Int. J. Human–Computer Studies 188 (2024) 103276

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Peer-supplied credibility labels as an online misinformation intervention

Saumya Pareek *, Jorge Goncalves

School of Computing and Information Systems, The University of Melbourne, Parkville, Melbourne, 3010, Victoria, Australia

ARTICLE INFO

Keywords:
Misinformation
Credibility assessment
Fake news
Social influence
Social media
Political homophily
Backfire effect

ABSTRACT

Misinformation is rampant on social media, and existing platform-supplied interventions offer limited effectiveness. In this study, we examine the effectiveness of credibility labels that dispute the accuracy of information when they are supplied by one's peers at different levels of relationship closeness and political agreement. We investigate four variants of these labels using a 2 (*strong* vs. *weak* tie strength) x 2 (*high* vs. *low* political agreement) between-subjects factorial design. We find that credibility disputes raised by one's co-partisans (peers with similar political beliefs) significantly reduced belief in misinformation, irrespective of one's relationship closeness with the peer. Our findings also reveal that in contrast to prior literature, a peer's knowledgeability may be more potent than trustworthiness in causing belief change, and that trust can sometimes manifest even in the credibility judgement of distant peers, when perceived to have expertise or a fact-checking tendency. We further highlight the dual nature of these credibility labels, discussing scenarios in which disputes by hyper-partisan members of the opposite party can enforce belief in misinformation. We conclude by discussing how peer-supplied credibility disputes can benefit social media, especially echo chambers with high political homophily, where disputes by a co-partisan may be met with less resistance and persuade significantly reduced belief in fake news.

1. Introduction

Misinformation is increasingly plaguing social media platforms (Karlova and Fisher, 2013). With over 4.7 billion social media users worldwide having unhindered access to creating, consuming, and disseminating information (Kemp, 2022), our media landscape has transitioned from one where a few select outlets serve as agents of information to one where the creation and diffusion of information is dictated by individual beliefs and partisan predilections (Lazer et al., 2018). This has fuelled the fabrication of misinformation from countless non-credible sources (Lewandowsky et al., 2012), making social media the primary platforms where online misinformation flourishes, and diffuses "farther, faster, deeper, and more broadly" than the truth (Vosoughi et al., 2018). Consequently, there is a growing fear of such platforms morphing into "weaponised propaganda machines" (Revell, 2017; Wardle and Derakhshan, 2018).

This rampant diffusion of inaccurate information disguised as the truth endangers social media users in diverse ways, from causing confusion and anxiety in emergency situations (Budak et al., 2011; Gupta et al., 2013), to fuelling violence (Haque et al., 2020) and promoting criminal activities (Chen and Sin, 2013), to sowing seeds of doubt during global health crises (Morozov, 2009; Wardle and Singerman, 2021). With more and more people using social media as their primary source of news discovery (Newman et al., 2021), misinformation has exacerbated public polarisation (Au et al., 2022), while also worsening the already fractionated political landscape by motivating grave distrust in politics and journalism (Balmas, 2014), and threatening the very foundation of democracy by influencing political outcomes (Allcott and Gentzkow, 2017).

In light of this, social networking sites have deployed algorithmic interventions like credibility labels to help users gauge the veracity of information they encounter. However, two fundamental problems with these approaches persist. First, misperceptions are *sticky* and difficult to revert (Thorson, 2016) — a resistance further heightened when corrective efforts challenge one's preexisting attitudes (Nyhan and Reifler, 2010; Ecker and Ang, 2019). Second, users often ignore platform-applied fact-checks due to the institutional distrust caused by these platforms' revenue-driven and politically-biased intentions (Saltz et al., 2021b), as the misinformation they appear to seek to combat is also what generates the most engagement (Sharma et al., 2017).

Belief formation literature argues that in addition to expertise and trustworthiness, *goodwill* – whether the source of information has the receiver's best interest at heart – is also a heuristic used to evaluate source credibility (Lupia and McCubbins, 1998). Since the persuasiveness of a source increases with its perceived credibility, *strong ties* – peers with whom one shares a close relationship – may help individuals

* Corresponding author.
E-mail address: saumya.pareek@student.unimelb.edu.au (S. Pareek).

https://doi.org/10.1016/j.ijhcs.2024.103276
Received 3 October 2023; Received in revised form 15 April 2024; Accepted 16 April 2024
Available online 17 April 2024

overcome motivated reasoning and be more willing to accept attitude-incongruent credibility disputes. Moreover, since messages shared by one's *in-group* members, like those from the same political party, are systematically processed (Chaiken, 1987), they carry a greater capacity for persuasion (Mackie et al., 1990), and may serve as an effective intervention for online misinformation. While the influence of source political slant has been examined in relation to *manifesting* belief in misinformation, there has not yet been a systematic investigation of whether it can promote *reduced* belief in misinformation.

Moreover, although debunking efforts may reduce belief in fake news, they may not necessarily compel individuals to share posts with a credibility label that disputes their accuracy (Pasquetto et al., 2022). Thus, examining how credibility disputes by one's peers can influence the re-sharing of misinformation is also critical to evaluate their overall de-biasing potential. Furthermore, when corrections challenge an individual's pre-existing belief system, they can strengthen misperceptions instead of eliminating them, a phenomenon termed as the *boomerang* or *backfire* effect (Byrne and Hart, 2009). It is crucial to identify whether any corrective interventions can backfire, as current literature provides mixed evidence of this phenomenon (Weeks, 2015; Wood and Porter, 2019).

This study aims to bridge the aforementioned gaps, and examines the effectiveness of credibility labels, supplied by one's social media peers instead of the platform, in reducing belief in misinformation. In particular, we investigate the following research questions:

- **RQ1:** How is the influence of peer-supplied credibility labels impacted by the relationship closeness and level of political agreement between their receiver and their source?
- **RQ2:** How do peer-supplied credibility labels influence users' inclination to re-share social media posts containing misinformation?
- **RQ3:** Under what conditions can peer-supplied credibility labels backfire and further entrench belief in misinformation?

To answer these research questions, we conducted a study where we first showed participants fake news headlines which reinforced their partisan beliefs and biases. We then presented credibility disputes supplied by their peers, which may have been difficult to accept because they contradict prior beliefs. Our experimental design manipulates the *tie strength* between the receiver (the participant) and source (the peer) of the credibility message (i.e. whether the relationship is strong, like with a family member, or weak, like with an acquaintance), and the *political agreement* between them (i.e. whether they both identify as supporting the same political party or not). Next, we measured participants' belief in the news headlines, how confident they were in their veracity evaluation, and their likelihood of sharing it on social media, before and after the addition of a credibility label supplied by their peers.

We found a main effect of political agreement on change in belief **(RQ1)**, which suggests that credibility disputes by members of one's *in-group* on social media can have a significant impact on reducing belief in even worldview–congruent misinformation. While the influence of tie strength was not statistically significant **(RQ1)**, we uncovered interesting insights about metrics such as the expertise and knowledgeability of a peer, which may be more potent than trustworthiness in engendering trust in a peer's veracity judgement, leading to higher perceived credibility. Further, our findings revealed no significant difference in the propensity to share news headlines before and after the application of credibility labels **(RQ2)**, suggesting that a decrease in belief in misinformation does not necessarily lead to an increased likelihood of sharing the disputed post (with an added credibility label) with one's social media peers. Finally, we found evidence of the backfire effect — when the source of the credibility label was perceived as hyper-partisan and from a different political party, or lacked expertise, beliefs in misinformation were bolstered rather than being unchanged or reduced **(RQ3)**.

There are four major contributions of this study. First, we provide evidence on the effectiveness of credibility labels supplied by one's online peers, suggesting that social media platforms and the social structures they support can be effectively leveraged to reduce misperceptions. Second, we indicate that belief in attitude-affirming misinformation, despite being more challenging to influence, can be reduced when credibility is disputed by one's co-partisans. This has important implications for social media platforms, especially echo chambers, where credibility disputes raised by one's co-partisans may be encountered with less resistance and facilitate belief change. Third, we emphasise that metrics which govern credibility assessment apart from trustworthiness, like source expertise and knowledgeability, should not be seen as external to trustworthiness, but rather as precursors to *enough* trustworthiness to persuade belief change. This challenges prior literature suggesting that trustworthiness alone may have a greater influence than expertise in evaluating the credibility of a source. Lastly, we highlight the dual nature of peer-supplied credibility disputes, outlining scenarios in which they could backfire, such as when offered by hyper-partisan ties with whom one shares a low political agreement.

