



Re-designing Social Media to Promote User Correction of Misinformation: Experimenting With Two Novel Techniques and Their Interplay with News Importance

Selin Gurgun , Emily Arden-Close , Keith Phalp & Raian Ali

To cite this article: Selin Gurgun , Emily Arden-Close , Keith Phalp & Raian Ali (04 Dec 2025): Re-designing Social Media to Promote User Correction of Misinformation: Experimenting With Two Novel Techniques and Their Interplay with News Importance, International Journal of Human-Computer Interaction, DOI: [10.1080/10447318.2025.2586835](https://doi.org/10.1080/10447318.2025.2586835)

To link to this article: <https://doi.org/10.1080/10447318.2025.2586835>



© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 04 Dec 2025.



Submit your article to this journal [↗](#)



Article views: 88




View related articles [↗](#)



View Crossmark data [↗](#)

Re-designing Social Media to Promote User Correction of Misinformation: Experimenting With Two Novel Techniques and Their Interplay with News Importance

Selin Gurgun^a, Emily Arden-Close^a, Keith Phalp^a and Raian Ali^b 

^aFaculty of Science and Technology, Bournemouth University, Poole, UK; ^bCollege of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

ABSTRACT

A previous co-design research study proposed various user interface designs (UIDs) aimed at motivating users to engage in corrective actions and challenge others who post misinformation. This paper assesses the effectiveness of two such UIDs in promoting user correction, compared to the existing designs of mainstream social media platforms. The UIDs are “privately challenge toggle” (enabling an easy switch to private conversation mode with the misinformation poster) and “discuss section” (a dedicated thread for arguments to avoid cluttering the main comment box). Additionally, the study considers the perceived importance of the misinformation content as a variable to be accounted for. An online within-subject vignette experimental study with 306 UK Facebook users (111 male, 194 female and 1 non-binary) compared the proposed designs to the existing one, focusing on usability parameters such as effectiveness, acceptability, and comfort. Results indicate that participants rated usability of both proposed UIDs favorably compared to the existing design. A two-way repeated measures ANOVA exploring the impact of the interaction between the UIDs and content importance on the likelihood of engaging in user correction revealed a significant interaction: when the content is not important, the “privately challenge toggle” is more effective than the existing design. Conversely, when the content is important, users are more inclined to challenge misinformation using the existing design rather than the “discuss section.” This research provides insights into re-designing social media to promote user correction and highlights the importance of considering both the suitability of the UIDs and the perceived importance of misinformation posts.

KEYWORDS

Misinformation; social media design; behaviour change; user corrections; Nudge

1. Introduction

The proliferation of misinformation on social media is a growing concern due to its rapid spread, particularly across various social media platforms (Aïmeur et al., 2023). Various solutions have been proposed to combat this issue, including machine learning (Ahmed et al., 2021; Ibrishimova & Li, 2020), natural language processing (de Oliveira et al., 2021; Nakov et al., 2021), blockchain (Jing & Murugesan, 2019) and algorithmic approaches (Figueira & Oliveira, 2017). While the focus often revolves around these technological approaches, to effectively combat misinformation it is important to also consider whether the design and architecture of online platforms facilitate and promote the act of reporting and debating amongst users.

User interface design (UID) in Human-Computer Interaction (HCI) has long been recognized as a tool to influence user behavior, often through techniques such as gamification (Deterding et al., 2011), nudges (Caraban et al., 2019; Thaler & Sunstein, 2008), and persuasive technologies (Fogg, 2002; Oinas-Kukkonen & Harjumaa, 2009). While Whilst these approaches have primarily focused on motivating people to improve their own behaviors e.g., in relation to wellbeing (Langrial et al., 2012) and

academic performance (Widyasari et al., 2019), there is limited research on motivating people to adopt positive and pro-social behaviors in interactions with others.

Previous research has demonstrated the positive impact of social media features in interventions leading to increased motivation and engagement (Elaheebocus et al., 2018; Ma et al., 2010). However, the incorporation of social media features in digital behavior change interventions is an evolving area that requires deeper investigation. Several studies suggest that improved UID can enhance the quality of online discussions (Seering et al., 2019), facilitating meaningful conversations (Sukumaran et al., 2011; Taylor et al., 2019) and creating constructive discussion environments (Baughan et al., 2021; Kriplean et al., 2012). Recent research also highlights design considerations that influence users' comments on online news (Kiskola et al., 2023).

In addressing misinformation, research also explored the effectiveness of UID through identification, flagging or sharing mechanisms. For instance, one study found that fact-checking flags, which are indicators used by platforms to indicate whether the content has been fact-checked, can impact users' identification of fake news (Gaozhao, 2021). Similarly, Kim et al. (2019) discuss how changes in UID, such as incorporating source ratings, can prompt users to be more critical toward news, thereby helping combat the spread of misinformation. Similarly, research has explored the efficacy of the design of symbols of fake news flags, such as the stop symbol and caution symbol in user interaction with misinformation on social media platforms (Figl et al., 2023). Additionally, one study demonstrated that flagging false news reduces false news-sharing intentions (Mena, 2020). These studies highlight the importance of platform design and architecture in influencing user behavior.

To date, few studies have investigated the importance of UID in motivating users to counter misinformation, i.e., in motivating user corrections, which is a complex and multifaceted behavior that entails a broad range of individual and social factors including altruism, bystander effect and fear of being attacked (Gurgun, Arden-Close, Phalp, et al., 2023). Regardless of the complexity of this behavior, research showed that user corrections, where social media users publicly challenge or refute misinformation shared by others (Bode & Vraga, 2018, 2021b; Vraga & Bode, 2020), are an effective way to combat misinformation. However, when people encounter misinformation, they often refrain from challenging others (Chadwick & Vaccari, 2019; Tandoc et al., 2020; Tully et al., 2020). This problem requires a thorough examination of the reasons (for a survey exploring the reasons, please refer to Gurgun et al. (2024)) and the development of solutions in order to overcome barriers to challenging misinformation. Designing interfaces that could encourage users to challenge misinformation is a potential solution to combat misinformation, based on evidence that interfaces that utilize persuasive techniques can influence likelihood of challenging misinformation on social media, relative to the existing Facebook interface (Gurgun et al., 2023). The literature highlights the potential of UID in shaping discussions, fostering critical engagement and encouraging challenging misinformation on social media.

Much like how design influences behavior, the importance of the content also plays a crucial role in influencing user engagement. When an issue is perceived as important, individuals are more likely to take action and discuss it frequently, motivated by a sense of civic responsibility (Gearhart & Zhang, 2014; Moy et al., 2001). While factors such as believability and alignment with preexisting opinions have been shown to impact user engagement (e.g., read, like, comment or share) (Kim et al., 2019; Kim & Dennis, 2019), this study seeks to explore the importance of the content in influencing the decision to challenge misinformation. For example, does it matter to users whether they challenge important and unimportant misinformation publicly or privately, or whether they address it in the comment section or require an additional section to address this content?

In addition to evaluating the perceived importance of the content, this study explores how design and content impact different user groups: those who are generally willing to engage in user corrections and those who are not. Research indicates that addressing misinformation on social media is not common, in the UK nearly 80% of users avoid informing others that content is false or exaggerated (Chadwick & Vaccari, 2019), and in a UK survey, although 58% of participants encountered misinformation, only 21% took action to rectify it (Vicol, 2020). For users already willing to challenge misinformation, social media design and content may have less influence, as their motivation is likely intrinsic. However, examining how design and content affect hesitant or unwilling users can provide deeper insights.

Previous study has identified user interface design requirements for encouraging users to challenge misinformation using co-design (Gurgun, Arden-Close, et al., 2024). Building on this, the current study evaluates two specific designs, which represent two main modalities. The first is “private toggle” which represents the distinction between private and public challenging. Private challenging refers to addressing misinformation in a manner where the identity of the person challenging is anonymous to everyone except the poster. The second is the “discuss section,” which allows users to challenge misinformation in a designated area. This represents the distinction between general comments on the post and in-depth discussions in a public setting. By evaluating the efficacy and user perceptions of these designs, this study assesses the impact of these approaches on users’ motivation to challenge misinformation on Facebook.

We investigate the following research questions:

1. Do the proposed designs differ from the existing design, where users have to comment publicly as part of a thread, in terms of usability, including effectiveness, acceptability, and comfort? If so, to what extent?
2. Is there an interaction between designs and perceived content importance with regard to impact on likelihood to challenge?
 2. To what extent do different designs (Private toggle, Discuss section, Existing) and perceived content importance (Unimportant, Important) influence people’s likelihood to challenge misinformation and how they interact with each other to influence likelihood to challenge?
3. Is there an interaction between designs and the perceived content importance regarding impact on likelihood to challenge *across two participants groups; those who are generally willing to challenge misinformation and those who are generally unwilling*?
 3. To what extent do different designs (Private toggle, Discuss section, Existing) and perceived content importance (Unimportant, Important) influence people’s likelihood to challenge misinformation *across the two willingness levels groups (willing and unwilling) and how do they compare to each other*?

Given the absence of prior research in this specific area, this study has the potential to lead to unique insights in terms of motivations to challenge misinformation and importance of content. Our results can guide future designs by enhancing understanding of designing more personalized and effective strategies for users to combat misinformation on social media.

2. Theoretical underpinning

User corrections, where social media users publicly challenge or refute misinformation shared by others, are generally perceived as effortful and difficult due to their potential to result in conflicts or damage relationships (Chadwick et al., 2022; Gurgun et al., 2024; Tandoc et al., 2020). Therefore, although user corrections have been suggested as an effective way to address misinformation (Bode, 2019; Emily K. Vraga & Bode, 2020), users typically do not challenge others when they encounter it (Bode & Vraga, 2021a; Tandoc et al., 2020; Tully et al., 2020). By not challenging misinformation, users contribute to its spread by remaining silent. While social media platforms have implemented measures to combat misinformation, there appears to be a paucity of research aimed at encouraging users to challenge misinformation.

Misinformation, which often spreads not only due to technical vulnerabilities but also because of social dynamics such as social influence (Gimpel et al., 2021; Lawson et al., 2023), can be considered a socio-technical issue. The absence of proactive efforts to challenge misinformation is related to both technological factors (technologies, algorithms and designs that do not allow or prioritize challenging misinformation online) and social factors (group dynamics such as interpersonal relationships and self-image concerns) (Gurgun et al., 2024; Tandoc et al., 2020). Therefore, two technical and social systems need to collaborate in order to effectively address these barriers. A narrow focus on a single perspective may lead to problems (Whitworth, 2011). For example, a solution that enables users to challenge misinformation without considering their desire to maintain their relationships may come at the cost of those relationships.

Research found that Facebook's policies on COVID – 19 vaccine misinformation such as removing anti-vaccine content were not successful due to the inherent design features (Broniatowski et al., 2023). It highlighted that Facebook is designed to motivate people to build communities and tailors users' feeds based on their existing beliefs, providing them with similar content that aligns with their beliefs. This can lead to the amplification of misinformation, making it difficult to control or stop. Therefore, there is a need to move beyond focusing solely on artificial intelligence and algorithms to address misinformation. In this context, user corrections can be an effective strategy to combat misinformation (Bode & Vraga, 2018; Emily K. Vraga & Bode, 2018; Walter & Murphy, 2018) and their comparable effectiveness relative to algorithmic corrections (Bode & Vraga, 2018). When developing a user interface design, there is a need to adopt a holistic approach that goes beyond mere visual esthetics by understanding the cognitive processes and social dynamics that shape users' interactions with social media platforms.

