can be used to develop metrics that not only measure immediate user responses to social explanations, such as engagement patterns and sentiment analysis using machine-learning algorithms, but also track changes in belief systems over time using network analysis.

G. Practical Challenges Faced by Social Platforms in Misinformation Management

Finally, we explore the practical challenges that social media platforms face in effectively managing misinformation using socio-contextual cues. Social platforms, despite the availability of various socio-contextual cues, face some challenges in mitigating misinformation's spread [116]. Among various factors contributing to this challenge is the "attention economy" inherent within these platforms [159]. The primary business model of many social platforms is built on user engagement metrics such as clicks, likes, shares, and the followers [159], [160]. This model often leads to a situation where sensational or controversial content, typically associated with misinformation, tends to receive higher user engagement [42]. Such a scenario presents a complex dilemma for platforms: balancing the need to keep users engaged (and thereby maintaining profitability) against the moral and social responsibility to prevent the spread of misinformation. Adding to this challenge is the opaque nature of the algorithms curating user content feeds [161]. These algorithms, focusing primarily on maximizing engagement, can inadvertently amplify misinformation by creating echo chambers where users are continuously exposed to and reinforced in their existing beliefs or misconceptions [162]. Given these challenges, it becomes clear that social media platforms should do a comprehensive reevaluation to the core mechanisms that govern content distribution and user interactions on these platforms. This reevaluation involves finding a delicate balance between user engagement, factual accuracy, and ethical responsibility, ensuring that the pursuit of profit does not override the imperative of delivering truthful and reliable information.

In addition to these inherent challenges, social platforms face difficulties in implementing and integrating sociocontextual cues effectively to combat misinformation. Implementation challenge, for example, revolves around the feasibility of tracking, storing, and processing the vast amount of sociocontextual cues within a unified system. Addressing this challenge requires the deployment of cutting-edge technologies capable of managing the substantial data generated by user interactions—ranging from individual attributes to complex network behaviors (see Section III-C for more details). Moreover, strict adherence to data privacy and legal standards, such as those outlined in the General Data Protection Regulation (GDPR), is imperative [163]. Compliance with such regulations requires the utilization of privacy-preserving data mining techniques, striking a delicate balance between anonymizing user data for privacy and preserving its analytical utility [163]. Thus, ongoing research endeavors are indispensable to ensure robust privacy protection without compromising analytical depth and utility [164]. Furthermore, the challenge extends to seamlessly integrating these processed cues into the existing frameworks of content moderation and misinformation management on social platforms [165]. Achieving this integration mandates a robust technical infrastructure capable of accommodating diverse data types, coupled with a strategic approach to embedding these insights into moderation policies and practices that uphold ethical standards [166].

The integration challenge in social media arises from its diverse content formats (e.g., text, images, videos, and user interactions) [167]. While this article's focus has been on textual misinformation, we recognize that misinformation also thrives in image, video, and audio, each requiring specific analytical techniques for processing [167]. For example, as detailed in Section III, textual content is analyzed through NLP, user behavior through social network analysis, and visual content via CNN [168]. Despite the development of sophisticated methods to process each data type, the primary challenge extends beyond individual data processing to the fusion of these insights, aiming for a comprehensive understanding of misinformation. Recent advancements in this line of work include the development of Text-Image CNN (TI-CNN) models, which consider both textual and visual content to detect fake news, underscoring the importance of multimodal analysis in contemporary fake news detection efforts [168]. Yet, effectively integrating sociocontextual cues requires expanding our focus beyond text and visuals to encompass the analysis of post timings, patterns of network dissemination, and metrics of user engagement. Therefore, a more holistic approach that utilizes various analytical techniques, such as temporal analysis and diffusion pattern modeling, beneficial to enhance the detection and mitigation strategies against misinformation.

Finally, the proliferation of misinformation remains a formidable challenge for social media platforms, as rapid propagation undermines content moderation and fact-checking endeavors [70]. This swift dissemination of misinformation can distort public discourse before any corrective measures are possible [70]. Furthermore, psychological barriers like confirmation bias and cognitive dissonance make it difficult to correct misinformation, with individuals often reluctant to accept information that contradicts their preexisting beliefs [14]. In response to this challenge, there have been recent research on developing early detection systems for misinformation, employing techniques such as linguistic analysis, examination of user attributes and comments, engagement metrics, and analysis of diffusion patterns [169], [170], [171], [172]. These emerging misinformation detection models have shown remarkable efficacy, showing an accuracy rate of over 90% within just 5 min of news release [169]. However, despite these technological improvements, there is a critical need for greater transparency and better understanding of these algorithms among users. The current gap in user-centric design within misinformation detection models restricts the model's utility in combating the spread of emerging misinformation. For instance, illustrating the diffusion patterns of misinformation in a manner that echoes past deceptive campaigns could lead users to critically evaluate and potentially disregard dubious sources. Therefore, addressing emerging misinformation opens up various research paths, including identifying key sociocontextual cues for detecting emerging misinformation and differentiating between the approaches used to explain traditional misinformation versus those required for emerging misinformation.