2. Related work

Referring to the definitions laid down by the Information Disorder Framework (Wardle and Derakhshan, 2018), this paper regards "misinformation" as a term encompassing both *misinformation*, the unintentional sharing of objectively false information, and *disinformation*, the deliberate dissemination of false information for malicious reasons. Misinformation has become pervasive in today's online environments, especially on social media platforms where users' beliefs determine what is shared and subsequently proliferated to entire networks (Vicario et al., 2016; Friggeri et al., 2014). Once misinformation is wrongfully accepted as the truth, misperceptions are forged and buttressed by repeated exposure to false information, which tend to be sticky and often highly resistant to corrections (Thorson, 2016).

To tackle misinformation, understanding the cognitive, affective, partisan, and social motivations behind its spread and believability, and how they can be leveraged to design effective interventions is critical. In the following sections, we summarise the research that has been conducted in this area. We begin by discussing how self-expression and affective responses to misinformation act as drivers of its conscious propagation. We then discuss the unique affordances of social media that can be leveraged to transform it into a promising antidote to misinformation, despite being the very grounds for the creation and propagation of it. Next, we highlight the current platform-based algorithmic interventions and outline reasons for their limited effectiveness in causing belief revision. Further, we examine *trust* in a correction's source, *goodwill* exhibited by them, and source-receiver *attitude homophily* – one's inclination to associate and form social connections with others who share similar political beliefs or attitudes (Gillani et al., 2018) – as heuristics for credibility assessment. We then identify scenarios in which even belief-challenging corrections may be effective when designed to overcome motivated reasoning and confirmation bias. Lastly, we also discuss conditions under which corrections may cause a belief change in an unintended direction.

2.1. Misinformation on social media and existing interventions

The tendency of misinformation to 'spread like wildfire' can be attributed to user motivations which are not entirely rooted in a lack of information literacy — empirical works demonstrate that sharing misinformation fulfils psychological needs which often leads to its virality (Acerbi, 2019). For example, Chen and Sin (2013) analysed motivators for sharing misinformation and found that 67.8% (n = 116) of their respondents willingly shared it as a way of obtaining others' opinions on the subject and of expressing their own. Similarly, Chen

et al. (2015) and Jahanbakhsh et al. (2022) also observed that individuals without malicious intentions shared misinformation for reasons grounded in social rather than informational needs — viewing it as a way of self-expression.

Notably, people who unknowingly share debunked misinformation often do so because of their emotional response to reading it, as misinformation tends to carry a greater affective footprint than the truth (Stieglitz and Dang-Xuan, 2013). Further, online communications containing misinformation often employ moral-emotional keywords (for example, fight, evil, punish, greed, etc.), which are more cognitively attractive to readers. Multiple recent studies have identified the tendency of emotional and moral content to frequently reach virality on social media, irrespective of their veracity (Brady et al., 2017; Stieglitz and Dang-Xuan, 2013). Brady et al. (2020b) attribute this trend to the ability of moral-emotional terms to capture more attention than neutral terms, an effect so influential that the presence of a single moral-emotional word in polarising Tweets increased retweeting by 20%.

2.1.1. Social media as an avenue for misinformation correction

In the context of political misinformation, the unique affordances of social media platforms make them a suitable avenue for refutation of misleading claims and reduction in misperceptions. Firstly, despite the prevalence of partisan and ideological echo chambers on social networking sites, the selective exposure to information is not as extreme as other media choices, which may be solely driven by hyper-partisan beliefs and cater to selective political narratives (Barberá et al., 2015). Consequently, social media platforms carry the potential of exposing their users to corrective information through platform-based interventions, like algorithmic corrections, and posts shared by politically diverse members of one's online network (Messing and Westwood, 2014). Secondly, a correction or credibility label appended to social media posts can influence not just the original poster, but also other members of their online community who come across the post. This strategy distributes corrective messages to those who may be unwilling to assess the veracity of posts by themselves, through a phenomenon referred to as "observational correction" (Vraga and Bode, 2020). Lastly, although the rate of diffusion of corrective messages is significantly lower than misinformation (Vosoughi et al., 2018), when corrections or credibility labels are attached to posts containing misinformation, they can travel through online networks at the same speed as the misinformation by *piggybacking* on it, which can alert individuals to the questionable veracity of posts.

2.1.2. Existing interventions

While seemingly promising, current algorithmic interventions on social media platforms to battle misinformation exist on a spectrum of severity and offer varying effectiveness (Saltz et al., 2021a). On one end, harsh interventions like the altogether removal of posts containing misinformation can prevent users from viewing disputed content, but may also impinge on their right to free speech, with some arguing it borders on censorship. More moderate approaches, like *downranking* false posts, and *shadowbanning* user profiles identified as spreading false information, reduce the spread of misinformation by decreasing the number of times it appears in users' feeds. However, they offer limited scalability because misinformation is created more rapidly than it can be assessed by platforms (Epstein et al., 2020), allowing dubious content to seep through networks unchecked.

Lenient interventions like "soft moderation" techniques which display credibility or contextual labels on posts do not curtail free speech or impact user autonomy, and instead attempt to empower individuals to evaluate information for themselves. Although such algorithmic labels may appear as a "more nuanced" strategy for gauging post veracity while encouraging freedom of expression when compared to blanket removal of disputed content (Morrow et al., 2022), there is conflicting evidence outlining their effectiveness. Both Yaqub et al. (2020) and Mena (2020) found that irrespective of partisanship, labels successfully reduced people's intention to share misinformation, while Oeldorf-Hirsch et al. (2020) reported the influence of labels to be almost non-existent. Interestingly, further highlighting the ineffectiveness of such credibility labels, Saltz et al. (2021a) found platform-applied fact-checking labels were perceived as "punitive and patronising", with several participants finding it offensive that the platform tried to inform them whether a displayed post was true or not.

Since platforms where misinformation tends to flourish, like Facebook, are also profit-driven, users are often generally resistant towards the misinformation labels they supply due to the institutional distrust that arises from a potential conflict of interest (Saltz et al., 2021b). These platforms face a critical choice — algorithmically promote engagement by proliferating misinformation which often gets more clicks than the objective truth (Sharma et al., 2017), or combat misinformation by dampening the spread of false posts at the expense of revenue. Moreover, this distrust is often exacerbated due to algorithmic moderation being perceived as inherently biased and lacking "human factors of cognition", and having the least perceived legitimacy and trustworthiness when compared to other moderation processes, like expert verification (Pan et al., 2022).

Building on the aforementioned literature which highlights the limited effectiveness of platform-supplied corrective labels, this study explores the influence of credibility labels when they are supplied by *one's social media peers*, rather than the platform.

2.2. The role of source credibility, goodwill, and tie strength in changing beliefs

Source credibility, in this context defined as the perceived trustworthiness of sources of misinformation and their corrections, heavily influences the willingness of individuals to accept new information (Ecker et al., 2022) and further propagate it (Flintham et al., 2018). While the influence of source credibility diminishes in case of media outlets as they are often perceived to have a political agenda (Dias et al., 2020; Wintersieck et al., 2021), its influence remains substantial in the acceptance of information originating from non-media sources (Nadarevic et al., 2020; Walter and Tukachinsky, 2020; Amazeen and Krishna, 2020). Thus, in this study, we seek to investigate the influence of source credibility when corrective messages come from a non-media source, i.e., one's social media peers.

For assessing source credibility, individuals tend to rely on two heuristics: source *expertise* and *trustworthiness* (Metzger et al., 2010). In addition to expertise and trustworthiness, some researchers argue for a third determinant of source credibility – *goodwill* (Lupia and McCubbins, 1998), which examines whether the source values the receiver's well-being (McCroskey and Teven, 1999). In the context of interpersonal communication, through direct messaging or sharing posts on social media, the credibility of the sender is often regarded as a proxy for the credibility of information they share (Kang et al., 2011; Metzger et al., 2010). For example, how individuals perceive the sender of information has a greater influence on belief and sharing decisions than the source of the information (Chakrabarti et al., 2018). Further, the decision to share information encountered online can often be driven by how the sender is perceived, rather than its content or journalistic source. Moreover, as corrective messages perceived to have a persuasive intent can often fail and cause counter-argumentation, individuals are more likely to follow the advice of their close peers, perceiving their corrections as less driven by persuasion compared to those from strangers (Van Noort et al., 2012; Petty and Cacioppo, 1979).