Design interventions that influence user behavior include diverse terminology such as persuasive techniques (Fogg, 2002; Oinas-Kukkonen & Harjumaa, 2009), nudging (Thaler & Sunstein, 2008) and affordances (Lockton et al., 2008). To prevent terminological confusion, we refer to design proposals in this study as UID interventions. These designs, namely the "Privately Challenge Toggle" (referred to as "private toggle" for brevity) and "Discuss Section" mechanisms, emerge as promising strategies to motivate users to challenge misinformation (Gurgun, Arden-Close, et al., 2024).

It is challenging to propose a one-size-fits-all solutions to promote confronting in Facebook. First, Facebook's system architecture, like many other social media platforms, does not provide dedicated features to facilitate the act of challenging misinformation, possibly because it considers user corrections doable with the current features such as commenting. On the contrary, social media allows users to access misinformation even after it has been removed (Broniatowski et al., 2023). Secondly, users have diverse motivations for using social media, like Facebook, such as passing time, information seeking and communication (McAndrew & Jeong, 2012; Nadkarni & Hofmann, 2012). Interventions must acknowledge the multi-layered nature of user interactions with social media design features. Promoting challenging misinformation through social media design requires the design features to be perceived as having benefits that outweigh their costs. These benefits can be emotional (accepted for providing an enjoyable experience such as gaming) or functional (used for their functionality such as protecting users from perceived negative consequences) (Venkatesh, 2000). In the latter case, users may use the features even if they are not enjoyable.

We used three parameters to assess usability. The International Organization for Standardization (ISO, 1999) defines usability as the extent to which specified users accomplish specific goals in particular environments, focusing on effectiveness, efficiency, and satisfaction. Frøkjær et al. (2000) similarly emphasize these three dimensions and advocate including them all in usability evaluations. In our context, efficiency was not measured because the interventions were not intended to minimize task completion time or optimize performance speed; instead, they aimed to encourage a socially sensitive behavior (challenging misinformation) where social and perceptual factors are more critical than operational speed. Although satisfaction is a core component of the ISO usability definition, in our study it was not included in its broad sense because it is typically measured as an overall post-use evaluation of a system, encompassing multiple experiential factors (e.g., enjoyment, fulfillment, ease. For the purposes of this research, we focused on constructs most relevant to changes in users' intentions and attitudes toward challenging misinformation. Drawing on Human-Computer Interaction (HCI) literature, we measured effectiveness, social acceptability, and comfort as key factors influencing whether users are willing to adopt and engage with such design (Davis, 1989). Effectiveness refers to the capability of the proposed design to accurately and completely achieve specified goals (ISO, 1999). In the Technology Acceptance Model (TAM) this aligns with the concept of perceived usefulness which measures the extent to which technology has improved potential users' performance toward specific goals (Davis, 1989). We measured effectiveness as a fundamental aspect of UID that contributes to the success and acceptance of a product or system within its' intended context. Acceptability in our context refers to how socially acceptable it is to use certain features, including whether users feel natural and whether the audience understands the users' actions and finds them weird or normal (Montero et al., 2010) The

social context plays a pivotal role in shaping expectations, and actual or anticipated disapproval from other people can influence the adoption and usage patterns of an interface (Koelle et al., 2019). Given that confrontation is often discouraged in online social media platforms, gaining insights regarding social acceptance can inform future designs in terms of users' expectations and social norms. The concept of comfort in HCI studies encompasses two main aspects: the level of satisfaction users feel regarding physical comfort whilst using the interface (ISO, 2011) and the social acceptability of the system (Koelle et al., 2020). When user interfaces lack quality, users may experience physical discomfort (Kivijärvi & Pärnänen, 2023). Hornbæk (2006) identified various examples of physical discomfort related to perceptions of interfaces, including sore eyes, upper body discomfort, overall muscular discomfort, and fatigue. Regardless of participants' perception of the concept, our aim is to assess the extent to which they would use these designs with ease.

Building on prior research employing co-design methods (Gurgun et al., 2024), our study evaluates the design mechanisms, "private toggle" and "discuss section," as subsets of a broader framework of UI interventions aimed at empowering users to critically engage with misinformation. We also provided the "existing Facebook user interface" to compare it to both design mechanisms.

2.1. Privately challenge toggle

The Privately Challenge Toggle is a feature that allows users to switch their comment visibility settings to private. When they choose "private," their comment and name will be visible to only the original poster. Others will only see that someone commented about the content's veracity (See Figure 1). The idea of semi-anonymity was thought to be effective in encouraging users to challenge misinformation in a previous co-design study (Gurgun, Arden-Close, et al., 2024). First, it is intended to make users feel safe by instilling a sense of protection against personal attacks. Research on online disinhibition suggests that when social cues are limited and interactions are not fully synchronous in online environments, people may experience a sense of anonymity. This can lead to reduced self-awareness, making it difficult for them to perceive themselves as others do and to regulate their behavior accordingly (Lapidot-Lefler & Barak, 2012; Suler, 2004; Wu et al., 2017). One of the main reasons users do not challenge misinformation is fear of others' negative reactions (Gurgun et al., 2024). The "privately challenge toggle" design aims to prevent anticipated backlash from other users while informing other people that this content is being challenged.

Secondly, social identity theory (Langrall et al.) suggests people define their self-concept based on the social groups they belong to (Tajfel and Turner (2004). People's social context, categorization, adherence to social norms and perceived status all impact their sense of self. To cultivate a positive identity, they use different impression management strategies to control or influence the impressions that others form (Fuller et al., 2007; Rosenfeld et al., 2001). Users often refrain from challenging due to concerns about how they may be perceived (Gurgun et al., 2024). Therefore, this design aims to create a sense of security for users when challenging misinformation, helping them feel more comfortable expressing their concerns.

Research established that observing others being corrected can lead to users updating their own attitudes (Bode & Vraga, 2015; Emily K Vraga & Bode, 2017). According to these studies, those who observe others being challenged are less susceptible to cognitive dissonance, as their identity is not directly threatened, making them more open to correction. Therefore, this design aims to both protect the commenter's identity from others while also indicating to the group that this topic has been challenged by a member.

2.2. Discuss section

The proposed "discuss section" encourages more in-depth engagement by separating comments from the discussion section. In the "discuss section," users can initiate and participate in extended conversations (See Figure 2). It provides a space for discussions, encouraging users to engage beyond brief comments and reactions. By default, when viewing a post, users see "Comments," which is the standard Facebook experience, allowing them to read and engage with comments. However, with the "Discuss" section option, users can participate in and read more extensive discussions.

The intention behind this separated section is to create a space for users to engage in critical discussions about the veracity of the content as social media platforms are typically not designed to effectively

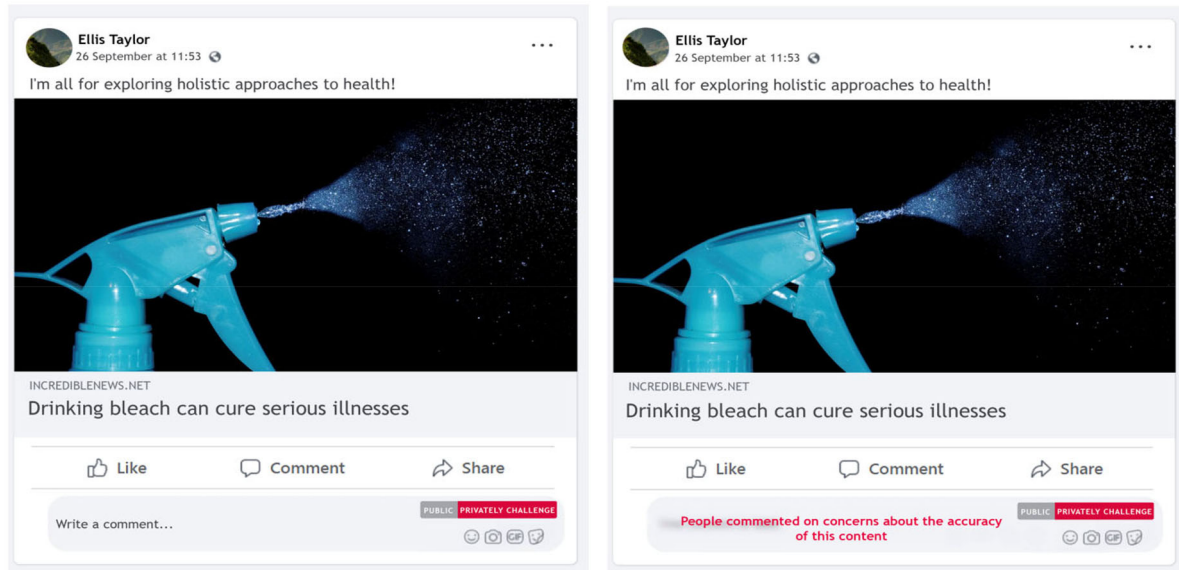


Figure 1. Design sketches for privately challenge Toggle.

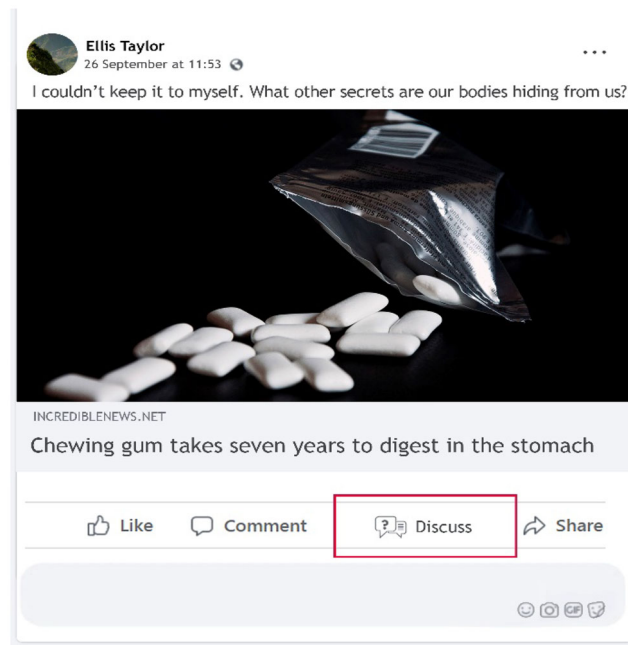


Figure 2. Design sketches for the discuss section.

address misinformation. The reasons people often mention for not challenging misinformation are primarily rooted in social concerns, including fears of being involved in heated arguments and anticipation that their relationship might be harmed (Gurgun et al., 2024). When users identify questionable content, the presence of a dedicated section for questioning may prompt them to evaluate the content. Leading them to seek clarifications and corrections may foster a critical thinking culture and enhance the quality and accuracy of the online environment. Over time, regular engagement within the “discuss section” can contribute to the development of social norms that prioritize sharing content only after verifying its accuracy and engaging in constructive discussions about it.

While this separated section could provide a space for users to engage in critical discussions about the veracity of the content, it may also lead to further spread of misinformation if not properly moderated. Therefore, while this section has the potential to address misinformation, it would require careful implementation and moderation to be effective.