VI. Conclusion

In this article, we envisioned integrating social explanations into XAI to combat misinformation. By leveraging sociocontextual cues such as user attributes, engagement metrics, diffusion patterns, and user comments, we provided a comprehensive overview of social explanations, highlighting both their unique benefits and the challenges they present. Drawing upon interdisciplinary techniques from fields like computer science, social computing, HCI, and psychology, our vision extended the current XAI literature by broadening its focus beyond content-based explanations. We highlighted the unique advantages of integrating social explanations into XAI, including enhanced user trust in AI systems and empowerment in combating misinformation. However, we also acknowledged several open research challenges associated with our vision, such as the dynamic and nonbinary characteristics of misinformation and the effective integration of social explanations into XAI frameworks. Additionally, we discussed the practical application of our vision in areas like education and journalism and address the specific challenges of realizing our vision on social media platforms, where misinformation is known to proliferate. Moving forward, we anticipate that this article will position our vision of integrating social explanation into XAI as a crucial pathway for future XAI research, particularly in the current era of combating online misinformation.

Acknowledgment

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] K. Kaur and S. Gupta, "Towards dissemination, detection and combating misinformation on social media: A literature review," *J. Bus. Ind. Marketing*, vol. 38, no. 8, pp. 1656–1674, 2022.

[2] A. J. Rodriguez-Morales and O. H. Franco, "Public trust, misinformation and COVID-19 vaccination willingness in Latin America and the Caribbean: Today's key challenges," *Lancet Reg. Health–Amer.*, vol. 3, Sep. 2021, pp. 1–2.

[3] G. Verma, A. Bhardwaj, T. Aledavood, M. De Choudhury, and S. Kumar, "Examining the impact of sharing COVID-19 misinformation online on mental health," *Sci. Rep.*, vol. 12, no. 1, 2022, Art. no. 8045.

[4] S. Gupta, G. Jain, and A. A. Tiwari, "Polarised social media discourse during COVID-19 pandemic: Evidence from YouTube," *Behav. Inf. Technol.*, vol. 42, no. 2, pp. 227–248, 2023.

[5] J. Roozenbeek, S. Van Der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky, "Psychological inoculation improves resilience against misinformation on social media," *Sci. Adv.*, vol. 8, no. 34, 2022, Art. no. eabo6254.

[6] A. M. Guess et al., "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India," *Proc. Nat. Acad. Sci.*, vol. 117, no. 27, pp. 15536–15545, 2020.

[7] N. Micallef, V. Armacost, N. Memon, and S. Patil, "True or false: Studying the work practices of professional fact-checkers," *Proc. ACM Human-Comput. Interact.*, vol. 6, no. CSCW1, pp. 1–44, 2022.

[8] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, "The spread of COVID-19 conspiracy theories on social media and the effect of content moderation," *Harvard Kennedy School Misinf. Rev.*, vol. 1, no. 3, 2020, doi: 10.37016/mr-2020-034.

[9] T. Ahmad, M. S. Faisal, A. Rizwan, R. Alkanhel, P. W. Khan, and A. Muthanna, "Efficient fake news detection mechanism using enhanced deep learning model," *Appl. Sci.*, vol. 12, no. 3, 2022, Art. no. 1743.

[10] A. Rafique, F. Rustam, M. Narra, A. Mehmood, E. Lee, and I. Ashraf, "Comparative analysis of machine learning methods to detect fake news in an Urdu language corpus," *PeerJ Comput. Sci.*, vol. 8, 2022, Art. no. e1004.

[11] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: A comprehensive review," *Artif. Intell. Rev.*, pp. 1–66, vol. 55, pp. 3503–3568. doi: 10.1007/s10462-021-10088-y.

[12] W. Shahid et al., "Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities," *IEEE Trans. Computat.* Social Syst., early access, Jun. 6, 2022.

[13] M. L. Fransen, E. G. Smit, and P. W. Verlegh, "Strategies and motives for resistance to persuasion: An integrative framework," *Frontiers Psychol.*, vol. 6, Aug. 2015, Art. no. 1201.

[14] U. K. Ecker et al., "The psychological drivers of misinformation belief and its resistance to correction," *Nature Rev. Psychol.*, vol. 1, no. 1, pp. 13–29, 2022.

[15] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," 2020, *arXiv:2006.11371*.

[16] B. Singh and D. K. Sharma, "SiteForge: Detecting and localizing forged images on microblogging platforms using deep convolutional neural network," *Comput. Ind. Eng.*, vol. 162, Dec. 2021, Art. no. 107733.

[17] F. Yang et al., "XFake: Explainable fake news detector with visualizations," in *Proc. World Wide Web Conf.*, 2019, pp. 3600–3604.

[18] Z. Kou, Y. Zhang, D. Zhang, and D. Wang, "CrowdGraph: A crowdsourcing multi-modal knowledge graph approach to explainable fauxtography detection," *Proc. ACM Human-Comput. Interact.*, vol. 6, no. CSCW2, pp. 1–28, 2022.

[19] S. Mohseni et al., "Machine learning explanations to prevent overtrust in fake news detection," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 15, 2021, pp. 421–431.

[20] S.-Y. Chien, C.-J. Yang, and F. Yu, "XFlag: Explainable fake news detection model on social media," *Int. J. Human–Comput. Interact.*, vol. 38, nos. 18–20, pp. 1808–1827, 2022.