These findings suggest that *strong ties* – individuals who have one's best interests at heart and with whom one shares an intimate relationship (Campbell et al., 1986) may have greater perceived credibility than weaker ties. Shahid et al. (2022) corroborate this notion by reporting that articles shared by one's family members engendered the highest

level of (misplaced) trust in fake news articles, when compared to other sources like celebrities and journalists. Even in cases where the news originates from an unfamiliar media outlet, individuals are inclined to share it with others if it is disseminated by someone they trust (Sterrett et al., 2019). Together, these works highlight how the trust individuals place in news encountered on social media is closely tied to the identity of the person sharing it. However, it remains to be understood whether this influence can be harnessed to foster acceptance of peer-supplied credibility disputes.

Our study builds upon this literature by systematically examining the role of the *source of a credibility dispute*, rather than the *source of misinformation*. Specifically, we seek to examine how characteristics of the source of a credibility dispute can enhance its acceptance by the receiver, subsequently diminishing belief in the misinformation it is trying to dispute. Since the persuasiveness of a source increases with its perceived credibility (Pornpitakpan, 2004), we investigate how relational closeness between the source and receiver of a credibility dispute may reduce belief in fake news. We also seek to understand the difference in the persuasiveness of credibility disputes when supplied by *strong* ties versus *weak* ties, who may have differing perceived trustworthiness and goodwill.

2.3. Corrections that challenge existing belief systems

In addition to source credibility, belief change can be impacted by corrections that violate one's pre-existing ideological or partisan attitudes, as people tend to prefer information that aligns with their long-held beliefs in order to avoid *cognitive dissonance* (Festinger, 1962) – a phenomenon referred to as selective exposure. In the following sections, we outline the fundamental reasons behind why belief-challenging corrections are difficult to ingrain, and outline how attitudinal homophily may be leveraged to overcome motivated reasoning, making corrections more persuasive.

2.3.1. Worldview

An individual's *worldview* – the ideology, values, and beliefs that underpin one's socio-cultural identity – can substantially impact willingness to accept corrections to closely held beliefs. Previous works have demonstrated that when corrections violate existing attitudes, people either ignore the worldview-challenging information (Nyhan and Reifler, 2010), or worse, focus exclusively on attitude-congruent information (Prasad et al., 2009; Lewandowsky and Oberauer, 2016). This reluctance to accept worldview-challenging information persists regardless of whether incorrect news reports are attributed to an honest mistake or deception — in either case, individuals are more inclined to discredit information that contradicts their worldview (Axt et al., 2020). The influence of worldview is significant, so much so that the perceived credibility of a news claim is influenced more by an individual's alignment with its political message than the political association of the media source publishing the news (Jakesch et al., 2018). In these cases, misinformation does not influence attitudes, but rather attitudes influence what people choose to believe or dismiss as truth, particularly when people are cognitively occupied and lean on heuristics to estimate the veracity of information encountered (Ecker et al., 2011). This is especially prevalent in a political context, where corrections invalidating one's worldview can simply fail (Ecker and Ang, 2019). Such corrections that contradict pre-existing perceptions, which one also shares with other group members, can be perceived to be attacking one's identity and undermining one's intellectual authority, hindering belief change (Hornsey and Fielding, 2017). For example, an anti-vaccination advocate may view information that disputes the notion that vaccines cause autism as opposing their identity and challenging their autonomy. Therefore, it is important to examine the effectiveness of peer-supplied credibility disputes when they are more likely to contradict one's pre-existing partisan biases and worldviews.

2.3.2. Motivated reasoning

To protect pre-existing attitudes, beliefs are often governed by motivated reasoning, where information that aligns with prior attitudes is evaluated to be more credible than attitude-challenging information (Taber and Lodge, 2006). This belief asymmetry runs across partisan lines too, predisposing both Republicans and Democrats to accept misinformation surrounding certain themes as truth, and disregard corrections to certain other communications as being invalid. For example, as per the American political landscape at the time of writing in October 2023, most Republicans and Democrats have contrasting beliefs in the urgency posed by the changing climate (Pew Research Center, 2020b), which can elicit different beliefs in climate change related communications, governed by their partisan predilections. Although partisanship and motivated reasoning cause selective exposure to misinformation and influence individuals' tendencies to dismiss or accept corrections, both Republicans and Democrats tend to believe misconceptions that either support their party or denigrate the opposition party (Kraft et al., 2015). Even adding politically-diverse news articles (with their political stance and credibility highlighted) to people's information diet is often insufficient to overcome motivated reasoning (Gao et al., 2018). Therefore, it is crucial to investigate whether encountering a credibility dispute raised by one's social media peer is influential enough to help overcome motivated reasoning, and reduce belief in attitude-congruent misinformation.

2.3.3. In-group ties

Interestingly, psychologically identifying with a group like a political party and subscribing to its beliefs can be a double-edged sword. While associating with a group – one's *in-group* – can promote (misplaced) beliefs, it may also be the very antidote to wrongful belief perseverance in the face of credibility disputes. Messages shared by in-group members tend to be systematically processed (Chaiken, 1987), and thus carry the potential of causing significant belief change (Mackie et al., 1990).

Overall, since source trustworthiness is a more important credibility indicator than expertise, as delineated in the seminal paper by McGinnies and Ward (1980), belief-challenging corrections, when received from in-group members, may be more persuasive than those shared by out-group members and serve as an effective intervention for online misinformation, which this study investigates. For instance, on WhatsApp, corrections to news headlines are more likely to be shared when sent by an in-group member compared to an out-group member (Pasquetto et al., 2022). However, it is important to further investigate how in-group ties on social media platforms can influence individuals' beliefs in news headlines, in addition to their sharing intentions. Moreover, it is unclear how credibility labels on online platforms can influence belief in news headlines when both tie-strength and political affiliation with their source are at play. This study aims to bridge that gap by examining the persuasiveness of credibility disputes from individuals at various intersections of relationship closeness and political agreement.

2.4. Unintended consequences of misinformation correction

The effectiveness of misinformation corrections has been a subject of extensive research in various fields, including psychology, political science, and journalism. While there is empirical support for the positive impact of fact-checking, such as causing a decrease in the sharing of misinformation (Henry et al., 2020; Pasquetto et al., 2022), the impact of fact-checking exists on a spectrum of effectiveness. A critical challenge arises because of the existence of *belief echoes*, where even after misinformation is corrected, it can continue to influence beliefs and attitudes (Thorson, 2016). This persistence is particularly pronounced among individuals whose worldview aligns with the misinformation being corrected, even if they are offered corrections immediately after exposure to the misinformation (Garrett and Weeks, 2013).

Political news introduces additional challenges, with corrections often proving ineffective against inaccurate but highly persuasive political narratives (Barrera et al., 2020).

Notably, for some individuals, corrections can also backfire: being exposed to belief-contradicting evidence makes them adopt said beliefs even more strongly. When corrective messages induce such a *boomerang* or *backfire* effect, they can promote an attitude change in a direction unintended by the message, further entrenching false beliefs (Byrne and Hart, 2009). Several researchers have theorised that people engaging in psychological rebellion or reactance kindles this phenomenon (Brehm, 1966; Byrne and Hart, 2009), wherein they perceive fact-checking messages as a threat to their intellectual abilities and to truths central to their identity, and attempt to reestablish their freedom by embracing incorrect beliefs with increased vigour.

However, there is limited literature that has found evidence for the backfire effect (Trevors and Duffy, 2020), a phenomenon that is challenging to reproduce (Weeks, 2015; Wood and Porter, 2019) and which recent reviews suggest may not be as prevalent as previously thought (Nyhan, 2021; Swire-Thompson et al., 2020). Nyhan (2021) state that the newest consensus in literature is that corrective information usually does not lead to a backfire effect; instead, it generally results in modest but significant improvements in belief accuracy. Swire-Thompson et al. (2020) further highlight the need for improved measurement and experimental designs to accurately assess this phenomenon. Thus, in this study, we investigate the existence of any backfire effects by presenting, as facts, misinformation that participants would be more likely to believe in based on their political leaning, and then showing corrections to this misinformation which would challenge deeply held prior beliefs.