3. Method

3.1. Preliminary work

We conducted two preliminary studies before conducting the pilot study (See [Figure 3](#)). The first one was a preliminary survey to assess the most appropriate news in terms of being perceived as misinformation and the importance of the content. We presented 10 headlines in a randomized order to 101 participants (Age range: 22–60, $M = 39.3$, 68 Female, 33 Male) and asked them to indicate how serious they found the potential consequences of the presented headlines when shared on social media using a 6-point Likert scale (1 = Not at all serious and 6 = Extremely serious). The headlines included the following: Drinking bleach can cure serious illnesses, WiFi signals linked to mass psychiatric distress, NASA: A giant asteroid will hit Earth next month, The existence of new technology empowering governments to secretly track citizens, All vaccines alter DNA, Space aliens responsible for global warming, Ice cream for breakfast boosts productivity, Bananas are going extinct, Chewing gum takes seven years to digest in the stomach, and Massive cyber-attack paralyzes global financial systems. In designing our study, we intentionally avoided using overtly partisan or politically charged misinformation headlines, in order to control for confounding factors in our experiment. Highly political content can trigger strong ideological responses and group-based biases (Taber & Lodge, 2006) that might overshadow the effects of our design interventions. By focusing on misinformation topics of lower political salience, we aimed to isolate the impact of the interface design changes on user behavior, without the results being clouded by partisan predispositions. This choice improved internal validity by ensuring that any differences we observed were more likely due to the design elements rather than participants' preexisting ideological loyalties.

Subsequently, we selected both the most and least important headlines to implement in the designs. The headline perceived as most important? was “Drinking bleach can cure serious illnesses” ($M = 5.16$), while the least important was “Chewing gum takes seven years to digest in the stomach” ($M = 2.19$). The second study aimed to assess participants' comprehension of the designs. We distributed the designs to 31 people (Age range: 26–69, $M = 45.8$, 16 Female, 14 Male) to gain insights regarding clarity, readability and presentation of the designs. Four participants had misconceptions, while the remaining participants correctly understood the objectives of the designs. Based on participant feedback we then adjusted the designs. For example, we changed the name of the proposed section on Facebook from “Analyse section” to “Discuss section” in response to their comments.

3.2. Experimental study

The vignette experiment was designed using QualtricsTM (<https://www.qualtrics.com>), a web-based survey platform, and included both closed-ended and open-ended questions. There were three main parts to the questionnaire.

The first part focused on participant demographics, social media usage and attitudes toward challenging misinformation (e.g., time spent on social media per day, and frequency of their challenging). The second part introduced participants to user interface design proposals and asked about their likelihood to challenge. Participants were instructed to read each user interface design carefully and answer questions for each UID. Participants indicated how likely they would be to challenge misinformation using a 5-point Likert scale (1 = very unlikely, and 5 = very likely) and their rationale for this response with open-ended questions. The third part of the questionnaire examined participants' evaluations regarding the designs in terms of effectiveness, social acceptability and comfort.

The experimental study followed a 3 (design: Privately challenge toggle; Discuss section; usual interface) \times 2 (content importance: important; unimportant) within-subject design. Each participant was randomly presented with the three designs with two examples of news content meaning they received a total of 12 social media posts. The designs used in the study are not the ultimate versions but were used to obtain insights following the speculative design approach (Kiskola et al., 2023). Speculative design proposals can be valuable research tools when they prompt informative responses from participants (Baumer et al., 2014; Kiskola et al., 2021). We chose to mimic the social media platform Facebook for the posts due to its wide usage in the UK with 56 million Facebook users in January 2024 (Statista, 2024). In addition, 67% of internet users claimed that they encountered fake news on Facebook (Centre for International Governance et al., 2019). It is a

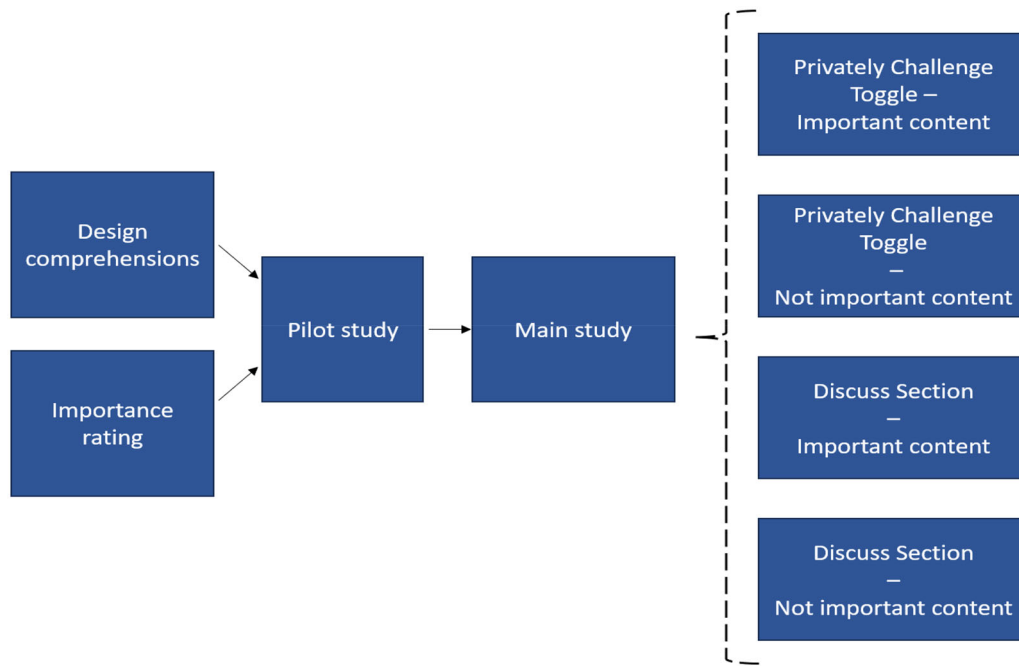


Figure 3. Study design.

semipublic space where comments and likes are visible to more people than just friends. For example, if someone comments on a friend's profile, this comment can be visible to even individuals who are not their friends, depending on privacy settings. This public visibility of social interactions is useful to observe and understand social interactions, which is a key aspect of our research. Facebook also decreased online anonymity through encouraging users to frequently update their profile pictures and even requiring its users to register their actual name (see Facebook, 2024). Therefore, it is an appropriate platform for considering social relationships.

The example posts contained the following: user picture, username, user comment, source name, and the news headline. We tried to eliminate as many confounding variables as possible with regard to posts. First, the profile picture of the source followed a generic look to keep the impact as low as possible. Second, to minimize any possible gender-specific effect, the names used were gender-neutral (Ellis Taylor and Alex Smith). Third, the study underwent face validation with six participants, including two experts, to ensure the clarity, relevance, and appropriateness of its methods and tools.

3.3. Participants

We collected data from 354 participants. We first carefully screened participants and removed those who unsuccessfully answered the two attention checks, provided gibberish or meaningless responses in open-ended questions, or completed the survey too quickly to be paying attention (less than 50% of the median duration, calculated after removing lengthy outliers). After this, 306 participants were left for further analysis (age range 20–60 years, 111 male, 194 female and 1 non-binary). Participants were recruited through ProlificTM (www.prolific.co), an established platform for online recruitment for research studies. Included participants were aged 18 years or older, UK-based English speakers, non-passive Facebook users (i.e., posted on Facebook, not just lurking), had encountered misinformation before and were using Facebook with their own identities, not anonymously.

3.4. Procedure

Bournemouth University Research Ethics Committee approved the study. Prior to data collection, a pilot test was conducted to check the usability of the questionnaire. The pilot test was active for two weeks, during which 12 participants completed the questionnaire. Following the pilot test, several changes were made

to improve the questionnaire. One question was changed due to participant confusion. To reduce the impact of fatigue and habituation (Porter et al., 2004), six design proposals were randomized. Participants were compensated in an attempt to ensure sufficient engagement with the study.

Data collection took place between 24 and 31 January 2024. Participants were invited to participate in an online survey that explored the impact of design on challenging misinformation on social media. Individuals who met the inclusion criteria were given the link to the anonymous questionnaire and asked to read the participant information sheet and consent to participate. Participants were informed that they were free to stop at any time. Participants took on average 17.6 min to complete the questionnaire. There were two attention checks within the questionnaire. The survey included eleven open-ended questions, and all participants were required to write a minimum of 30 characters per question (which are not analyzed in this paper). Among these questions, participants were prompted to explain the rationale behind their ratings and to offer recommendations regarding the strengths and weaknesses of the designs. Participants who did not provide sensible answers were excluded from the analysis. Eligible participants were compensated for their participation.

3.5. Measured variables

3.5.1. Preliminary studies

Perceived seriousness: In the first preliminary survey, participants rated the seriousness of the potential consequences of 10 news headlines when shared on social media. Ratings were given on a 6-point Likert scale (1 = Not at all serious, 6 = Extremely serious). This measure was used to identify the most and least important headlines for use in the main experimental study.

Design comprehension: In the second preliminary study, participants reviewed the proposed designs and provided open-ended responses regarding their clarity, readability, and presentation. Responses were coded for correct understanding versus misconceptions.

3.5.2. Experimental study

Perceived importance of the headline: For each headline shown in the main study, participants rated its importance on a 6-point Likert scale (1 = Not at all important, 6 = Extremely important) to verify that the predefined “important” and “unimportant” categories matched participants’ perceptions.

Likelihood to challenge misinformation: After viewing each design and content combination, participants indicated how likely they were to challenge the misinformation using a 5-point Likert scale (1 = Very unlikely, 5 = Very likely).

Willingness to challenge misinformation in general: Participants rated their overall willingness to challenge misinformation on a 6-point Likert scale (1 = Very unwilling, 6 = Very willing). For analysis, responses were grouped into two categories: willing (somewhat willing, willing, very willing) and unwilling (somewhat unwilling, unwilling, very unwilling).

Usability evaluations: Participants rated each design on three parameters: effectiveness, social acceptability, and comfort, using Likert scales. Effectiveness was defined as the perceived usefulness of the design in enabling challenges to misinformation. Social acceptability referred to the perceived appropriateness of using the design in the social media context. Comfort referred to the ease and satisfaction experienced when using the design (See “Theoretical Underpinning” section for detailed definitions).

Demographic and contextual measures: Participants reported age, gender, education level, time spent on social media per day, and frequency of challenging misinformation.

4. Data analysis

The data was analyzed using SPSS software version 28. To detect an effect of partial eta squared = 0.01 with 95% power in a two-way within-subjects ANOVA (three groups, alpha = 0.05, non-sphericity correction = 0.8), G*Power suggested we would need 252 participants. Descriptive statistics were used to describe the data and report frequencies. As our study is a within-subject design, we used a two-way repeated measures ANOVA to examine the effects of designs and perceived importance of the content on participants’ likelihood to challenge. The changes in continuous variables within groups, such as the difference between the usability parameters in proposed designs and existing designs, were analyzed

using paired-samples *t*-tests. The vignettes, study design, and dataset are available on the Open Science Framework through the link provided in the Supplementary Materials section.

5. Results

5.1. Participant demographics

The demographics of the 306 participants are summarized in Table 1. This breakdown includes age distribution, gender representation, education level and willingness to challenge.