[21] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[22] A. Asudeh, H. V. Jagadish, Y. Wu, and C. Yu, "On detecting cherry-picked trendlines," *Proc. VLDB Endowment*, vol. 13, no. 6, pp. 939–952, 2020.

[23] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding explainability: Towards social transparency in AI systems," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–19.

[24] U. Ehsan, K. Saha, M. De Choudhury, and M. O. Riedl, "Charting the sociotechnical gap in explainable AI: A framework to address the gap in XAI," *Proc. ACM Human-Comput. Interact.*, vol. 7, no. CSCW1, pp. 1–32, 2023.

[25] P. Meel and D. K. Vishwakarma, "Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities," *Expert Syst. Appl.*, vol. 153, 2020, Art. no. 112986, doi: 10.1016/j.eswa.2019.112986.

[26] A. Joy, A. Shrestha, and F. Spezzano, "Are you influenced? Modeling the diffusion of fake news in social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2021, pp. 184–188.

[27] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016, pp. 2972–2978.

[28] H. Guo, J. Cao, Y. Zhang, J. Guo, and J. Li, "Rumor detection with hierarchical social attention network," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 943–951.

[29] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 354–361.

[30] R. Sridhar, "Prediction of social influence for provenance of misinformation in online social network using big data approach," *Comput. J.*, vol. 64, no. 3, pp. 391–407, 2021.

[31] D. Y. Zhang et al., "FauxBuster: A content-free fauxtography detector using social media comments," in *Proc. IEEE Int. Conf. Big Data (Big Bata)*. Piscataway, NJ, USA: IEEE Press, 2018, pp. 891–900.

[32] L. Shang, Z. Kou, Y. Zhang, and D. Wang, "A duo-generative approach to explainable multimodal COVID-19 misinformation detection," in *Proc. ACM Web Conf.*, 2022, pp. 3623–3631.

[33] J. Wanner, L.-V. Herm, K. Heinrich, and C. Janiesch, "The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study," *Electron. Markets*, vol. 32, no. 4, pp. 2079–2102, 2022.

[34] H. C. Kelman, "Compliance, identification, and internalization three processes of attitude change," *J. Conflict Resolution*, vol. 2, no. 1, pp. 51–60, 1958.

[35] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 430–435.

[36] K. Shu, X. Zhou, S. Wang, R. Zafarani, and H. Liu, "The role of user profiles for fake news detection," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2019, pp. 436–439.

[37] Y. Long, Q. Lu, R. Xiang, M. Li, and C.-R. Huang, "Fake news detection through multi-perspective speaker profiles," in *Proc. 8th Int. Joint Conf. Natural Lang. Process. (Short Papers)*, vol. 2, 2017, pp. 252–256.

[38] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 675–684.

[39] S. K. Uppada, K. Manasa, B. Vidhathri, R. Harini, and B. Sivaselvan, "Novel approaches to fake news and fake account detection in OSNS: User social engagement and visual content centric model," *Social Netw. Anal. Mining*, vol. 12, no. 1, 2022, Art. no. 52.

[40] Y. Zhao, J. Da, and J. Yan, "Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches," *Inf. Process. Manage.*, vol. 58, no. 1, 2021, Art. no. 102390.

[41] M. Glenski, T. Weninger, and S. Volkova, "Propagation from deceptive news sources who shares, how much, how evenly, and how quickly?" *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 4, pp. 1071–1082, Dec. 2018.

[42] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[43] S. Kwon and M. Cha, "Modeling bursty temporal pattern of rumors," in *Proc. Int. AAAI Conf. Web Social Media*, 2014, vol. 8, no. 1, pp. 650–651.

[44] Z. Zhao et al., "Fake news propagates differently from real news even at early stages of spreading," *EPJ Data Sci.*, vol. 9, no. 1, 2020, Art. no. 7.

[45] L. Wu, F. Morstatter, K. M. Carley, and H. Liu, "Misinformation in social media: Definition, manipulation, and detection," *ACM SIGKDD Explor. Newslett.*, vol. 21, no. 2, pp. 80–90, 2019.

[46] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 637–645.

[47] S. Jiang and C. Wilson, "Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media," *Proc. ACM Human-Comput. Interact.*, 2018, vol. 2, no. CSCW, pp. 1–23.

[48] A. Sharma and D. Cosley, "Do social explanations work? Studying and modeling the effects of social explanations in recommender systems," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 1133–1144.

[49] I. Guy et al., "Personalized recommendation of social software items based on social relations," in *Proc. 3rd ACM Conf. Recommender Syst.*, 2009, pp. 53–60.

[50] P. Bonhard and M. A. Sasse, "'Knowing me, knowing you'—Using profiles and social networking to improve recommender systems," *BT Technol. J.*, vol. 24, no. 3, pp. 84–98, 2006.

[51] A. Starke, M. Willemsen, and C. Snijders, "Promoting energy-efficient behavior by depicting social norms in a recommender interface," *ACM Trans. Interactive Intell. Syst. (TiiS)*, vol. 11, nos. 3–4, pp. 1–32, 2021.

[52] M. Trunfio and S. Rossi, "Conceptualising and measuring social media engagement: A systematic literature review," *Italian J. Marketing*, vol. 2021, pp. 267–292, Aug. 2021.

[53] H. B. Kang et al., "From who you know to what you read: Augmenting scientific recommendations with implicit social networks," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2022, pp. 1–23.