2.5. Summary

In summary, studies investigating online misinformation reveal a pervasive distrust among users toward platform-supplied credibility labels, ascribed to the platforms' engagement-driven and politically-biased motivations (Saltz et al., 2021b). Individuals may also resist accepting corrections that contradict prior beliefs. However, the effectiveness of credibility disputes raised by a source associated with trust and goodwill, such as a close social media peer, remains an open question. This study seeks to address this gap by investigating the effectiveness of credibility labels not supplied by platforms, but by one's online peers. This intervention resembles the "social media" credibility label examined by Yaqub et al. (2020), although our investigation extends beyond theirs in several ways. While Yaqub et al. (2020) test a singular "social media" credibility label which states that participants' "social media friends" dispute an article's credibility, they do not consider the characteristics of one's relationship with said "social media friends", which may heavily impact the persuasiveness of credibility disputes they raise. Our study takes a more comprehensive approach by considering the strength of the relationship and the political affiliation between the participant and their social media peer raising the credibility dispute, which remain to be examined despite the high likelihood of these factors existing simultaneously in any relationship. Moreover, our examination expands beyond Yaqub et al. (2020)'s work by assessing not only participants' intention to share a headline after a credibility label is applied, but also how participants' perceived accuracy of the headline is influenced by the presence of a credibility label.

Furthermore, while the impact of source partisanship on *fostering* belief in misinformation has been studied, existing work does not examine whether it can also *reduce* belief in misinformation. Additionally, little attention has been given to understanding how the effectiveness of peer-supplied credibility disputes may be influenced by how strongly their receiver identifies with their political orientation. Thus, our study aims to thoroughly examine the effectiveness of peer–based credibility labels as a misinformation intervention, evaluating how their impact may vary based on the relationship closeness and political similarity between the source and the receiver, while also considering the strength of the receiver's identification with their political party.

3. Method

To assess the influence of source tie strength and political agreement on reducing belief in misinformation, we deployed an online survey-based experiment. This survey measured belief in misinformation before and after viewing credibility labels sourced from one's social network peers. Such pretest-posttest experimental design approaches have been widely utilised in multiple studies seeking to measure belief change caused by an intervention (Garrett et al., 2013; Thorson, 2016; Park et al., 2021). This approach also allowed us to simulate a scenario where users may come across news headlines on social media and are required to rely on heuristics or visual cues to gauge their veracity, with our label being one of these credibility cues that they may come across. In the sections that follow, we first describe the recruitment details, followed by our survey questions, then the overall experiment design, and lastly, the methods we used to perform quantitative analysis.

3.1. Measures and participants

Fig. 1 provides an overview of the overall experiment design. We utilised a 2 (Source Tie Strength: Strong or Weak) x 2 (Source Political Agreement: High or Low) between-subjects factorial design, as represented in Fig. 1(f). We used Qualtrics to create the survey and Prolific to recruit participants, restricting their political affiliation to the two major parties that dominate American electoral politics. Hence, participants were chosen based on a screening criteria which outlined that they had to be located in and be residents of the United States, and identify either as a Republican or a Democrat. We also only presented the survey to participants with a minimum approval rate of 85 on Prolific, and ensured that no participant took part in more than one experimental condition. We determined the sample size using G*Power (Faul et al., 2007), with a medium effect size ($f^2 = 0.2$), $\alpha = 0.05$, and a power of 0.8 (following established methodological recommendations by Cohen (1992)). The suggested sample size was 75 participants. To uphold reliability and ensure a balance of participants across our four conditions, we conservatively recruited 96 participants overall, equally divided between Republicans and Democrats for representativeness.

3.2. Procedure

The survey began by collecting participants' demographic data and political orientation (Fig. 1(a)). We also asked participants to rate how strongly they identified as associating with their political orientation, and recorded their most frequently used social networking and instant messaging platforms. Next, we asked participants to provide the first name of a peer who satisfied a two-part description, each part varying with a participant's assigned **Tie Strength (TS)** and **Political Agreement (PA)** treatment, as shown in Fig. 1(a), following similar studies investigating such "social debunking" approaches (Pasquetto et al., 2022). One half of the participants were asked to provide the name of a *strong* tie (i.e., with whom they shared a close relationship, like a family member), while the other half were asked to name a *weak* tie (i.e. with whom they had a distant relationship, like an acquaintance). Within each of these two treatments, half the participants were asked to name a tie who shared the same political beliefs as them (*high* PA), and the other half named someone who disagreed with their political beliefs (*low* PA). We note that in this work, *low PA* represents political disagreement between the participant and their social media peer, while *high PA* represents a similar political affiliation between the two. By considering the political agreement between the source and receiver of credibility disputes rather than their absolute political orientations, we aim to increase the likelihood of our findings being applicable to a broader range of political identities, beyond Republicans and Democrats. Together, these two descriptions assigned each participant into one of the four possible experimental conditions, namely *Strong TS* × *High PA*, *Strong TS* × *Low PA*, *Weak TS* × *High PA*, and *Weak TS* × *Low PA*.

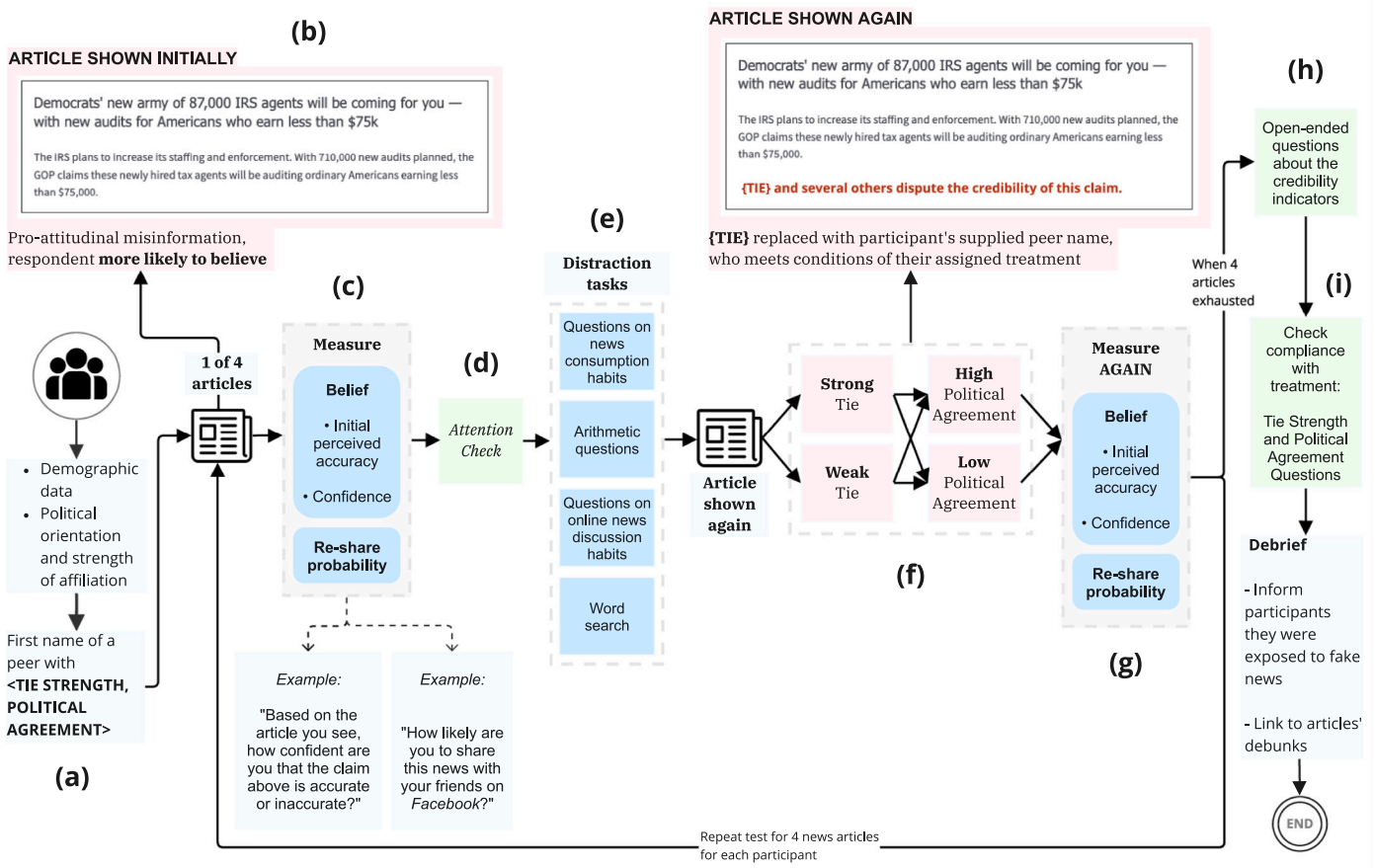

Fig. 1. The full experiment design. TIE STRENGTH and POLITICAL AGREEMENT vary for participants based on their assigned treatment, and TIE is replaced by the name of the peer the participant provides. **(a)**: Initial demographic and political belief questionnaire, and procurement of peer name. **(b)**: The original belief-affirming misinformation chosen based on respondent partisanship. **(c)**: Measurement of initial belief in misinformation and sharing intent. **(d)**: Attention check questions to identify inattentive respondents. **(e)**: Short distractor tasks between belief change measurements, and their types. **(f)**: Same misinformation presented again with a credibility label sourced from the peer participants named. **(g)**: Re-measurement of belief in misinformation and sharing intent. **(h)**: Open-ended questions to enquire about the effect of these labels. **(i)**: Self-reported tie strength and political agreement evaluations by participants to identify non-compliance with their assigned treatment, and debriefing related to participants' exposure to misinformation.