5.2. Descriptive statistics

Prior to conducting the study, we selected the most and least important headlines to implement in the designs using a survey. In the main study, we controlled if these ratings accurately reflected how participants perceived the headlines. By asking participants to evaluate the importance of each headline on a 6-point Likert scale, we were able to confirm that the headlines' importance ratings matched participants' perceptions. ("Drinking bleach can cure serious illnesses" $M = 4.3$, $SD = 1.9$, "Chewing gum takes seven years to digest in the stomach" $M = 2.6$, $SD = 1.17$).

Means and standard deviations (SD) of likelihood to challenge were calculated for each design. Participants rated how likely they were to use the design provided to challenge the misinformation on a 6-point Likert scale (1- Very unlikely, 6- Very likely). In our study, "willingness groups" were operationalized by categorizing participants into two distinct groups based on their responses to the question of how willing they are to challenge misinformation on social media on a 6-point Likert scale. We grouped them as willing (somewhat willing, willing, very willing) and unwilling (somewhat unwilling, unwilling, very unwilling). The results are summarized in Table 2 below. These values provide insights into perceived likelihood to challenge associated with each design based on the importance of the content across willingness groups.

5.3. Differences between proposed and existing designs in terms of usability: Effectiveness, acceptability, and comfort

A paired-samples *t*-test was used to determine whether there was a statistically significant difference between the two design proposals compared to the existing design in terms of comfort, effectiveness, and acceptability (See Table 3). Data are mean \pm standard deviation unless otherwise stated. There were three outliers in each of the four pairs that were more than 1.5 box-lengths from the edge of the box in a boxplot. Inspection of their values did not reveal them to be extreme and they were kept in the analysis. The difference scores were normally distributed, as assessed by visual inspection of a Normal Q-Q Plot. As shown in Table 3, the difference was highly significant ($p < 0.001$) for both designs and all parameters. The results indicate that users' evaluation of comfort, effectiveness and acceptability are more positive for the Private Toggle and the Discuss Section than those of the existing design.

Table 1. Participant demographics.

<i>N</i>		306
Age: <i>M</i> (<i>SD</i>)	40.2 (9.9)	<i>N</i> * (%)
Age	18 – 24	13 (4.2)
	25 – 34	83 (27.1)
	35 – 44	107 (35)
	45+	103 (33.7)
	Male	194 (63.4)
Gender	Female (%)	111 (36.3)
	Non-binary (%)	1 (0.3)
	Primary (%)	52 (17)
Education	Further (%)	49 (16)
	Higher (%)	205 (67)
Willingness to challenge (%)	Unwilling	153 (50)
	Willing	153 (50)

Table 2. Mean (standard deviation), of likelihood to challenge for each design across willingness groups.

	Unimportant content	Important content	Willing		Unwilling	
			Unimportant content	Important content	Unimportant content	Important content
<i>Private toggle</i>	2.77 (1.5)	3.42 (1.84)	3.10 (1.52)	3.71 (1.9)	2.45 (1.4)	3.14 (1.72)
<i>Discuss section</i>	2.45 (1.41)	3.09 (1.76)	2.79 (1.46)	3.85 (1.69)	2.11 (1.26)	2.33 (1.47)
<i>Existing design</i>	2.48 (1.46)	3.56 (1.81)	2.98 (1.51)	4.29 (1.61)	1.99 (1.21)	2.84 (1.7)

Table 3. Paired samples *t*-test results.

The usability parameters	Paired differences				Cohen's <i>d</i>
	<i>M</i>	<i>SD</i>	<i>t</i> (305)	<i>p</i>	
<i>Comfort</i>					
Private Toggle - Existing	0.57	1.9	5.23	<0.001	0.29
Discuss Section - Existing	0.22	1.39	2.8	0.005	0.16
<i>Effectiveness</i>					
Private Toggle - Existing	0.79	1.7	8.11	<0.001	0.46
Discuss Section - Existing	0.57	1.5	6.63	<0.001	0.37
<i>Acceptability</i>					
Private Toggle - Existing	0.59	1.66	6.28	<0.001	0.35
Discuss Section - Existing	0.39	1.37	5.01	<0.001	0.29

For **comfort**, the Private Toggle showed a statistically significant positive difference compared to the existing design, with a mean difference of 0.57 (95% CI, 0.35 to 0.78), $t(305) = 5.23$, $p < 0.001$ and Cohen's $d = 0.29$, indicating a small effect size. Discuss Section, also showed a statistically significant positive difference compared to the existing design, with a mean difference of 0.22 (95% CI, 0.07 to 0.38), $t(305) = 2.80$, $p = 0.005$ and Cohen's $d = 0.16$, indicating a very small effect size. In terms of **effectiveness**, the Private Toggle demonstrated a statistically significant positive difference compared to the existing design, with a mean difference of 0.79 (95% CI, 0.60 to 0.98), $t(305) = 8.11$, $p < 0.001$ and Cohen's $d = 0.46$, representing a moderate effect size. The Discuss Section showed a statistically significant positive difference compared to the existing design, with a mean difference of 0.57 (95% CI, 0.40 to 0.74), $t(305) = 6.63$, $p < 0.001$ and Cohen's $d = 0.37$, corresponding to a small to moderate effect size. Regarding **acceptability**, the Private Toggle exhibited a statistically significant positive difference compared to the existing design, with a mean difference of 0.59 (95% CI, 0.41 to 0.78), $t(305) = 6.28$, $p < 0.001$ and Cohen's $d = 0.35$, indicating a small to moderate effect size. The Discuss Section displayed a statistically significant positive difference compared to the existing design, with a mean difference of 0.39 (95% CI, 0.24 to 0.55), $t(305) = 5.01$, $p < 0.001$. and Cohen's $d = 0.129$, reflecting a small effect size.

5.4. The interaction between design and perceived content importance in predicting the likelihood to challenge

5.4.1. Design techniques effectiveness vs news importance

A two-way repeated measures ANOVA was run to determine the effects of designs and content types on participants' likelihood to challenge (See Table 4). Design type had three levels (Private toggle, Discuss section and Existing) and the perceived importance of the content had two levels (Unimportant, Important). There were no outliers, as assessed by examination of studentized residuals for values greater than ± 3 . The likelihood to challenge for each of the designs was normally distributed, as assessed by visual inspection of a Normal Q-Q Plot. Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for the two-way interaction and the simple main effects ($p < 0.5$). Therefore, a Greenhouse-Geisser correction was reported.

Data are mean \pm standard deviation unless otherwise stated. There was a statistically significant two-way interaction between design and content suggesting a moderate effect size with a partial eta squared value of 0.030 [Greenhouse-Geisser's correction: $F(1.94, 592.9) = 9.32$, $p < 0.001$, $\epsilon = 0.972$]. Therefore, simple main effects were run.

Table 4. Repeated measures analysis of variance results.

		<i>df</i>	<i>F</i>	Partial eta squared
Design	Sphericity Assumed	2.00	7.19***	0.023
	Greenhouse-Geisser	1.80	7.19*	0.023
Content	Sphericity Assumed	1.00	133.93***	0.305
	Greenhouse-Geisser	1.00	133.93***	0.305
Design * Content	Sphericity Assumed	2.00	9.32***	0.030
	Greenhouse-Geisser	1.94	9.32***	0.030

** $p < 0.01$; *** $p < 0.001$.

Table 5. Pairwise comparisons.

					95% Confidence interval for difference ^b	
					Lower bound	Upper bound
Mean difference						
Std. Error						
Sig ^b						
<i>Important Content</i>						
Private Toggle	Existing Interface	−0.14	0.13	NS	−0.39	0.12
Discuss Section	Existing Interface	−0.474*	0.10	<0.001	−0.67	−0.27
<i>Unimportant Content</i>						
Private Toggle	Existing Interface	0.291*	0.10	0.01	0.08	0.49
Discuss Section	Existing Interface	−0.03	0.09	0.71	−0.20	0.14

Based on estimated marginal means.

*The mean difference is significant at the 0.05 level.

^bAdjustment for multiple comparisons: Bonferroni.

We compared each design proposal with the existing interface design (See Table 5 and Figure 4). For the multiple simple main effects, a Bonferroni correction was applied, and the p value was set at 0.025 for the comparisons.

For the unimportant content type, only “private toggle” ($M = 2.77$, $SE = 0.08$) showed a positive significant difference in comparison to the existing design ($M = 2.48$, $SE = 0.08$) [Greenhouse-Geisser’s correction $F(1, 305) = 7.77$ $p = 0.006$] with a mean difference of 0.291 (95% CI, 0.08 to 0.496). This indicates that people were more likely to challenge using “private toggle” relative to existing design when the content is unimportant. The difference between “discuss section” ($M = 2.45$, $SE = 0.08$) and the existing interface design ($M = 2.48$, $SE = 0.08$) was not significantly different [Greenhouse-Geisser’s correction $F(1, 305) = 0.137$ $p = 0.711$] with a mean difference of −0.033 (95% CI, −0.206 to 0.141).

For the important content, only the “discuss section” ($M = 3.09$, $SE = 0.1$) showed a negative significant difference in comparison to the existing design ($M = 3.56$, $SE = 0.1$) [Greenhouse-Geisser’s correction $F(1, 305) = 21.64$ $p < 0.001$] with a mean difference of −0.474 (95% CI, −0.674 to −0.273). This indicates that people were more likely to challenge using existing design relative to the “discuss section” when the content is important. “Private toggle” ($M = 3.42$, $SE = 0.1$) did not show a statistically significant difference [Greenhouse-Geisser’s correction $F(1, 305) = 1.06$ $p = 0.303$] with a mean difference of −0.137 (95% CI, −0.399 to 0.124).

5.5. Interaction between design and perceived content importance on the likelihood to challenge across different willingness groups

5.5.1. The influence of different designs (private Toggle, Discuss section, existing) and perceived content importance (unimportant, important) on the likelihood to challenge across different willingness groups

In our study, “willingness groups” were operationalized by categorizing participants into two groups based on their responses to the question regarding how willing they were to challenge misinformation on social media on a 6-point Likert scale. After grouping them as willing (somewhat willing, willing, very willing) and unwilling (somewhat unwilling, unwilling, very unwilling), a two-way repeated measures ANOVA was run to examine the differences in likelihood to challenge misinformation across willingness groups (See Table 6). Mauchly’s test of sphericity indicated that the assumption of sphericity had been violated for the two-way interaction and the simple main effects ($p < 0.5$). Therefore, a Greenhouse-Geisser correction was reported. There was a statistically significant two-way interaction between design and content in both unwilling and willing groups [Greenhouse-Geisser’s correction: F

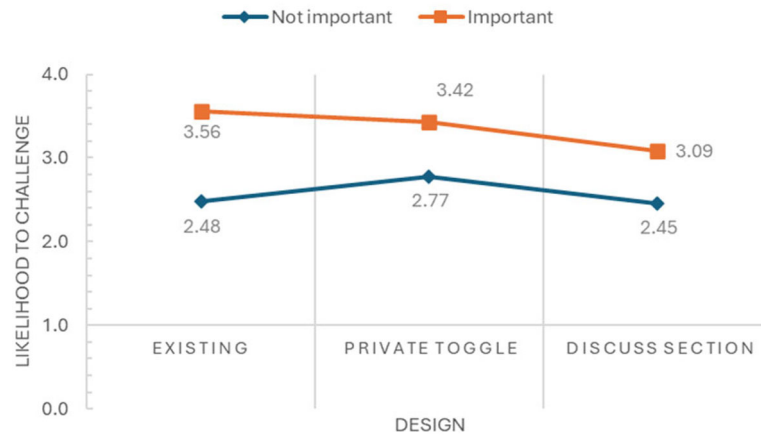


Figure 4. Likelihood to challenge for important and important content across three designs.