[54] C. Moser, S. Y. Schoenebeck, and P. Resnick, "Impulse buying: Design practices and consumer needs," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2019, pp. 1–15.

[55] I. Krsek, K. Wenzel, S. Das, J. I. Hong, and L. Dabbish, "To self-persuade or be persuaded: Examining interventions for users' privacy setting selection," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2022, pp. 1–17.

[56] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, "Is seeing believing? How recommender system interfaces affect users' opinions," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2003, pp. 585–592.

[57] H. Zhu, B. Huberman, and Y. Luon, "To switch or not to switch: Understanding social influence in online choices," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2012, pp. 2257–2266.

[58] R. B. Cialdini, *Influence: Science and Practice*, vol. 4. Boston, MA, USA: Pearson Education, 2009.

[59] J. T. Buchanan, E. J. Henig, and M. I. Henig, "Objectivity and subjectivity in the decision making process," *Ann. Oper. Res.*, vol. 80, nos. 1–4, pp. 333–345, 1998.

[60] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[61] M. Deutsch and H. B. Gerard, "A study of normative and informational social influences upon individual judgment," *J. Abnormal Social Psychol.*, vol. 51, no. 3, pp. 629–636, 1955.

[62] M. Avram, N. Micallef, S. Patil, and F. Menczer, "Exposure to social engagement metrics increases vulnerability to misinformation," 2020, *arXiv:2005.04682*.

[63] A. Kim, P. L. Moravec, and A. R. Dennis, "Combating fake news on social media with source ratings: The effects of user and expert reputation ratings," *J. Manage. Inf. Syst.*, vol. 36, no. 3, pp. 931–968, 2019.

[64] J. Colliander, "'This is fake news': Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media," *Comput. Human Behav.*, vol. 97, pp. 202–215, Aug. 2019.

[65] C. I. Hovland and W. Weiss, "The influence of source credibility on communication effectiveness," *Public Opinion Quart.*, vol. 15, no. 4, pp. 635–650, 1951.

[66] S. Winter and N. C. Krämer, "A question of credibility–effects of source cues and recommendations on information selection on news sites and blogs," *Communications*, vol. 39, no. 4, pp. 435–456, 2014.

[67] W. Duan, B. Gu, and A. B. Whinston, "Informational cascades and software adoption on the internet: An empirical investigation," *MIS Quart.*, vol. 33, no. 1, pp. 23–48, 2009.

[68] M. D. Kearney, S. C. Chiang, and P. M. Massey, "The Twitter origins and evolution of the COVID-19 "pandemic" conspiracy theory," *Harvard Kennedy School Misinf. Rev.*, vol. 1, no. 3, pp. 1–18, 2020, doi: 10.37016/mr-2020-42.

[69] W. Goffman and V. Newill, "Generalization of epidemic theory," *Nature*, vol. 204, no. 4955, pp. 225–228, 1964.

[70] M. Del Vicario et al., "The spreading of misinformation online," *Proc. Nat. Acad. Sci.*, vol. 113, no. 3, pp. 554–559, 2016.

[71] M. Himelein-Wachowiak et al., "Bots and misinformation spread on social media: Implications for COVID-19," *J. Med. Internet Res.*, vol. 23, no. 5, 2021, Art. no. e26933.

[72] P. Majerczak and A. Strzelecki, "Trust, media credibility, social ties, and the intention to share towards information verification in an age of fake news," *Behav. Sci.*, vol. 12, no. 2, 2022, Art. no. 51.

[73] I. K. Mensah, M. K. Khan, J. Liang, N. Zhu, L.-W. Lin, and D. S. Mwakapesa, "The moderating influence of perceived government information transparency on COVID-19 pandemic information adoption on social media systems," *Frontiers Psychol.*, vol. 14, Jun. 2023, Art. no. 1172094.

[74] M. A. Ruani and M. J. Reiss, "Susceptibility to COVID-19 nutrition misinformation and eating behavior change during lockdowns: An international web-based survey," *Nutrients*, vol. 15, no. 2, 2023, Art. no. 451.

[75] P. Cadet and M. D. Feldman, "Pretense of a paradox: Factitious intersex conditions on the internet," *Int. J. Sexual Health*, vol. 24, no. 2, pp. 91–96, 2012.

[76] J. Shin, L. Jian, K. Driscoll, and F. Bar, "The diffusion of misinformation on social media: Temporal pattern, message, and source," *Comput. Human Behav.*, vol. 83, pp. 278–287, Jun. 2018.

[77] N. Geraee, M. H. Kaveh, D. Shojaeizadeh, and H. R. Tabatabaee, "Impact of media literacy education on knowledge and behavioral intention of adolescents in dealing with media messages according to stages of change," *J. Adv. Med. Educ. Professionalism*, vol. 3, no. 1, pp. 9–14, 2015.

[78] N. Pröllochs, "Community-based fact-checking on Twitter's birdwatch platform," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 16, 2022, pp. 794–805.

[79] O. Vitman, Y. Kostiuk, G. Sidorov, and A. Gelbukh, "Sarcasm detection framework using context, emotion and sentiment features," *Expert Syst. Appl.*, vol. 234, 2023, Art. no. 121068, doi: 10.1016/j.eswa.2023.121068.