3.2.1. News headlines

After naming a peer, participants were shown a brief news headline containing misinformation, an example of which is presented in Fig. 1(b). We deliberately chose not to emulate a "Facebook-style" news post (Pennycook et al., 2021a)—complete with a headline, byline, related photograph, media source, and user engagement metrics such as likes and comments. Instead, we displayed the article's headline and a brief description. This presentation decision stemmed from various considerations. Presenting headlines without these additional details allowed us to establish a controlled setting where we could isolate the influence of our credibility labels. Further, the absence of extraneous details also allowed for a clearer examination of the causal relationship between our manipulated variables, *tie strength* and *political agreement*, and our variable of interest, the *degree of belief change*. Lastly, we recognise that not all individuals use Facebook, and employing a format tied to a specific platform may introduce confounding variables related to user familiarity. By avoiding a platform-specific format, our work aims for broader generalisability of our findings across diverse social media contexts.

To ensure that the headlines contained verifiably false information relevant to the political climate at the time of writing in October 2023, we sourced headlines from Politifact, a comprehensive fact-checking website. In doing so, we ensured that our stimuli consisted of content that had, to some extent, already spread on social media. This approach allowed us to examine headlines that are more likely to gain traction online and warrant fact-checking, ensuring a more realistic representation of content circulating on social media platforms (Pennycook et al., 2021a).

To accurately assess the effectiveness of peer-supplied credibility indicators in reducing belief in misinformation, we evaluated their influence in scenarios where they had to do significant *corrective work* — specifically when misinformation reaffirmed participants' existing partisan biases, making it highly believable due to motivated reasoning. In such scenarios, our credibility labels questioned misinformation that aligned with participants' partisan views, making belief change more challenging compared to situations where misinformation contradicted participants' worldview and was inherently less believable from the outset. Thus, the fake news articles displayed for each participant were chosen based on their reported political leaning, presenting Democrats with a different set of articles than Republicans.

All participants saw attitude-congruent misinformation in the headlines – i.e., misinformation that they were more likely to believe in due to their preexisting beliefs – to engender as many misperceptions as we could, for our corrective efforts later. The authors systematically identified issues characterised by significant belief polarisation between Republicans and Democrats by cross-checking several datasets on U.S. political polarisation, such as comprehensive surveys conducted by the Pew Research Center (see Pew Research Center (2020a, 2014, 2019)). This informed the selection of topics such as scepticism towards COVID-19 vaccinations for Republican-leaning participants, and anti-Trump sentiments for Democrat-leaning participants. We then curated misinformation pertaining to these topics from Politifact. The aim was to ensure that the misinformation presented to participants was congruent with their political beliefs, making it more likely to be perceived as credible, so we could subsequently test the effectiveness of our

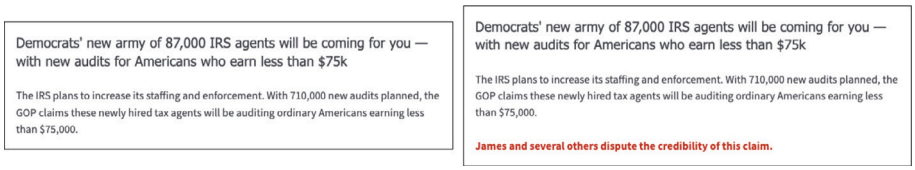

Fig. 2. One of the four headlines shown to Republican respondents containing attitude-congruent misinformation. *Left:* The original misinformation; *Right:* Misinformation with a peer–supplied credibility label; participant-supplied peer name replaces "{Tie}".

credibility labels. Fig. 2(a) shows an example of misinformation shown to Republican-leaning participants. Over the course of the experiment, each respondent was shown a total of four news headlines containing believable misinformation. By measuring the effectiveness of credibility labels over four news headlines, we increase the likelihood of any observed change in belief being attributed to our experimental manipulations, rather than specific characteristics of the news headline. This approach enables us to draw causal inferences about the impact of peer-supplied credibility labels. The full list of headlines presented to both Republican and Democrat respondents is presented in Appendix A.

3.2.2. Belief measurement

After participants were shown a fake news headline, they answered several questions designed to measure their belief in the headline and their intention to re-share, as illustrated in Fig. 1(c). We measured belief using a 5-point Likert scale ranging from *"Definitely Inaccurate"* to *"Definitely Accurate"*. We also measured their confidence in their reported belief, i.e., how certain they were about their evaluation of the article's credibility, on a sliding scale of 1 to 100, with a higher score indicating higher confidence in their belief. In an attempt to nullify the bias generated by the starting position of the anchor on a slider (Sellers, 2013), our sliders started unmarked, with an anchor appearing only after users clicked on the slider's range. Lastly, to investigate whether peer–supplied credibility labels had an influence on re-sharing probability, we asked respondents how likely they were to share the headline on their most used social networking and instant messaging platforms.

In order to measure the efficacy of our credibility labels, we measured participants' belief and confidence in misinformation *twice* – once before showing the credibility labels and once after presenting it. After their initial belief and re-sharing intent were measured, participants were shown the news headline again, with the addition of a credibility label, as displayed in Fig. 1(f). We asked participants to imagine that they had come across this news headline on their most used social networking site, and that their named peer had disputed its credibility, following similar studies investigating social corrections (Pasquetto et al., 2022). This variant of the headline was designed to make it seem that a credibility dispute had been supplied by the peer whose first name participants shared with us at the start of the experiment, as shown in Fig. 2. These disputes, presented in red to make them visually apparent, highlighted that the article's credibility was being disputed by a peer. This process was repeated for each of the four news headlines for every participant, and the order of the headlines was randomised to account for any ordering effects. As shown in Fig. 1(d), to screen inattentive participants and ensure comprehension, we inserted two attention check questions with clear answers throughout the survey, such as "Please enter the word "cherry" when asked for your favourite colour. What is your favourite colour?". They were adapted from existing methods to gauge survey participant inattentiveness (Huang et al., 2012), and those who failed both checks ($n = 7$) were removed from our final dataset. Additional participants were recruited until we obtained 96 valid responses from individuals who successfully passed at least one attention check.

Between reading the uncorrected misinformation and the corrected misinformation, participants also completed a short distractor task, as represented in Fig. 1(e). Distractor tasks are commonly employed in studies which measure belief or attitude before and after a manipulation, to allow for the manipulation's effects to manifest, and to prevent immediate recall of initial stimuli (LaPaglia et al., 2013; Fulton et al., 2022; Brown, 1958). In line with previous studies that incorporate distractor tasks when investigating misinformation interventions (Pasquetto et al., 2022; Thorson, 2016), we utilised tasks, such as arithmetic questions and word searches, between the presentation of the uncorrected and corrected misinformation. By engaging participants in distractor tasks, we aimed to minimise demand characteristics and prevent participants from consciously altering their responses based on their awareness of the study's objectives. Overall, utilising distractor tasks enabled us to gather more naturalistic measurements of participants' updated belief after the credibility disputes were presented to them. We deployed four distinct distractor tasks, one for each fake news headline presented to the participants, and the order of these tasks was randomised to counter any ordering biases.

At the end of the survey, we probed participants through open-ended questions to gain insights into how they felt about receiving credibility disputes from their peers, and what the main reasons were behind any change in belief (Fig. 1(h)). We were also interested in understanding whether the presence of credibility labels influenced their decision to share the fake news headline with their social networks. Moreover, since participants were exposed to misinformation, they were informed about it at the end of the study, and were presented with links from Politifact that thoroughly debunked the misinformation presented to them.

The Ethics Committee of our university approved the study. Participants took a median time of approximately 15 min to complete the survey and received around US$4 for participation.