Table 6. Repeated measures analysis of variance results across willingness groups.

			<i>df</i>	<i>MS</i>	<i>F</i>	<i>p</i>	Partial η^2
Unwilling	Design	Sphericity Assumed	52.447	2	26.223	11.623	<0.001
		Greenhouse-Geisser	52.447	1.770	29.639	11.623	<0.001
Content		Sphericity Assumed	78.240	1	78.240	46.988	<0.001
		Greenhouse-Geisser	78.240	1.000	78.240	46.988	<0.001
Design * Content		Sphericity Assumed	16.577	2	8.289	10.539	<0.001
		Greenhouse-Geisser	16.577	1.996	8.306	10.539	<0.001
Willing	Design	Sphericity Assumed	16.113	2	8.057	2.949	00.054
		Greenhouse-Geisser	16.113	1.833	8.792	2.949	00.059
Content		Sphericity Assumed	226.510	1	226.510	91.126	<0.001
		Greenhouse-Geisser	226.510	1.000	226.510	91.126	<0.001
Design * Content		Sphericity Assumed	18.850	2	9.425	7.624	<0.001
		Greenhouse-Geisser	18.850	1.911	9.865	7.624	<0.001

Table 7. Pairwise comparisons across willingness groups.

					95% Confidence interval for difference ^a	
		Mean difference	Std. Error	Sig ^a	Lower bound	Upper bound
<i>Unwilling</i>						
<i>Unimportant Content</i>						
Private Toggle	Existing Interface	0.464*	0.14	0.001	0.19	0.74
Discuss Section	Existing Interface	0.12	0.11	0.27	-0.1	0.35
<i>Important Content</i>						
Private Toggle	Existing Interface	0.3	0.17	0.08	-0.04	0.64
Discuss Section	Existing Interface	-0.510*	0.13	<0.001	-0.77	-0.25
<i>Willing</i>						
<i>Unimportant Content</i>						
Private Toggle	Existing Interface	0.12	0.15	0.45	-0.19	0.42
Discuss Section	Existing Interface	-0.19	0.14	0.16	-0.46	0.08
<i>Important Content</i>						
Private Toggle	Existing Interface	-0.575*	0.2	0.004	-0.96	-0.19
Discuss Section	Existing Interface	-0.438*	0.15	0.005	-0.74	-0.13

Based on estimated marginal means.

*The mean difference is significant at the 0.05 level.

^aAdjustment for multiple comparisons: Bonferroni.

(1.9, 303.3) = 10.53, $p < 0.001$ $\varepsilon = 0.998$ and $F(1.9, 290.4) = 7.62$, $p < 0.001$ $\varepsilon = 0.955$ respectively]. Therefore, simple main effects were run.

We compared each design proposal to the existing interface design for the willingness groups. For the multiple simple main effects, a Bonferroni correction was applied, and the p -value set at 0.025 for the comparisons (See Table 7).

In the unwilling group, for the unimportant content, only “privately challenge toggle” ($M = 2.45$, $SE = 0.11$) showed a positive significant relative to the existing design ($M = 1.98$, $SE = 0.99$) [Greenhouse-Geisser’s correction $F(1, 152) = 11.12$ $p = 0.001$] with a mean difference of 0.464 (95% CI, 0.189 to

0.739). This indicates that people who are unwilling to challenge were more likely to challenge using “privately challenge toggle” relative to the existing design when the content was unimportant. For the important content type, the “discuss section” ($M=2.32$, $SE=0.11$) showed a negative significant difference relative to the existing design ($M=2.87$, $SE=0.13$) [Greenhouse-Geisser’s correction $F(1, 152) = 14.73$ $p < 0.001$] with a mean difference of -0.510 (95% CI, -0.772 to -0.247). This indicates that people who are unwilling to challenge were more likely to challenge using the existing design relative to the “discuss section” when the content was important.

In the willing group, for important content, “privately challenge toggle” ($M=3.71$, $SE=0.15$) and “discuss section” ($M=3.85$, $SE=0.13$) showed a negative significant difference relative to the existing design ($M=4.28$, $SE=0.13$) [Greenhouse-Geisser’s correction: $F(1, 152) = 8.59$, $p=0.004$ and $F(1, 152) = 7.99$, $p=0.005$ respectively]. This indicates that people who are willing to challenge were more likely to challenge when using the existing design relative to “privately challenge toggle” and “discuss section.”

6. Discussion

In this study, we examined two user interface design proposals that we introduced in a previous study. Our aim was to determine whether these designs would enhance participants’ likelihood to challenge misinformation. While existing research has explored the impact of user interface design (UID) on users’ critical thinking and consequently their tendency to share misinformation (Kim et al., 2019; Ozturk et al., 2015; Tanaka et al., 2013), to the best of our knowledge this is the first study to evaluate UID proposals in terms of encouraging users to challenge misinformation.

In our study, we employed two approaches to evaluate the effectiveness of different designs. First, we compared the proposed designs to the existing design on Facebook, focusing on key usability parameters including effectiveness, acceptability, and comfort. Secondly, we conducted a two-way repeated measures ANOVA to examine the potential interaction effects of design and content type on users’ likelihood to challenge misinformation. These findings contribute to our understanding of user interaction with different user interfaces and content types. They also offer practical implications for the development of more effective and user-friendly design strategies for the future.

6.1. Usability: Effectiveness, acceptability, and comfort

For each design, we assessed three parameters in terms of usability: effectiveness, acceptability, and comfort. Participants rated the proposed designs as more effective in challenging misinformation as well as more acceptable and comfortable than the existing design, which suggested that the proposed designs had the potential to encourage users to challenge misinformation. However, participants’ self-reported likelihood to challenge the content did not align with these favorable evaluations. When participants were presented with content they perceived as important or unimportant, their Willingness to challenge varied depending on whether they perceived the content as important or unimportant, regardless of the design’s overall positive ratings. These findings suggest that the perceived importance of the topic plays a significant role in influencing likelihood to engage with or challenge the content, beyond the qualities of the design itself. This result aligns with recent research showing that people are more likely to correct misinformation when they perceive its consequences as severe (Luo et al., 2024).

6.2. The interplay between design and content on the likelihood to challenge misinformation

Our findings indicate a significant two-way interaction between the design and the content type on the likelihood to challenge. Previous research suggests that demographic factors such as age, openness to experience, and perspective-taking influence the persuasiveness of interfaces in terms of challenging misinformation (Gurgun et al., 2023). Evidence from other domains shows that tailoring interventions to individual characteristics can increase their effectiveness, for example, in mobile health applications (Rivera-Romero et al., 2023) alcohol reduction strategies among university students (Bewick et al., 2008; Miller et al., 2015). Research also established that the believability of the content and its alignment with preexisting opinions influence users’ engagement (e.g., whether they read, like, comment or

share it) (Kim et al., 2019; Kim & Dennis, 2019). Our results support earlier studies that emphasize the need to consider both content characteristics and user traits, as both shape how individuals perceive and respond to interface designs (Gearhart & Zhang, 2014; Moy et al., 2001). In particular, we found that users' perception of the content's importance, significantly influenced how they evaluated the design.

Although the designs were perceived as more comfortable and acceptable by users, they did not directly influence likelihood to challenge. This discrepancy can be explained in several ways. One plausible explanation could be the social desirability bias, which suggests that research participants choose responses they believe are more socially desirable (Grimm, 2010). This concept also relates to demand effects which refer to participants' tendency to respond to questions in a way they believe the researcher desires (Grimm, 2010). Therefore, their evaluation in terms of usability may be influenced by perceived social expectations. Another explanation for this discrepancy might be that participants may have relied on superficial cues, heuristics, habits and intuition when engaging with the designs. The Elaboration Likelihood Model (Petty & Cacioppo, 1986) proposes that people can process information and form attitudes through two distinct routes: the central route and the peripheral route. While the central route refers to a high degree of thought (e.g., critically evaluating the message and considering arguments and counterarguments) the peripheral route involves a low degree of thought and is based on elements such as the appearance of the speaker or the length of the message. In our case, participants may have been influenced by peripheral cues, such as the esthetic appeal of the proposed designs, rather than engaging in critical thinking or deep cognitive processing related to challenging misinformation. This tendency to rely on peripheral cues could have led them to conform to the norm or default choices based on habit. Research suggests that knowledge does not always lead to congruent behavior. The knowledge-behavior gap describes a situation where there is incoherency and inconsistency between what people know or could know based on available information and how they actually behave in response to that knowledge (Hornik, 1989; Rimal, 2000). Factors such as self-efficacy can mediate the relationship between knowledge and behavior. Also, according to Theory of Planned Behavior (TPB) (Ajzen, 1991), while attitudes, subjective norms and perceived behavioral control influence intention, factors such as perceived barriers, past experience and situational constraints can interfere with the translation of intent to action. Overall, the discrepancy may stem from the preliminary nature of the designs.

6.2.1. *Privately challenge toggle*

We found that users demonstrated a higher likelihood of challenging misinformation with the "private toggle" we proposed (which allows them to hide their comments and names from everyone except the original poster) compared to the existing design when the content is unimportant. This result shows that, people believe that they can address misinformation with the poster in a semi-private way on social media even though they are not completely anonymous. Complete anonymity where people confront without revealing their identities to anyone, may lead to more arguments and contribute to a toxic discussion environment (Barlett, 2015; Guo & Yu, 2020; Joinson, 2007). Research found that people often refrain from confronting misinformation online due to the fear of facing negative repercussions, such as potential attacks (Gurgun et al., 2024; Gurgun et al., 2025). This fear of being attacked can negatively impact people's willingness to engage in online discussions (Urbaniak et al., 2022). For instance, the 32% of users who never or rarely share content related to political or social issues refrain from doing so primarily due to fear of being attacked (McClain, 2021). Therefore, allowing users to switch the visibility of their comments to private could help address their concerns about facing negative consequences when they challenge misinformation online. By providing a private setting, people may feel more comfortable expressing their views and confronting misinformation without the fear of public backlash. This approach could potentially encourage more open and safer environment, leading to increased engagement in challenging misinformation.

Our findings highlight a significant focus on motivating people who are hesitant or unwilling to confront misinformation. The privately challenge toggle showed a positive impact on users' likelihood to

challenge in the unwilling group. This finding underscores the potential of design interventions in empowering users to actively combat misinformation.

The result regarding the private toggle increasing the likelihood to challenge more than the existing one in the less important content provides useful insights. Research showed that people may refrain from challenging others due to pro-social attitudes, such as avoiding causing offense or preventing others from feeling embarrassed (Gurgun et al., 2024; Tandoc et al., 2020). Consequently, in less significant contexts, participants may have wanted to save the poster from embarrassment as the topic is not considered critical. In addition, people are often concerned about their image and reputation when sharing (Altay et al., 2022) and correcting misinformation (Gurgun et al., 2024). This concern may lead them to be less willing to engage on topics they find unimportant.