[80] A. Piepenbrink and A. S. Gaur, "Topic models as a novel approach to identify themes in content analysis," *Acad. Manage. Proc.*, vol. 2017, no. 1, 2017, Art. no. 11335.

[81] P. Bose, S. Srinivasan, W. C. Sleeman IV, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts," *Appl. Sci.*, vol. 11, no. 18, 2021, Art. no. 8319.

[82] M. S. Mredula, N. Dey, M. S. Rahman, I. Mahmud, and Y.-Z. Cho, "A review on the trends in event detection by analyzing social media platforms' data," *Sensors*, vol. 22, no. 12, 2022, Art. no. 4531.

[83] M. Yu et al., "Spatiotemporal event detection: A review," *Int. J. Digit. Earth*, vol. 13, no. 12, pp. 1339–1365, 2020.

[84] L. Huang, P. Shi, H. Zhu, and T. Chen, "Early detection of emergency events from social media: A new text clustering approach," *Natural Hazards*, vol. 111, no. 1, pp. 851–875, 2022.

[85] M. A. C. Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 32, no. 6, pp. 635–646, 2020.

[86] Y. Liu et al., "Summary of ChatGPT/GPT-4 research and perspective towards the future of large language models," 2023, *arXiv:2304.01852*.

[87] W. Hariri, "Unlocking the potential of ChatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing," 2023, *arXiv:2304.02017*.

[88] D. K. Sharma, B. Singh, S. Agarwal, H. Kim, and R. Sharma, "Sarcasm detection over social media platforms using hybrid auto-encoder-based model," *Electronics*, vol. 11, no. 18, 2022, Art. no. 2844.

[89] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[90] M. Azzimonti and M. Fernandes, "Social media networks, fake news, and polarization," Nat. Bureau of Econ. Res., Cambridge, MA, USA, Tech. Rep., 2018. [Online]. Available: https://doi.org/10.1016/j.ejpoleco.2022.102256

[91] K. Das, S. Samanta, and M. Pal, "Study on centrality measures in social networks: A survey," *Social Netw. Anal. Mining*, vol. 8, pp. 1–11, Feb. 2018, Art. no. 28.

[92] K. Batool and M. A. Niazi, "Towards a methodology for validation of centrality measures in complex networks," *PLoS One*, vol. 9, no. 4, 2014, Art. no. p. e90283.

[93] P. Bedi and C. Sharma, "Community detection in social networks," *Wiley Interdisciplinary Rev. Data Mining Knowl. Discovery*, vol. 6, no. 3, pp. 115–135, 2016.

[94] A. Amira, A. Derhab, S. Hadjar, M. Merazka, M. G. R. Alam, and M. M. Hassan, "Detection and analysis of fake news users' communities in social media," *IEEE Trans. Comput. Social Syst.*, early access, Jun. 15, 2023.

[95] D. F. Tokojima Machado, A. Fioravante de Siqueira, N. Rallo Shimizu, and L. Gitahy, "It-which-must-not-be-named: COVID-19 misinformation, tactics to profit from it and to evade content moderation on YouTube," *Frontiers Commun.*, vol. 7, Nov. 2022, pp. 1–14.

[96] N. R. de Oliveira, P. S. Pisa, M. A. Lopez, D. S. V. de Medeiros, and D. M. Mattos, "Identifying fake news on social networks based

on natural language processing: Trends and challenges," *Information*, vol. 12, no. 1, 2021, Art. no. 38.

[97] I. Lymperopoulos and G. Lekakos, "Analysis of social network dynamics with models from the theory of complex adaptive systems," in *Proc. Collaborative, Trusted Privacy-Aware e/m-Services: 12th IFIP WG 6.11 Conf. e-Bus., e-Services, e-Soc. (I3E)*, Athens, Greece. New York, NY, USA: Springer-Verlag, 2013, pp. 124–140.

[98] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. Part II," *Annu. Rev. Control*, vol. 45, pp. 166–190, Apr. 2018.

[99] R. Jain et al., "Explaining sentiment analysis results on social media texts through visualization," *Multimedia Tools Appl.*, vol. 82, no. 15, pp. 1–17, 2023.

[100] R. Akula and I. Garibay, "VizTract: Visualization of complex social networks for easy user perception," *Big Data Cogn. Comput.*, vol. 3, no. 1, 2019, Art. no. 17.

[101] D. Sehnan, V. Goel, S. Masud, C. Jain, V. Goyal, and T. Chakraborty, "DiVA: A scalable, interactive and customizable visual analytics platform for information diffusion on large networks," *ACM Trans. Knowl. Discovery Data*, vol. 17, no. 4, pp. 1–33, 2023.

[102] J. Fulda, M. Brehmer, and T. Munzner, "TimeLineCurator: Interactive authoring of visual timelines from unstructured text," *IEEE Trans. Visualization Comput. Graph.*, vol. 22, no. 1, pp. 300–309, Jan. 2015.

[103] L. T. Becker and E. M. Gould, "Microsoft power BI: Extending excel to manipulate, analyze, and visualize diverse data," *Serials Rev.*, vol. 45, no. 3, pp. 184–188, 2019.