3.2.3. Accounting for non-compliance with assigned treatment

We recognise that the validity of our analysis depends on our participants' ability to name a peer that matches their assigned treatment — when asked to name a weak tie, participants may nevertheless mention a strong tie. To account for any non-compliance, we presented a set of standardised questions at the end of the survey to measure participants' relationship closeness and political agreement with the peer they had named at the start of the experiment, shown in Fig. 1(i). These compliance checks were in line with those undertaken by similar experiments (Pasquetto et al., 2022). We measured tie strength using questions from the Uni-dimensional Relationship Closeness Scale (URCS), which measures perceived trustworthiness and connection strength (Dibble et al., 2012), as relationship closeness is an effective indicator of tie strength (Marsden and Campbell, 1984). Questions were adapted by replacing the words "my [tie]" with the name of the peer as supplied by individuals. For example, "My relationship with my [tie] is close" was presented as "My relationship with {TIE} is close", on a 7-point Likert scale ranging from 1 (Strongly Disagree) to 7 (Strongly Agree), with {TIE} being substituted with the name provided by the participant. The items were averaged to create a single overall closeness score (range: 1–7) per participant, which was used to check tie strength compliance. Participants with a closeness score less than or equal to 4 in the *Strong TS* condition, or greater than or equal to 5 in the *Weak TS* condition, were removed for non-compliance. Further, political agreement between the respondent and their named tie was ascertained using the question "How much does {TIE} agree or disagree with your

political beliefs?", on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), with {TIE} again being substituted with the name provided by the participant. Responses with discrepancies between their assigned and self-reported levels of political agreement were discarded. After undertaking these two compliance checks, new participants were recruited until we received 96 valid responses divided equally between the four conditions and respondent partisanship.

4. Results

The final valid dataset contained 96 participants' reported beliefs in 4 headlines, yielding 384 initial and final belief measurements, split equally between Democrats and Republicans. The demographic information of participants is presented in Table B.3 in Appendix A. In the following sections, we report the findings from our quantitative analysis, along with our qualitative analysis procedure and subsequent findings.

4.1. Quantitative analysis

We first undertook a manipulation check to verify whether the misinformation we presented was indeed perceived by participants to be attitude-congruent. We then built three generalised linear mixed-effects models (GLMM) to investigate how (1) belief in misinformation, (2) confidence in one's credibility evaluation, and (3) intent to share misinformation is influenced after seeing credibility disputes by peers at different intersections of tie strength and political agreement.

Our results indicate that being alerted about the disputed credibility by one's in-group peers greatly reduces belief in misinformation, irrespective of tie strength **(RQ1)**. Furthermore, the strength of one's political orientation, i.e., whether one identifies strongly or weakly with their political party, also plays a statistically significant role in belief change. However, in this study we neither find significant differences in participants' confidence in their belief, nor likelihood to share news headlines **(RQ2)**, before and after the credibility label is presented to them. Lastly, we underscore the dual nature of peer-supplied labels, outlining scenarios where they may backfire for certain individuals **(RQ3)**. In the next sections, we describe these findings in detail, starting with a manipulation check analysis.

4.2. Manipulation check

To verify whether the misinformation we presented to participants was indeed perceived to be attitude-congruent, we analysed participants' initial belief in the headlines. We hypothesised that participants would be more likely to report higher belief in headlines if they were attitude-congruent. As expected, Republican participants reported the misinformation presented to them to be Moderately Accurate (coded as 4) or Definitely Accurate (coded as 5) 66.3% of the time (mean confidence = 73%), while Democratic participants did so 70% of the time (mean confidence = 74.5%). These results firmly establish that attitudinal congruence manifested as intended, and user behaviours observed did not occur randomly, but under the influence of our experimental conditions, confirming the validity of our results.

4.3. Model construction

We define the following three outcome variables of our analyses:

- **Change in Belief:** Difference in belief before and after exposure to the credibility label, defined as $\text{Belief}_{final} - \text{Belief}_{initial}$. Belief was measured on a 5-point Likert scale from "Definitely Inaccurate" (coded as 1) to "Definitely Accurate" (coded as 5). A negative value of Change in Belief represents a reduction in belief in misinformation, while a positive value represents an increase in misperceptions.
- **Change in Confidence:** Difference in participants' confidence in their evaluated credibility before and after exposure to the credibility label, defined as $\text{Confidence}_{final} - \text{Confidence}_{initial}$. Range 0 to 100.
- **Change in Sharing Intent**: Difference in participants' willingness to share news headlines on their most used social media and instant messaging platforms, before and after applying the credibility label. Sharing Intent was measured on a 5-point Likert scale from "Extremely Unlikely" (coded as 1) to "Extremely Likely" (coded as 5).

We investigated the impact of the following six predictor variables on persuading changes in participants' belief, confidence, and sharing intent:

- **Tie Strength (TS):** A binary categorical variable, (1) strong (2) weak, indicating the level of relationship closeness between the receiver and the source of the credibility dispute.
- **Political Agreement (PA):** A binary categorical variable, (1) high (2) low, indicating whether the source and the receiver of the credibility dispute identify as belonging to the same political party.
- **Self-Reported Strength of Political Orientation:** An ordinal variable, measured on a 3-point Likert scale ("Mildly", "Moderately", "Strongly"), in response to the question "How strongly do you identify as a {POLITICAL ORIENTATION}?"
- **Political Interest:** A continuous variable, calculated after aggregating responses to questions that inquire about a participant's interest in politics and current affairs. Range 1–100.
- **Online News Discussion Frequency:** A continuous variable, calculated from a participant's frequency of sharing and discussing news on their most used social networking and instant messaging platforms. Range 1–100.
- **Confidence**$_{initial}$: A continuous variable, reflecting how sure a participant is about their credibility evaluation of the article, before being shown the peer-supplied credibility labels. Range 1–100. Adding initial confidence to the model allowed us to account for participants' initial (un)certainty in their belief on the effectiveness of our credibility labels, as well as any prior knowledge of the headline. In research exploring shifts in beliefs or attitudes, such as investigations into misinformation (Jahanbakhsh et al., 2023; Wijenayake et al., 2021) and conformity (Wijenayake et al., 2019, 2020), researchers often use participants' initial confidence as a proxy for their prior knowledge or uncertainty regarding the given task.

We employed the statistical R package `lme4` (Bates et al., 2015) to build three generalised linear mixed-effects models (GLMMs) of the relationships between the aforementioned predictor variables and each outcome variable. This enabled us to determine the impact of a group of predictor variables on each of our outcome variables. We specified participant IDs as a random effect in our models, to account for individual differences as well as any correlation amongst repeated measurements from the same participant.

4.3.1. The effect of political agreement and tie strength on change in belief *(RQ1)*

As shown in Table 1, we found a statistically significant main effect of Political Agreement on Change in Belief ($\beta = 0.928$, $SE = 0.104$, $p < 0.001$), with an odds ratio of 2.47 (95% CI between [2.02, 3.02]). This indicates that when disputes were offered by individuals with whom one has high Political Agreement, the odds of experiencing a change in belief in misinformation were 2.47 times higher compared to instances of low Political Agreement. We performed a post-hoc analysis using Bonferroni correction to obtain the estimated marginal means (emmeans) for the two levels of Political Agreement – *high* and *low*.

Table 1
Effect of predictors on participants' change in belief. Statistically significant main effects ($p < 0.05$) are in bold. The sign of the estimate (+/−) denotes the direction of the relationship between the predictor and change in belief.

Variable	Estimate	Std. Error	Odds Ratio	95% CI	p value
Baselines: Tie Strength = Strong, Political Agreement = High, Party Strength = Mild					
Tie Strength = Weak	−0.137	0.104	0.87	[0.72, 1.07]	0.191
Political Agreement = Low	0.928	0.104	2.47	[2.02, 3.02]	<0.001
$\text{Confidence}_{initial}$	0.003	0.002	1.07	[0.98, 1.18]	0.148
Party Strength = Moderate	−0.199	0.140	0.82	[0.63, 1.08]	0.162
Party Strength = Strong	−0.518	0.143	0.60	[0.46, 0.79]	<0.001
Political Interest	−0.004	0.003	0.93	[0.85, 1.03]	0.190
News Discussion Frequency	−0.108	0.012	0.91	[0.81, 1.03]	0.358

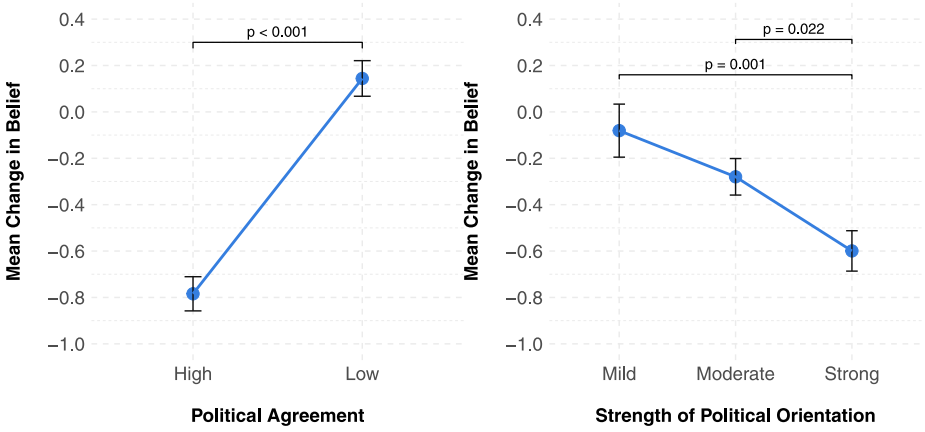

Fig. 3. Estimated marginal means. *Left:* Mean change in belief in misinformation through credibility disputes offered by peers with *high* vs. *low* Political Agreement with the receiver, irrespective of Tie Strength; *Right:* Mean change in belief in misinformation depending on how strongly receivers identify with their political parties. Error bars denote standard error (SE).