6.2.2. “Discuss” section

Within the context of important content, the likelihood of challenging misinformation in the existing design was higher compared to “Discuss Section” which allows users to engage in extended conversations in a separate section. Participants were reluctant to use a separate section for questioning the content, potentially influenced by the nature of social media. Tagg et al. (2017) highlighted a paradox where people acknowledge that Facebook may not be ideal for serious discussions yet still use it for political expression while avoiding controversial topics. People primarily use Facebook to maintain relationships and engage in social interactions (Nadkarni & Hofmann, 2012; Quan-Haase & Young, 2010). Therefore, social media platforms like Facebook might not be the most suitable medium for discussing topics as people would on a traditional discussion forum.

The results relating to important content suggests that participants preferred to confront important issues in a more visible way. They may approach discussions on important topics with greater caution and deliberation, preferring to engage in a more structured and normative way rather than utilizing private comments or separate discussion threads. Moreover, the social dynamics surrounding important content may influence user behavior. Users may feel a greater sense of responsibility to contribute meaningfully publicly in relation to important topics. This could lead them to challenge in the main comment thread rather than participating in separate discussions. Research showed that the third person perception (TPP) hypothesis (Davison, 1983), which suggests that people tend to believe that others are more influenced by media messages than they are (particularly when the messages are considered undesirable) plays a role in motivating users to correct misinformation (Koo et al., 2021). In line with previous work, it is possible that people are more likely to confront misinformation publicly when they perceive it as significant and influential to others.

The observed differences in likelihood to challenge between the important and unimportant news highlight the need for content-specific design considerations. Design elements that are effective for encouraging users to challenge for important news may not be equally effective for unimportant news, and vice versa. The differences identified show the importance of considering contexts, user motivations and social dynamics in designing features to encourage challenge misinformation.

6.3. Design implications

Our findings highlight that design is a continuous and iterative process and there is no one-size-fits-all solution for preventing users from skipping past misinformation content due to various social, individual and technical factors (Gurgun et al., 2024) influencing their engagement. The primary finding of this research is that design can help people, particularly who are hesitant, confront misinformation and overcome barriers. As much as design is a valuable solution to encourage people to challenge misinformation additional research is essential to improve the efficacy. One notable development in this space is the shift toward crowdsourced counter-misinformation systems, such as Community Notes, which Meta has recently announced plans to adopt (BBC, 2025). These systems enable large-scale peer-led fact-checking (Righes et al., 2023) and may be perceived as more trustworthy because they come from “people like them” rather than from platform-imposed labels (Sharevski et al., 2025). However, evidence is mixed: large-scale observational analyses found no overall reduction in engagement with misleading posts (Chuai et al., 2024), and while pairing notes with warnings can reduce perceived accuracy, sharing

intentions often remain unchanged (Sharevski et al., 2025). These outcomes reflect broader misperceptions users hold about the futility or social risks of correction, such as fears of damaging relationships or doubts about its usefulness (Gurgun et al., 2025). Starting with a co-design approach could be a valuable way to address these issues. Involving a broad and diverse group of users in their creation would help tailor corrections to the narratives, concerns, and values of different communities, making them more relevant and harder to dismiss (Lee et al., 2023). Co-design can also improve framing and credibility: peer-generated corrections often outperform top-down warnings in shifting beliefs (Zeng et al., 2024), and users tend to trust community-contributed notes more than expert fact-checker interventions (Sharevski et al., 2025). A transparent, consensus-driven process gives corrections greater legitimacy, making people more likely to accept and act on them compared to those issued by perceived partisan authorities.

Our proposed designs target a complementary mechanism, reducing the social and relational costs of challenging misinformation, for example through semi-private confrontation or dedicated discussion spaces. Future research could explore integrating these with crowdsourced annotation systems to test whether combining socially safer correction methods with peer-generated context can help translate credibility shifts into actual corrective behavior and reduced content propagation.

Several design considerations emerge from the study results.

Introduction of semi-private settings: In our study we found that users prefer a private way to confront others. However, complete privacy is not desired on social media due to its potential to encourage disinhibited behavior (Lapidot-Lefler & Barak, 2012). Therefore, implementing semi-private solutions that protect both the user and the poster could serve as an initial step. In the context of social media, providing a secure environment is essential not only for safeguarding users' financial capital, which encompasses information security such as sensitive data and potentially monetary assets, but also for protecting users' social capital (Field, 2016). Users invest time and effort to build and maintain their social capital by cultivating relationships, engaging with others, and managing their online personas. This involves creating a favorable image or impression that aligns with their desired identity and values within their social network (Marwick & Boyd, 2011; Rosenberg & Egbert, 2011). Therefore, when offering design solutions, it is essential to go beyond just functionality and consider users' needs holistically. This involves considering factors such as social relationships. Providing users with options for privacy can empower them to engage more confidently in challenging misinformation online.

Consideration of the content: One of the most important findings from this research is that there is a high interaction between design and content. When individuals evaluated the proposed designs for a single neutral content (for usability assessment), they were rated positively compared to existing designs. However, when two of the proposed designs were evaluated with specific content, users were not more likely to challenge misinformation relative to the existing interface. Content plays a significant role in how individuals evaluate designs. When the incorrect content is important, users showed a greater tendency to confront publicly (as is currently most common), using the comment box for important content. Having a separate discussion area for important content seems not to be the most appealing option for users. The negative connotations associated with the term "discussion" may deter users from engaging with content on social media, as they often seek spaces for social connection rather than critical debate. However, it is essential for users to cultivate critical thinking skills to navigate the rapid spread of misinformation effectively. Research showed that social media platforms can significantly enhance individuals' ability to think critically and respond thoughtfully to information (Daniels & Billingsley, 2014; Luo et al., 2024). Social media, in this sense, can facilitate and contribute to this process by facilitating discussions within the main comment thread, as users may feel a heightened sense of responsibility to contribute publicly on important issues (Ausat, 2023).

Refinement and habit formation: Participants rated the proposed designs more positively than the existing design in terms of usability factors like effectiveness, acceptability, and comfort when evaluating them in isolation. However, when these designs were presented with real content, they did not significantly increase participants' likelihood to challenge misinformation compared to the existing design. This suggests that usability alone is insufficient to drive behavioral change in risky actions like challenging misinformation. Drawing on habit formation theory, which emphasizes the importance of

consistent repetition of behavior in stable contexts to achieve automaticity (Lally & Gardner, 2013), expecting an immediate behavioral shift overlooks the complexity of behavior change, which requires more than usability enhancements alone. Providing certain tools, whether cognitive or environmental, may lead to consequent behavioral change (Rapp et al., 2019). For instance, according to Bandura's self-regulation theory, change occurs when monitored behavior is compared to a standard or goal (Bandura, 1991). Future HCI designers should focus more on establishing new effective habits such as incorporating prompts or integrating privacy settings.

7. Limitations and future work

This study has several limitations. First, the designs presented were not fully developed, as they were intended to be speculative (Kiskola et al., 2023) and were not claimed to represent ultimate solutions. Future research should consider fully developing such design modalities in collaboration with professional designers to gain further insights into their impact on users' behaviors.

Participant selection also poses validity concern, as the sample was recruited online and exclusively from the UK, potentially limiting the generalizability of the findings internationally. The study predominantly represents Western viewpoints. Future research would benefit from data from more collectivistic cultures and cultures that typically have different views/approaches regarding conflict management (Morris et al., 1998; Tjosvold & Sun, 2002). We also acknowledge that politically divisive misinformation remains one of the most challenging and socially consequential domains of the misinformation problem (Guess et al., 2023). Avoiding explicit political content is a limitation, as it remains unclear whether our findings apply to high-stakes contexts where misinformation engages partisan identities. Politically charged falsehoods can provoke identity-protective reasoning and even backfire effects, reinforcing rather than correcting false beliefs (Nyhan & Reifler, 2010). In such settings, group loyalty may strongly shape willingness to correct, meaning our interventions could be received differently when content aligns with users' political affiliations or social norms.

Moreover, the study employed predetermined content types to assess users' likelihood to challenge misinformation, potentially overlooking the variety of misinformation encountered on social media platforms. Research showed that, the believability of the content and its alignment with preexisting opinions influence users' engagement (Kim et al., 2019; Kim & Dennis, 2019). Future research could explore the impact of the designs with different content types.

Additionally, while the study focused on Facebook user interface designs, future research should examine whether these design concepts are applicable and effective on other social media platforms, many of which differ significantly to Facebook in terms of the way in which content is presented and their user populations. Furthermore, the study measured participants' intentions (e.g., willingness to challenge misinformation) rather than observing their actual behaviors. While intention is a useful predictor, behavioral science consistently shows an intention-behavior gap, with intentions explaining only a modest proportion of real-world actions (Conner & Norman, 2022). Intentions may not fully capture how individuals would respond in live social media environments or how they might adapt their correction strategies over time. However, intention measures are a valuable first step in early-stage or speculative design research, providing insight into potential user acceptance before investing in fully functional prototypes. As design is iterative, these findings can guide follow-up studies using observational or longitudinal methods to understand real-world behaviors and adaptation.

8. Conclusion

In conclusion, this study explored the potential of user interface design interventions to encourage users to challenge misinformation on social media platforms. By introducing two novel features, "private toggle" and "discuss section" the study highlighted the complex interplay between design, content importance, and user willingness to engage. While the proposed designs were perceived as more usable than existing alternatives, their effectiveness varied depending on the perceived importance of the content and users' willingness to challenge. These findings underscore the necessity of tailoring interventions to user needs and content characteristics. Future research should refine these designs,

expand to other platforms and cultural contexts, and examine real-world application scenarios to further enhance our understanding of how social media can empower users to combat misinformation.

Author contributions

CRedit: **Selin Gurgun**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft; **Emily Arden-Close**: Conceptualization, Investigation, Methodology, Project administration, Supervision, Writing – review & editing; **Keith Phalp**: Conceptualization, Methodology, Supervision, Writing – review & editing; **Raian Ali**: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Supervision, Writing – review & editing.

Disclosure statement

No potential competing interest was reported by the author(s).

Funding

This research was made possible by an NPRP-14-Cluster Grant # NPRP 14C-0916-210015 from the Qatar National Research Fund (a member of the Qatar Foundation). The findings herein reflect the work and are solely the authors' responsibility. Open Access fund has been provided by the Qatar National Library.

Supplementary material

The vignettes, study design, and dataset are available on the Open Science Framework through the link: <https://osf.io/s9ega/>

ORCID

Raian Ali  <http://orcid.org/0000-0002-5285-7829>

References

- Ahmed, A. A. A., Aljabouh, A., Donepudi, P. K., & Choi, M. S. (2021). Detecting fake news using machine learning: A systematic literature review. *Psychology and Education*, 58(1), 1932–1939. <https://doi.org/10.48550/arXiv.2102.04458>
- Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining*, 13(1), 30. <https://doi.org/10.1007/s13278-023-01028-5>
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Altay, S., Hacquin, A.-S., & Mercier, H. (2022). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 24(6), 1461444820969893. <https://doi.org/10.1177/1461444820969893>
- Ausat, A. M. A. (2023). The role of social media in shaping public opinion and its influence on economic decisions. *Technology and Society Perspectives (TACIT)*, 1(1), 35–44.
- Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50(2), 248–287. [https://doi.org/10.1016/0749-5978\(91\)90022-L](https://doi.org/10.1016/0749-5978(91)90022-L)
- Barlett, C. P. (2015). Anonymously hurting others online: The effect of anonymity on cyberbullying frequency. *Psychology of Popular Media Culture*, 4(2), 70–79. <https://doi.org/10.1037/a0034335>
- Baughan, A., Petelka, J., Yoo, C. J., Lo, J., Wang, S., Paramasivam, A., Zhou, A., & Hiniker, A. (2021). Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–22. <https://doi.org/10.1145/344923>
- Baumer, E. P. S., Khovanskaya, V., Matthews, M., Reynolds, L., Schwanda Sosik, V., & Gay, G. (2014). Reviewing reflection: On the use of reflection in interactive system design. In *DIS '14: Proceedings of the 2014 Conference on Designing Interactive Systems*. Association for Computing Machinery.
- Bewick, B. M., Trusler, K., Mulhern, B., Barkham, M., & Hill, A. J. (2008). The feasibility and effectiveness of a web-based personalised feedback and social norms alcohol intervention in UK university students: A randomised control trial. *Addictive Behaviors*, 33(9), 1192–1198. <https://doi.org/10.1016/j.addbeh.2008.05.002>
- Bode, L. (2019). User correction as a tool in the battle against social media misinformation. *Georgetown Law Technology Review*, 4(2), 367.

- Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4), 619–638. <https://doi.org/10.1111/jcom.12166>
- Bode, L., & Vraga, E. K. (2018). See something, say something: correction of global health misinformation on social media. *Health Communication*, 33(9), 1131–1140. <https://doi.org/10.1080/10410236.2017.1331312>
- Bode, L., & Vraga, E. K. (2021a). Correction experiences on social media during COVID-19. *Social Media + Society*, 7(2). <https://doi.org/10.1177/20563051211008829>
- Bode, L., & Vraga, E. K. (2021b). People-powered correction: Fixing misinformation on social media. In *The Routledge companion to media disinformation and populism* (pp. 498–506). Routledge.
- Broniatowski, D. A., Simons, J. R., Gu, J., Jamison, A. M., & Abroms, L. C. (2023). The efficacy of Facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. *Science Advances*, 9(37), eadh2132. <https://doi.org/10.1126/sciadv.adh2132>
- Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Centre for International Governance Innovation, & Ipsos (2019). *CIGI-Ipsos Global Survey on Internet Security and Trust*.
- Chadwick, A., & Vaccari, C. (2019). *News sharing on UK social media: Misinformation, disinformation, and correction*. Online Civic Culture Centre at Loughborough University.
- Chadwick, A., Vaccari, C., & Hall, N.-A. (2022). *Covid vaccines and online personal messaging: The challenge of challenging everyday misinformation*. Online Civic Culture Centre at Loughborough University.
- Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). Did the roll-out of community notes reduce engagement with misinformation on X/Twitter? *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2), 1–52. <https://doi.org/10.1145/3686967>
- Conner, M., & Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology*, 13, 923464. <https://doi.org/10.3389/fpsyg.2022.923464>
- Daniels, K. N., & Billingsley, K. Y. (2014). “Facebook”—it’s not just for pictures anymore: The impact of social media on cooperative learning. *Journal of Educational Technology*, 11(3), 34–44.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>
- Davison, W. P. (1983). The third-person effect in communication. *Public Opinion Quarterly*, 47(1), 1–15. <https://doi.org/10.1086/268763>
- de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., & Mattos, D. M. F. (2021). Identifying fake news on social networks based on natural language processing: Trends and challenges. *Information*, 12(1), 38. <https://doi.org/10.3390/info12010038>
- Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining” gamification. In *MindTrek '11: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*. Association for Computing Machinery.
- Elaheebocus, S. M. R. A., Weal, M., Morrison, L., & Yardley, L. (2018). Peer-based social media features in behavior change interventions: Systematic review. *Journal of Medical Internet Research*, 20(2), e8342. <https://doi.org/10.2196/jmir.8342>
- Facebook (2024). *Names allowed on Facebook*. <https://www.facebook.com/help/112146705538576>
- Field, J. (2016). *Social capital*. Routledge.
- Figl, K., Kießling, S., & Remus, U. (2023). Do symbol and device matter? The effects of symbol choice of fake news flags and device on human interaction with fake news on social media platforms. *Computers in Human Behavior*, 144, 107704. <https://doi.org/10.1016/j.chb.2023.107704>
- Figueira, Á., & Oliveira, L. (2017). The current state of fake news: Challenges and opportunities. *Procedia Computer Science*, 121, 817–825. <https://doi.org/10.1016/j.procs.2017.11.106>
- Fogg, B. J. (2002). Persuasive technology: Using computers to change what we think and do. *Ubiquity*, 2002(December), 2. <https://doi.org/10.1145/764008.763957>
- Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? *CHI '00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Fuller, J. B., Barnett, T., Hester, K., Relyea, C., & Frey, L. (2007). An exploratory examination of voice behavior from an impression management perspective. *Journal of Managerial Issues*, 19(1), 134–151.
- Gaozhao, D. (2021). Flagging fake news on social media: An experimental study of media consumers’ identification of fake news. *Government Information Quarterly*, 38(3), 101591. <https://doi.org/10.1016/j.giq.2021.101591>
- Gearhart, S., & Zhang, W. (2014). Gay bullying and online opinion expression. *Social Science Computer Review*, 32(1), 18–36. <https://doi.org/10.1177/0894439313504261>
- Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, 38(1), 196–221. <https://doi.org/10.1080/07421222.2021.1870389>

- Grimm, P. (2010). Social desirability bias. In *Wiley international encyclopedia of marketing*. Wiley.
- Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., ... Tucker, J. A. (2023). Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science (New York, N.Y.)*, 381(6656), 404–408. <https://doi.org/10.1126/science.add8424>
- Guo, K. H., & Yu, X. (2020). The anonymous online self: Toward an understanding of the tension between discipline and online anonymity. *Information Systems Journal*, 30(1), 48–69. <https://doi.org/10.1111/isj.12242>
- Gurgun, S., Arden-Close, E., McAlaney, J., Phalp, K., & Ali, R. (2023). *Can We Re-design Social Media to Persuade People to Challenge Misinformation? An Exploratory Study*. PERSUASIVE 2023: Persuasive Technology. Lecture Notes in Computer Science.
- Gurgun, S., Arden-Close, E., Phalp, K., & Ali, R. (2023). Online silence: Why do people not challenge others when posting misinformation? *Internet Research*, 33(5), 1928–1948. <https://doi.org/10.1108/INTR-06-2022-0407>
- Gurgun, S., Arden-Close, E., Phalp, K., & Ali, R. (2024). Motivated by Design: A Co-Design Study to Promote Challenging Misinformation on Social Media. *Human Behavior and Emerging Technologies*, 2024(1). <https://doi.org/10.1155/2024/5595339>
- Gurgun, S., Arden-Close, E., Phalp, K., & Ali, R. (2025). *User Correction of Misinformation on Social Media: Perceived and Actual Social Norms*. Research Challenges in Information Science. RCIS 2025. Lecture Notes in Business Information Processing. Springer.
- Gurgun, S., Cemiloglu, D., Arden-Close, E., Phalp, K., Nakov, P., & Ali, R. (2024). Why Do We Not Stand Up to Misinformation? Factors Influencing the Likelihood of Challenging Misinformation on Social Media and the Role of Demographics. *Technology in Society*, 76, 102444. <https://doi.org/10.1016/j.techsoc.2023.102444>
- Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*, 64(2), 79–102. <https://doi.org/10.1016/j.ijhcs.2005.06.002>
- Hornik, R. (1989). The knowledge-behavior gap in public information campaigns: A development communication view. In *Information campaigns: Balancing social values and social change* (pp. 113–138). SAGE Publications, Inc.
- Ibrishimova, M. D., & Li, K. F. (2020). A machine learning approach to fake news detection using knowledge verification and natural language processing. In L. Barolli, H. Nishino, & H. Miwa (Eds.), *Advances in Intelligent Networking and Collaborative Systems. INCoS 2019*. Advances in Intelligent Systems and Computing. Springer.
- International Organization for Standardization (ISO) (1999). ISO 13407:1999(en) Human-centred design processes for interactive systems. Retrieved February 2, 2024 from <https://www.iso.org/obp/ui/#iso:std:iso:13407:ed-1:v1:en>
- International Organization for Standardization (ISO) (2011). ISO/IEC 25010:2011(en) Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models. <https://www.iso.org/obp/ui/#iso:std:iso-iec:25010:ed-1:v1:en>
- Jing, T. W., & Murugesan, R. K. (2019). A theoretical framework to build trust and prevent fake news in social media using blockchain. In F. Saeed, N. Gazem, F. Mohammed, & A. Busalim (Eds.), *Recent Trends in Data Science and Soft Computing. IRICT 2018. Advances in Intelligent Systems and Computing*. Springer.
- Joinson, A. N. (2007). Chapter 4 - Disinhibition and the Internet. In J. Gackenbach (Ed.), *Psychology and the Internet* (2nd ed., pp. 75–92). Academic Press.
- Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS Quarterly*, 43(3), 1025–1039. <https://doi.org/10.25300/MISQ/2019/15188>
- Kim, A., Moravec, P. L., & Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, 36(3), 931–968. <https://doi.org/10.1080/07421222.2019.1628921>
- Kiskola, J., Olsson, T., Rantasila, A., Syrjämäki, A. H., Ilves, M., Isokoski, P., & Surakka, V. (2023). User-centred quality of UI interventions aiming to influence online news commenting behaviour. *Behaviour & Information Technology*, 42(12), 2060–2092. <https://doi.org/10.1080/0144929X.2022.2108723>
- Kiskola, J., Olsson, T., Väättäjä, H., Syrjämäki, A. H., Rantasila, A., Isokoski, P., Ilves, M., & Surakka, V. (2021). Applying Critical Voice in Design of User Interfaces for Supporting Self-Reflection and Emotion Regulation in Online News Commenting. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Kivijärvi, H., & Pärnänen, K. (2023). Instrumental usability and effective user experience: Interwoven drivers and outcomes of Human-Computer interaction. *International Journal of Human-Computer Interaction*, 39(1), 34–51. <https://doi.org/10.1080/10447318.2021.2016236>
- Koelle, M., Ananthanarayan, S., & Boll, S. (2020). Social acceptability in HCI: A survey of methods, measures, and design strategies. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Koelle, M., Olsson, T., Mitchell, R., Williamson, J., & Boll, S. (2019). What is (un) acceptable? Thoughts on social acceptability in HCI research. *Interactions*, 26(3), 36–40. <https://doi.org/10.1145/3319073>