[104] Y. Jiang et al., "Computational approaches for understanding, generating, and adapting user interfaces," in *Proc. CHI Conf. Human Factors Comput. Syst. Extended Abstr.*, 2022, pp. 1–6.

[105] Y. Jiang, Y. Lu, C. Lutteroth, T. J.-J. Li, J. Nichols, and W. Stuerzlinger, "The future of computational approaches for understanding and adapting user interfaces," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, 2023, p. 1–5.

[106] M. Short and A. G. Cole, "Factors associated with e-cigarette escalation among high school students: A review of the literature," *Int. J. Environ. Res. Public Health*, vol. 18, no. 19, 2021, Art. no. 10067.

[107] Z. Epstein, N. Sirlin, A. Arechar, G. Pennycook, and D. Rand, "The social media context interferes with truth discernment," *Sci. Adv.*, vol. 9, no. 9, 2023, Art. no. eabo6169.

[108] J. Liu, S. M. Gaiha, and B. Halpern-Felsher, "A breath of knowledge: Overview of current adolescent e-cigarette prevention and cessation programs," *Current Addiction Rep.*, vol. 7, pp. 520–532, Dec. 2020.

[109] A. Pérez-Escoda, L. M. Pedrero-Esteban, J. Rubio-Romero, and C. Jiménez-Narros, "Fake news reaching young people on social networks: Distrust challenging media literacy," *Publications*, vol. 9, no. 2, 2021, Art. no. 24.

[110] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Apr. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[111] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 27730–27744, Dec. 2022.

[112] T. Nelson, N. Kagan, C. Critchlow, A. Hillard, and A. Hsu, "The danger of misinformation in the COVID-19 crisis," *Missouri Med.*, vol. 117, no. 6, pp. 510–512, 2020.

[113] R. Chen, B.-J. Fwu, T.-R. Yang, Y.-K. Chen, and Q.-A. N. Tran, "To mask or not to mask: Debunking the myths of mask-wearing during Covid-19 across cultures," *PLoS One*, vol. 17, no. 9, 2022, Art. no. e0270160.

[114] S. Rajeh and H. Cherifi, "Ranking influential nodes in complex networks with community structure," *PLoS One*, vol. 17, no. 8, 2022, Art. no. e0273610.

[115] P. Juneja, D. Rama Subramanian, and T. Mitra, "Through the looking glass: Study of transparency in reddit's moderation practices," *Proc. ACM Human-Comput. Interact.*, 2020, vol. 4, no. GROUP, pp. 1–35.

[116] S. Muhammed T and S. K. Mathew, "The disaster of misinformation: A review of research in social media," *Int. J. Data Sci. Analytics*, vol. 13, no. 4, pp. 271–285, 2022.

[117] J. Alghamdi, Y. Lin, and S. Luo, "A comparative study of machine learning and deep learning techniques for fake news detection," *Information*, vol. 13, no. 12, 2022, Art. no. 576.

[118] M. Lease, "On quality control and machine learning in crowdsourcing," in *Proc. Workshops 25th AAAI Conf. Artif. Intell.*, 2011, pp. 97-102.

[119] J. Lu, W. Li, Q. Wang, and Y. Zhang, "Research on data quality control of crowdsourcing annotation: A survey," in *Proc. IEEE Int. Conf. Dependable, Autonomic Secure Comput./Int. Conf. Pervasive Intell. Comput./Int. Conf Cloud Big Data Comput./Int. Conf. Cyber Sci. Technol. Congr. (DASC/PiCom/CBDCom/CyberSciTech)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 201–208.

[120] C. Doss et al., "Deepfakes and scientific knowledge dissemination," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 13429.

[121] M. R. Islam, S. Liu, X. Wang, and G. Xu, "Deep learning for misinformation detection on online social networks: A survey and new perspectives," *Social Netw. Anal. Mining*, vol. 10, pp. 1–20, Sep. 2020, Art. no. 82.

[122] D. Wang et al., "Using humans as sensors: An estimation-theoretic perspective," in *Proc. 13th Int. Symp. Inf. Process. Sensor Netw. (IPSN-14)*, Piscataway, NJ, USA: IEEE Press, 2014, pp. 35–46.

[123] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 195–208, Feb. 2018.

[124] C. Huang, D. Wang, and N. V. Chawla, "Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems," *IEEE Trans. Big Data*, vol. 6, no. 4, pp. 702–713, Apr. 2017.

[125] X. Wu, L. Xiao, Y. Sun, J. Zhang, T. Ma, and L. He, "A survey of human-in-the-loop for machine learning," *Future Gener. Comput. Syst.*, vol. 135, pp. 364–381, Oct. 2022.

[126] J. Lu, "Themes and evolution of misinformation during the early phases of the COVID-19 outbreak in China—An application of the crisis and emergency risk communication model," *Frontiers Commun.*, vol. 5, Aug. 2020, Art. no. 57.

[127] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychol. Sci. Public Interest*, vol. 13, no. 3, pp. 106–131, 2012.

[128] L. Du et al., "Gordie howe's 'miraculous treatment': Case study of Twitter users' reactions to a sport celebrity's stem cell treatment," *JMIR Public Health Surveillance*, vol. 2, no. 1, 2016, Art. no. p. e5264.

[129] M. Scriven, "An essential unpredictability in human behavior," in *Scientific Psychology*: Principles and Approaches. New York, NY, USA: US Information Agency, Voice of America Forum, 1964.