The average reduction in belief in misinformation was significantly greater in scenarios of *high* Political Agreement ($M = -0.784$, $SE = 0.073$), compared to *low* Political Agreement between the source and receiver ($M = 0.144$, $SE = 0.076$). In other words, participants felt a notable decrease in belief in misinformation when credibility labels were supplied by members of the same political party as them, irrespective of their level of relationship closeness (Fig. 3). Moreover, we did not find a statistically significant difference in belief change behaviour between Republican and Democratic participants.

In contrast to Political Agreement, our results do not indicate a statistically significant main effect of Tie Strength on belief change ($\beta = -0.137$, $SE = 0.104$, $p = 0.191$). In other words, credibility disputes by relationally close peers were no more effective in reducing misperceptions than those by relationally distant peers.

4.3.2. The effect of strength of political orientation on change in belief

The predictor Strength of Political Orientation had a statistically significant main effect on belief change. We performed a post hoc analysis using Bonferroni adjustment to obtain the estimated marginal means for the three levels of Strength of Political Orientation (*mild*, *moderate*, and *strong*). Due to the uncontrolled nature of this predictor variable, there were $n = 20$ who identified as mild, $n = 42$ who identified as moderate, and $n = 34$ who reported being strong partisans. Results show that there is a significant difference in belief change between *mild* ($M = -0.080$, $SE = 0.114$) and *strong* ($M = -0.599$, $SE = 0.087$) partisans, $p = 0.001$. Furthermore, a similar significant effect is identified between *moderate* ($M = -0.279$, $SE = 0.078$) and *strong* partisans ($p = 0.022$), but not between *mild* and *moderate* partisans ($p = 0.334$), as shown in Fig. 3. The plot indicates that those who are ardent supporters of their party are more likely to be influenced by credibility labels compared to participants who do not identify as strongly with their partisan beliefs.

We constructed another model investigating participants' Change in Sharing Intent, with the same predictors as before: Tie Strength, Political Agreement, $\text{Confidence}_{initial}$, Strength of Political Orientation, Political Interest, and News Discussion Frequency. News Discussion Frequency represented participants' baseline sharing behaviour collected at the study's outset: 20 rarely shared news articles, 42 shared them sometimes, and 34 frequently shared news articles on social media platforms, indicating a diversity in sharing propensity. Although credibility labels decreased misperceptions in 39.5% of the instances (Table 2), we found no statistically significant differences between participants' intention to share headlines on their most used social networking and instant messaging platforms before ($M = 2.05$, $SD = 1.25$) and after ($M = 1.93$, $SD = 1.16$) the labels **(RQ2)**.

We constructed a final model investigating participants' Change in Confidence after the credibility indicators, with the same six predictors minus $\text{Confidence}_{initial}$. We also report no statistically significant differences in participants' confidence in their belief after being exposed to the credibility labels.

4.3.3. The backfire effect

Notably, in 13% of cases, credibility labels caused an *increase* in participants' belief in misinformation. This increased belief was more frequent in scenarios of *low* Political Agreement compared to *high*, as illustrated in Fig. 4. Specifically, as depicted in the last row of Table 2, *high* Political Agreement caused an increase in belief in misinformation in 3.1% of cases, while *low* Political Agreement caused belief in misinformation to increase in 22.9% of cases. This increased belief in misinformation after seeing credibility disputes that challenged one's pre-existing partisan beliefs is indicative of the backfire effect **(RQ3)**, and we find that it manifested when the source of the dispute held a different political orientation compared to the receiver. We further explore this finding through our qualitative analysis.

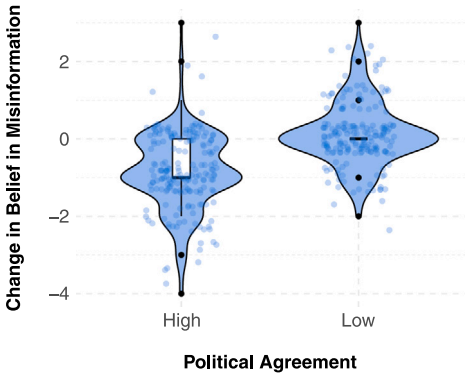

Fig. 4. Violin plot depicting the distribution of changes in belief after the credibility label, comparing Political Agreement levels *high* and *low*. The width of each violin represents the probability density, and the overlaid boxplots provide additional summary statistics. A positive *y*-axis value indicates *increased* belief in misinformation, and vice versa.

Table 2
Distribution of belief change in misinformation after viewing the credibility labels, across the four experimental conditions; TS = Tie Strength, PA = Political Agreement.

		Proportion (%) of cases where participants' belief		
		Decreased	Did not change	Increased
Strong TS	High PA	66.6%	31.2%	2.0%
	Low PA	9.3%	56.2%	34.3%
Weak TS	High PA	60.4%	35.4%	4.1%
	Low PA	21.8%	66.6%	11.4%
Total	High PA	63.5%	33.3%	3.1%
	Low PA	15.6%	61.4%	22.9%
	Total	**39.5%**	**47.3%**	**13.0%**

4.4. Qualitative analysis

At the end of the survey, participants were asked three open-ended questions delving into factors related to their relationship with the tie, specifically exploring aspects of tie strength and political agreement that may have influenced belief change. We were also interested in learning how the presence of credibility disputes influenced participants' intention to share headlines.

To systematically analyse these responses, we employed a deductive thematic analysis approach, following the methodology outlined by Braun and Clarke (2006). We developed a coding framework by defining key themes aligning with our research objectives, which guided our subsequent coding and categorisation process. There were three main themes: how the persuasiveness of peer-supplied credibility labels was influenced by (1) relational closeness and (2) political similarity between the participant and their peer; and (3) why in some cases participants doubled down and embraced their challenged beliefs with increased conviction, i.e., exhibited the backfire effect. We gained a holistic understanding of our qualitative data through multiple readings, and initiated the initial coding phase. This also involved breaking down longer responses into smaller units if they contained multiple themes. Two authors independently coded the transcripts and conducted multiple rounds of coding, labelling responses according to our pre-established themes. Response segments which raised doubts or were ambiguous were highlighted and discussed extensively in consensus meetings between the coders to ensure uniformity and accuracy in the interpretation of the data. This approach allowed us to undertake a rich analysis of our qualitative data grounded in our themes, offering valuable insights into the dynamics of belief change in response to peer-supplied credibility labels.

In the following sections, we present these themes in detail. We first explore the influence of tie strength (and specific factors that compose it, like trustworthiness and expertise), followed by a discussion of political homophily and its persuasiveness. We also report evidence of the backfire effect — discussing the attributes of a source of credibility dispute which may trigger it.

4.4.1. Tie strength

When credibility was disputed by strong ties – individuals with whom one has a close relationship and exhibit trustworthiness – participants reported being willing to reduce their misperceptions, even if it challenged their prior beliefs; *"I trust [tie's] thoughts and experience, so it would at least make me want to confirm my own interpretation of the data being presented in the news. I would be fully willing to admit that I was wrong in this situation."* (P16, *Strong TS × Low PA*), and; *"I'm more likely to trust what [tie] trusts."* (P88, *Strong TS × High PA*).