- Koo, A. Z.-X., Su, M.-H., Lee, S., Ahn, S.-Y., & Rojas, H. (2021). What Motivates People to Correct Misinformation? Examining the Effects of Third-person Perceptions and Perceived Norms. *Journal of Broadcasting & Electronic Media*, 65(1), 111–134. <https://doi.org/10.1080/08838151.2021.1903896>
- Kriplean, T., Morgan, J., Freelon, D., Borning, A., & Bennett, L. (2012). Supporting reflective public thought with considerit. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. ACM.
- Lally, P., & Gardner, B. (2013). Promoting habit formation. *Health Psychology Review*, 7(sup1), S137–S158. <https://doi.org/10.1080/17437199.2011.603640>
- Langrial, S., Lehto, T., Oinas-Kukkonen, H., Harjumaa, M., & Karppinen, P. (2012). *Native mobile applications for personal well-being: A persuasive systems design evaluation* [Paper presentation]. 16th Pacific Asia Conference on Information Systems (PACIS 2012).
- Lapidot-Lefler, N., & Barak, A. (2012). Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior*, 28(2), 434–443. <https://doi.org/10.1016/j.chb.2011.10.014>
- Lawson, M. A., Anand, S., & Kakkar, H. (2023). Tribalism and tribulations: The social costs of not sharing fake news. *Journal of Experimental Psychology: General*, 152(3), 611–631. <https://doi.org/10.1037/xge0001374>
- Lee, A. Y., Moore, R. C., & Hancock, J. T. (2023). Designing misinformation interventions for all: Perspectives from AAPI, Black, Latino, and Native American community leaders on misinformation educational efforts. *Harvard Kennedy School Misinformation Review*, 4(1). <https://doi.org/10.37016/mr-2020-111>
- Lockton, D., Harrison, D., & Stanton, N. (2008). Making the user more efficient: Design for sustainable behaviour. *International Journal of Sustainable Engineering*, 1(1), 3–8. <https://doi.org/10.1080/19397030802131068>
- Luo, C., Zhu, Y., & Chen, A. (2024). What motivates people to counter misinformation on social media? Unpacking the roles of perceived consequences, third-person perception and social media use. *Online Information Review*, 48(1), 105–122. <https://doi.org/10.1108/OIR-09-2022-0507>
- Ma, X., Chen, G., & Xiao, J. (2010). Analysis of an online health social network. In *IHI '10: Proceedings of the 1st ACM International Health Informatics Symposium*. ACM.
- Marwick, A. E., & Boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133. <https://doi.org/10.1177/1461444810365313>
- McAndrew, F. T., & Jeong, H. S. (2012). Who does what on Facebook? Age, sex, and relationship status as predictors of Facebook use. *Computers in Human Behavior*, 28(6), 2359–2365. <https://doi.org/10.1016/j.chb.2012.07.007>
- McClain, C. (2021). *American Trend Panel*. <https://www.pewresearch.org/fact-tank/2021/05/04/70-of-u-s-social-media-users-never-or-rarely-post-or-share-about-political-social-issues/>
- Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, 12(2), 165–183. <https://doi.org/10.1002/poi3.214>
- Miller, M. B., Meier, E., Lombardi, N., & Leffingwell, T. R. (2015). Theories of behaviour change and personalised feedback interventions for college student drinking. *Addiction Research & Theory*, 23(4), 322–335. <https://doi.org/10.3109/16066359.2014.1001840>
- Montero, C. S., Alexander, J., Marshall, M. T., & Subramanian, S. (2010). Would you do that? Understanding social acceptance of gestural interfaces. In *MobileHCI '10: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*. Association for Computing Machinery.
- Morris, M. W., Williams, K. Y., Leung, K., Larrick, R., Mendoza, M. T., Bhatnagar, D., Li, J., Kondo, M., Luo, J.-L., & Hu, J.-C. (1998). Conflict management style: Accounting for cross-national differences. *Journal of International Business Studies*, 29(4), 729–747. <https://doi.org/10.1057/palgrave.jibs.8490050>
- Moy, P., Domke, D., & Stamm, K. (2001). The spiral of silence and public opinion on affirmative action. *Journalism & Mass Communication Quarterly*, 78(1), 7–25. <https://doi.org/10.1177/107769900107800102>
- Nadkarni, A., & Hofmann, S. G. (2012). Why do people use Facebook? *Personality and Individual Differences*, 52(3), 243–249. <https://doi.org/10.1016/j.paid.2011.11.007>
- Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., & Mansour, W. (2021). Overview of the clef-2021 checkthat! Lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In K. S. Candan et al. (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science*. Springer Science + Business Media Deutschland GmbH.
- Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 303–330. <https://doi.org/10.1007/s11109-010-9112-2>
- Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, 24(1), 28. <https://doi.org/10.17705/1CAIS.02428>
- Ozturk, P., Li, H., & Sakamoto, Y. (2015). *Combating rumor spread on social media: The effectiveness of refutation and warning* [Paper presentation]. 2015 48th Hawaii International Conference on System Sciences (HICSS 2015).
- Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion*. Springer.
- Porter, S. R., Whitcomb, M. E., & Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121), 63–73. <https://doi.org/10.1002/ir.101>
- Quan-Haase, A., & Young, A. L. (2010). Uses and gratifications of social media: A comparison of Facebook and instant messaging. *Bulletin of Science, Technology & Society*, 30(5), 350–361. <https://doi.org/10.1177/0270467610380009>

- Rapp, A., Tirassa, M., & Tirabeni, L. (2019). Rethinking technologies for behavior change: A view from the inside of Human change. *ACM Transactions on Computer-Human Interaction*, 26(4), 1–30. <https://doi.org/10.1145/3318142>
- Righes, L., Saeed, M., Demartini, G., & Papotti, P. (2023). The Community Notes Observatory: Can Crowdsourced Fact-Checking be Trusted in Practice? In *The Web Conference (WWW '23) Companion: Companion Proceedings of the ACM Web Conference 2023*.
- Rimal, R. N. (2000). Closing the knowledge-behavior gap in health promotion: The mediating role of self-efficacy. *Health Communication*, 12(3), 219–237. https://doi.org/10.1207/S15327027HC1203_01
- Rivera-Romero, O., Gabarron, E., Roperio, J., & Denecke, K. (2023). Designing personalised mHealth solutions: An overview. *Journal of Biomedical Informatics*, 146(C), 104500. <https://doi.org/10.1016/j.jbi.2023.104500>
- Rosenberg, J., & Egbert, N. (2011). Online impression management: Personality traits and concerns for secondary goals as predictors of self-presentation tactics on Facebook. *Journal of Computer-Mediated Communication*, 17(1), 1–18. <https://doi.org/10.1111/j.1083-6101.2011.01560.x>
- Rosenfeld, P., Giacalone, R. A., & Riordan, C. A. (2001). *Impression management: Building and enhancing reputations at work*. Thomson Learning.
- Seering, J., Fang, T., Damasco, L., Chen, M. C., Sun, L., & Kaufman, G. (2019). Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.
- Sharevski, F., Distler, V., & Alt, F. (2025). Helps me Take the Post With a Grain of Salt:” Soft Moderation Effects on Accuracy Perceptions and Sharing Intentions of Inauthentic Political Content on X. *SEC '25: Proceedings of the 34th USENIX Conference on Security Symposium*. USENIX Association.
- Statista (2024). Number of Facebook users in the United Kingdom from September 2018 to January 2024. Statista. <https://www.statista.com/statistics/1012080/uk-monthly-numbers-facebook-users/>
- Sukumaran, A., Vezich, S., McHugh, M., & Nass, C. (2011). Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Suler, J. (2004). The Online Disinhibition Effect [Periodical]. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, 7(3), 321–326. <https://doi.org/10.1089/1094931041291295>
- Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, 50(3), 755–769. <https://doi.org/10.1111/j.1540-5907.2006.00214.x>
- Tagg, C., Seargeant, P., & Brown, A. A. (2017). *Taking offence on social media: Conviviality and communication on Facebook*. Palgrave Macmillan, imprint published by Springer Nature.
- Tajfel, H., & Turner, J. C. (2004). *The social identity theory of intergroup behavior (Political psychology)* (pp. 276–293). Psychology Press.
- Tanaka, Y., Sakamoto, Y., & Matsuka, T. (2013). *Toward a Social-Technological System that Inactivates False Rumors through the Critical Thinking of Crowds* [Paper presentation]. 48th Hawaii International Conference on System Sciences (HICSS 2015).
- Tandoc, E. C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: How social media users respond to fake news and why. *Journalism*, 21(3), 381–398. <https://doi.org/10.1177/1464884919868325>
- Taylor, S. H., DiFranzo, D., Choi, Y. H., Sannon, S., & Bazarova, N. N. (2019). Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–26.
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Tjosvold, D., & Sun, H. F. (2002). Understanding conflict avoidance: Relationship, motivations, actions, and consequences. *International Journal of Conflict Management*, 13(2), 142–164. <https://doi.org/10.1108/eb022872>
- Tully, M., Bode, L., & Vraga, E. K. (2020). Mobilizing Users: Does Exposure to Misinformation and Its Correction Affect Users’ Responses to a Health Misinformation Post? *Social Media + Society*, 6(4). <https://doi.org/10.1177/2056305120978377>
- Urbaniak, R., Ptaszyński, M., Tempska, P., Leliwa, G., Brochocki, M., & Wroczynski, M. (2022). Personal attacks decrease user activity in social networking platforms. *Computers in Human Behavior*, 126, 106972. <https://doi.org/10.1016/j.chb.2021.106972>
- Venkatesh, V. (2000). Determinants of perceived ease of use: Integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research*, 11(4), 342–365. <https://doi.org/10.1287/isre.11.4.342.11872>
- Vicol, D. O. (2020). *Who is most likely to believe and to share misinformation?* <https://fullfact.org/media/uploads/who-believes-shares-misinformation.pdf>
- Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, 39(5), 621–645. <https://doi.org/10.1177/1075547017731776>
- Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, 21(10), 1337–1353. <https://doi.org/10.1080/1369118X.2017.1313883>

- Vraga, E. K., & Bode, L. (2020). Correction as a Solution for Health Misinformation on Social Media. *American Journal of Public Health*, 110(S3), S278–S280. <https://doi.org/10.2105/AJPH.2020.305916>
- Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, 85(3), 423–441. <https://doi.org/10.1080/03637751.2018.1467564>
- Whitworth, B. (2011). The social requirements of technical systems. (*Virtual Communities: Concepts, Methodologies, Tools and Applications* (pp. 1461–1481). IGI Global.
- Widyasari, Y. D. L., Nugroho, L. E., & Permanasari, A. E. (2019). Persuasive technology for enhanced learning behavior in higher education. *International Journal of Educational Technology in Higher Education*, 16(1), 1–16. <https://doi.org/10.1186/s41239-019-0142-5>
- Wu, S., Lin, T.-C., & Shih, J.-F. (2017). Examining the antecedents of online disinhibition. *Information Technology & People*, 30(1), 189–209. <https://doi.org/10.1108/ITP-07-2015-0167>
- Zeng, H.-K., Lo, S.-Y., & Li, S.-C. S. (2024). Credibility of misinformation source moderates the effectiveness of corrective messages on social media. *Public Understanding of Science (Bristol, England)*, 33(5), 587–603. <https://doi.org/10.1177/09636625231215979>

About the authors

Selin Gurgun is a UX Researcher with a PhD in Human-Computer Interaction from Bournemouth University. She specializes in mixed-method UX research, exploring how people navigate digital platforms and make decisions online. Her work bridges research and practice, turning behavioral insights into meaningful, user-centred digital experiences.

Emily Arden-Close is a Health Psychologist and Principal Academic in Psychology at Bournemouth University, UK. She received an MSc in Health Psychology and a PhD in Health Psychology Research and Professional Practice from the University of Southampton, UK. Her research focuses on developing and evaluating digital interventions to improve wellbeing.

Keith Phalp is a Professor of Computing at Bournemouth University, UK. His research interests include software engineering, particularly the early phases of software projects and the relationship between business and software models, model-driven development, applications of AI, and social computing, including digital addiction and online gambling.

Raian Ali is a Professor at Hamad Bin Khalifa University (HBKU), Qatar. His research focuses on technology and human behavior, including digital addiction, responsible technology use, and the impact of technology design on human well-being.