[130] A. Malik, F. Bashir, and K. Mahmood, "Antecedents and consequences of misinformation sharing behavior among adults on social media during COVID-19," *SAGE Open*, vol. 13, no. 1, 2023, Art. no. 21582440221147022.

[131] E. Dujeancourt and M. Garz, "The effects of algorithmic content selection on user engagement with news on Twitter," *Inf. Soc.*, vol. 39, no. 5, pp. 1–19, 2023.

[132] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Rev. General Psychol.*, vol. 2, no. 2, pp. 175–220, 1998.

[133] H. Song et al., "Trusting social media as a source of health information: Online surveys comparing the United States, Korea, and Hong Kong," *J. Med. Internet Res.*, vol. 18, no. 3, 2016, Art. no. e25.

[134] F. Petroni et al., "Language models as knowledge bases?" 2019, *arXiv:1909.01066*.

[135] Q. V. Liao and J. W. Vaughan, "AI transparency in the age of LLMS: A human-centered research roadmap," 2023, *arXiv:2306.01941*.

[136] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," 2023, *arXiv:2302.00487*.

[137] R.-K. Sheu and M. S. Pardeshi, "A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system," *Sensors*, vol. 22, no. 20, 2022, Art. no. 8068.

[138] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 2065–2073, 2014.

[139] A. Panteli and B. Boutsinas, "Addressing the cold-start problem in recommender systems based on frequent patterns," *Algorithms*, vol. 16, no. 4, 2023, Art. no. 182.

[140] C. Zheng, W. Shao, X. Chen, B. Zhang, G. Wang, and W. Zhang, "Real-world effectiveness of COVID-19 vaccines: A literature review and meta-analysis," *Int. J. Infectious Diseases*, vol. 114, pp. 252–260, Jan. 2022.

[141] A. M. Cavanaugh, K. B. Spicer, D. Thoroughman, C. Glick, and K. Winter, "Reduced risk of reinfection with SARS-Cov-2 after COVID-19 vaccination—Kentucky, May–Jun. 2021," *Morbidity Mortality Weekly Rep.*, vol. 70, no. 32, 2021, Art. no. 1081.

[142] S. T. Parker, "Measuring gun violence in police data sources: Transitioning to NIBRS," *Injury Epidemiol.*, vol. 9, no. 1, 2022, Art. no. 15.

[143] G. Pennycook and D. G. Rand, "Nudging social media toward accuracy," *Ann. Amer. Acad. Political Social Sci.*, vol. 700, no. 1, pp. 152–164, 2022.

[144] C. Carrasco-Farré, "The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions," *Humanities Social Sci. Commun.*, vol. 9, no. 1, pp. 1–18, 2022.

[145] L. Shang et al., "MMAdapt: A knowledge-guided multi-source multi-class domain adaptive framework for early health misinformation detection," in *Proc. ACM Web Conf. (WWW)*, Special Track on Web4Good, 2024, pp. 4653–4663.

[146] L. Shang, Y. Zhang, Z. Yue, J. Choi, H. Zeng, and D. Wang, "A domain adaptive graph learning framework to early detection of emergent healthcare misinformation on social media," in *Proc. Int. AAAI Conf. Web Social Media (ICWSM)*, 2024, pp. 1408–1421.

[147] J. Rabelo, R. B. Prudêncio, and F. Barros, "Collective classification for sentiment analysis in social networks," in *Proc. IEEE 24th Int. Conf. Tools Artif. Intell.*, vol. 1, Piscataway, NJ, USA: IEEE Press, 2012, pp. 958–963.

[148] B. Sluban, J. Smailović, S. Battiston, and I. Mozetič, "Sentiment leaning of influential communities in social networks," *Comput. Social Netw.*, vol. 2, pp. 1–21, Jul. 2015.

[149] M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray, *Algorithms for Decision Making*. Cambridge, MA, USA: MIT Press, 2022.

[150] M. Riveiro and S. Thill, "The challenges of providing explanations of AI systems when they do not behave like users expect," in *Proc. 30th ACM Conf. User Model., Adaptation Personalization*, 2022, pp. 110–120.

[151] L. Chimoyi et al., "The geography of COVID-19 misinformation: Using geospatial maps for targeted messaging to combat misinformation on COVID-19, South Africa," *BMC Res. Notes*, vol. 14, no. 1, 2021, Art. no. 468.

[152] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *Proc. 25th USENIX Secur. Symp. (USENIX Secur. 16)*, 2016, pp. 601–618.

[153] Z. Yang and Z. Liang, "Automated identification of sensitive data from implicit user specification," *Cybersecurity*, vol. 1, pp. 1–15, Sep. 2018.

[154] P. Fleming, A. P. Bayliss, S. G. Edwards, and C. R. Seger, "The role of personal data value, culture and self-construal in online privacy behaviour," *PLoS One*, vol. 16, no. 7, 2021, Art. no. e0253568.

[155] Y. Li, "Cross-cultural privacy differences," in *Modern Socio-Technical Perspectives on Privacy*. Cham, Switzerland: Springer International Publishing-Verlag, 2022, pp. 267–292.

[156] K. Bryanov and V. Vziatysheva, "Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news," *PLoS One*, vol. 16, no. 6, 2021, Art. no. e0253717.

[157] M. Gupta, D. Dennehy, C. M. Parra, M. Mäntymäki, and Y. K. Dwivedi, "Fake news believability: The effects of political beliefs and espoused cultural values," *Inf. Manage.*, vol. 60, no. 2, 2023, Art. no. 103745.

[158] W. Y. Wang, "'Liar, liar pants on fire': A new benchmark dataset for fake news detection," 2017, *arXiv:1705.00648*.

[159] V. R. Bhargava and M. Velasquez, "Ethics of the attention economy: The problem of social media addiction," *Bus. Ethics Quart.*, vol. 31, no. 3, pp. 321–359, 2021.

[160] C. Gerlitz and A. Helmond, "The like economy: Social buttons and the data-intensive web," *New Media Soc.*, vol. 15, no. 8, pp. 1348–1365, 2013.

[161] M. Fernández, A. Bellogín, and I. Cantador, "Analysing the effect of recommendation algorithms on the amplification of misinformation," 2021, *arXiv:2103.14748*.

[162] S. C. Rhodes, "Filter bubbles, echo chambers, and fake news: How social media conditions individuals to be less critical of political misinformation," *Political Commun.*, vol. 39, no. 1, pp. 1–22, 2022.

[163] K. Kenthapadi, I. Mironov, and A. G. Thakurta, "Privacy-preserving data mining in industry," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, 2019, pp. 840–841.

[164] G. Shailaja and C. G. Rao, "Robust and lossless data privacy preservation: optimal key based data sanitization," *Evol. Intell.*, vol. 15, no. 2, pp. 1123–1134, 2022.

[165] A. Kozyreva et al., "Resolving content moderation dilemmas between free speech and harmful misinformation," *Proc. Nat. Acad. Sci.*, vol. 120, no. 7, 2023, Art. no. e2210666120.

[166] J. A. Gallo and C. Y. Cho, "Social media: Misinformation and content moderation issues for congress," *Congressional Res. Service Rep.*, vol. 46662, pp. 1–32, Jan. 2021.

[167] P.-L. Chen, Y.-C. Cheng, and K. Chen, "Analysis of social media data: An introduction to the characteristics and chronological process," *Big Data Comput. Social Sci. Humanities*, pp. 297–321, Nov. 2018.

[168] M. Tajrian, A. Rahman, M. A. Kabir, and M. R. Islam, "A review of methodologies for fake news analysis," *IEEE Access*, vol. 11, pp. 73879–73893, 2023, doi: 10.1109/ACCESS.2023.3294989.

[169] S. Ghosh and P. Mitra, "Catching lies in the act: A framework for early misinformation detection on social media," in *Proc. 34th ACM Conf. Hypertext Social Media*, 2023, pp. 1–12.

[170] P. M. Konkobo, R. Zhang, S. Huang, T. T. Minoungou, J. A. Ouedraogo, and L. Li, "A deep learning model for early detection of fake news on social media," in *Proc. 7th Int. Conf. Behav. Social Comput. (BESC)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 1–6.

[171] L. Shang, Y. Zhang, Z. Yue, Y. Choi, H. Zeng, and D. Wang, "A knowledge-driven domain adaptive approach to early misinformation detection in an emergent health domain on social media," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 34–41.

[172] Z. Yue, H. Zeng, Z. Kou, L. Shang, and D. Wang, "Contrastive domain adaptation for early misinformation detection: A case study on COVID-19," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 2423–2433.

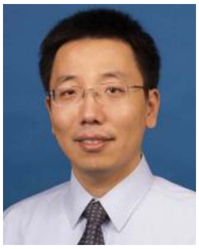
Yeaeun Gong received the B.A. degree in psychology from Yonsei University, and the M.S. degree in cognitive science from Seoul National University. She is working toward the Ph.D. degree with the School of Information Sciences, University of Illinois Urbana-Champaign (UIUC), Champaign, IL, USA.

Her research interests include human-centered AI and human–AI interaction.

Lanyu Shang received the B.S. degree in applied mathematics from the University of California, Los Angeles (UCLA), and the M.S. degree in data science from New York University. She is working toward the Ph.D. degree with the School of Information Sciences, University of Illinois Urbana-Champaign (UIUC), Champaign, IL, USA.

Her research interest includes online misinformation detection using social media data.

Mr. Lanyu was the recipient of the Outstanding Graduate Teaching Award at the University of Notre Dame, the Best Paper Award at ACM/IEEE ASONAM 2022, and the Best Paper Honorable Mention at IEEE SmartComp 2022.



Dong Wang (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Illinois Urbana-Champaign (UIUC), Champaign, IL, USA, in 2012, where he is now an Associate Professor with the School of Information Sciences.

His research interests include the area of social (human-centric) sensing, intelligence and computing, human-centered AI, AI for social good, data quality, and big data analytics.

Dr. Wang was the recipient of the NSF CAREER Award, the Google Faculty Research Award, the Young Investigator Program (YIP) Award from the ARO, The Wing Kai Cheng Fellowship from the University of Illinois, the Best Paper Award of 2022 ACM/IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM), the Best Paper Award of 16th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), and the Best Paper Honorable Mention of 8th IEEE SmartComp.