Furthermore, some participants who displayed no change in belief after seeing the credibility labels reported reduced confidence in their belief, conveying that while credibility labels may have been unsuccessful at reducing their misperceptions, at the very least they sowed some seeds of doubt; *"I usually stuck with my gut feeling, but didn't feel as confident in my answer."* (P49, *Weak TS x High PA*). In contrast, some participants in the weak tie strength condition mentioned either disregarding the credibility label because they could not trust the peer, or worse, using it to affirm their contrasting beliefs. *"I didn't take it into account because I cannot trust [tie's] judgement in either direction."* (P72, *Weak TS × Low PA*) and; *"I don't trust [tie's] judgement so I assumed that the opposite was true."* (P42, *Weak TS × Low PA*).

Notably, in some cases where trust in the source did not exist previously, it was manifested during the course of the experiment when the credibility labels supplied by the source reflected participants' own evaluation of the headlines' credibility; *"Anytime [tie] disputed the credibility of an article which I also disputed, it gave me a more confident feeling as [tie] and my views were the same. However, when [tie] disputed stories which I viewed as true, it made me pause to question if I was really right."* (P71, *Weak TS × High PA*).

Trustworthiness was not the only factor being consulted to estimate credibility, participants also reported source expertise and knowledge as metrics that engendered trust in credibility labels; *"I don't know much about what is going on, but [tie] does, so I trust his judgement."* (P23, *Strong TS × Low PA*) and; *"[Tie] always fact checks, so I'm quick to believe her."* (P19, *Strong TS × High PA*). Further, expertise also influenced individuals who felt confident in their evaluation of the headlines' truthfulness; *"[Tie] is educated, so while I believe in myself and my own stance, I don't think they would discredit something without reason or a good argument so I would be curious as to why they thought that and their reasoning."* (P20, *Weak TS × High PA*). Perceived tie expertise also nudged participants to do independent research; *"Since [tie] is older than me and very knowledgeable about life in general, if she disputed the credibility of a news article I would probably rethink whether or not it was legit and do some further research."* (P48, *Strong TS × High PA*).

Interestingly, participants described the public visibility of the credibility labels to have motivated belief change for them, as the reputation of the source would be at stake if they wrongly disputed headlines in front of their entire social network; *"All articles lost credibility if [tie] disputed them. He is never wrong, and wouldn't state something publicly to be true or false unless he was sure."* (P55, *Weak TS × High PA*), and; *"I felt [tie] had fact checked enough to dispute them publicly."* (P5, *Weak TS × High PA*).

4.4.2. Political agreement

A large majority of participants in the high political agreement condition mentioned that the credibility labels they received made them re-evaluate their beliefs; *"I felt like [tie] and I share similar values and morals, so if she disputed them I would be less confident in my*

initial assertion about the credibility and accuracy of the facts presented in the article." (P54, *High PA × Strong TS*). This phenomenon was also prevalent in cases of weak ties, with participants questioning their (misplaced) beliefs in the headlines; *"It marks a change in perception. Without [tie's] name there I feel more concrete in what I believe to be or not to be the truth. When you add someone into the mix it can shake things up and make you second guess or question it."* (P25, *High PA × Weak TS*).

In contrast, participants in the low political agreement condition expressed feeling unaffected by credibility labels when supplied by hyper-partisan peers of the opposite party, irrespective of partisanship; *"[Tie] is a right wing extremist. I do not respect her choices of political statements."* (P22, *Low PA × Weak TS*). This phenomenon outweighed the influence of expertise; *"I know that [tie] uses fact checkers regularly, but he also believes every news that is left leaning. So I took those two facts into account."* (P2, *High PA × Strong TS*).

Interestingly, a few participants who saw credibility labels from a tie of the opposite party reported feeling selectively influenced by it, based on their assessment of the tie's stance on the headline's content; *"In some cases, thinking about the bias of [tie], it made me change the way I thought about the article, in other cases, I disregarded because I knew of [tie's] bias."* (P27, *Low PA × Strong TS*).

Conversely, some participants felt compelled to re-evaluate their belief even when credibility messages came from members of the opposite party, because they perceived the source as knowledgeable; *"[Tie] is a smart guy, even though we have different beliefs, if he debunked a story I had to think twice about whether it was real."* (P12, *Low PA × Weak TS*).

We also observed evidence for selective exposure, regardless of respondent partisanship, with some expressing greater willingness to accept attitude-congruent disputes irrespective of their veracity, but seeking fact-check rebuttals to disprove news which violated their existing ideologies; *"If a friend I know who is pro-gun control shares a topic on gun violence, I may be more ready to accept it due to my own beliefs in the need for gun-control. But if it was the other way around, I may be more apt to fact check and try and disprove it to maintain my comfort in my own position."* (P92, *Low PA × Weak TS*).

4.4.3. The backfire effect

Analysis of our qualitative data identified two triggers underpinning the backfire effect observed. While we do not claim that these triggers always provoked an increase in misperceptions, we do emphasise that every reported instance of increased belief in misinformation was accompanied by one or both of these triggers.

The first trigger was lack of source expertise — when ties who were not perceived as knowledgeable disputed credibility, some participants reported a perceived increase in accuracy; *"I felt [the headlines] were possibly more credible. [Tie] usually isn't the sharpest tool in the shed."* (P93, *Low PA × Weak TS*), and; *"[Tie] is so little informed that if I saw that she disputed the credibility of something, depending on what it was, I may actually be less inclined to believe [the credibility label], simply because it had her endorsement."* (P66, *Low PA × Strong TS*).

Participants also cited absence of a fact-checking habit as a reason for misperceptions being further entrenched; *"I still thought all of [the headlines] were true despite what [tie] had to say about him. He never fact checks anything so if he thinks something is fake, it's probably true."* (P9, *Low PA × Weak TS*).

The second factor was the source being perceived as extremely partisan, with participants indicating that credibility disputes by such polarised ties further increased their belief in the (fake) news headlines, substantiating the backfire effect; *"If [tie] thinks something is true then it is probably the opposite. He just parrots anything they say on Fox."* (P1, *Low PA × Strong TS*), and; *"I just knew [tie] was biased a certain way, so it made me think even more differently than him."* (P10, *Low PA × Strong TS*).

5. Findings and implications

Thus far, the majority of efforts to moderate misinformation in online environments have centred around algorithmic interventions and their variants, such as platform-supplied veracity labels and indicators. Although more nuanced than the blunt "leave up/take down" approach to unverified posts, there exists mixed evidence of their effectiveness in reducing misperceptions (Oeldorf-Hirsch et al., 2020). Our study extended previous work by examining credibility labels that are supplied by sources other than the platform — members of one's social media network. To systematically examine how such credibility labels can reduce misperceptions, and understand the specifics of the underlying social structure that could trigger belief change, we took a novel approach. We investigated scenarios in which both the relational closeness and the political agreement with the source of a credibility label were at play. Overall, this study examined the effectiveness of credibility messages applied to attitude-congruent misinformation, when supplied by ties at different combinations of high vs. low political agreement and strong vs. weak tie strength.

In the following sections, we discuss the effects of tie strength and political agreement as revealed by our analyses, explore the persuasive impact of credibility disputes at different combinations of political agreement and tie strength, and highlight the unintended consequences that may arise when trying to offer credibility labels in a politically charged environment. Based on our findings, we also discuss potential political homophily based approaches to mitigate misinformation on social media platforms.

5.1. Tie strength and related (latent) factors that persuade belief change

Although corrections that oppose one's deeply held beliefs and ideologies are difficult to accept (Thorson, 2016), literature strongly suggests that new information offered by one's close ties – those perceived as being *trustworthy* and exhibiting *goodwill* towards the receiver – may appear more credible (Ecker et al., 2022; Metzger et al., 2010). However, contrary to expectations, our statistical analyses did not indicate a significant difference between belief reduction caused by *weak* vs. *strong* ties **(RQ1)**.

This observed inability of strong ties to persuade greater belief change than weak ties resonates with a recent study by Schaewitz et al. (2022), who found no evidence of disinformation being perceived as more credible when forwarded by strong ties compared to weak ties. However, their study investigates belief change in the *opposite* direction as ours — analysing whether stronger tie strength manifests *greater* belief in misinformation, while we investigate whether stronger tie strength promotes *decreased* belief in misinformation. Thus, our findings provide evidence towards a different perspective on the debiasing potential of tie strength – even though close ties are regarded as trustworthy, they may not be more persuasive than weak ties in influencing belief in misinformation in either direction – be it a reinforcement or a reduction.

There are multiple plausible explanations for the lack of impact of tie strength on belief change in our study. Firstly, and perhaps most importantly, the causal relationship between perceived trustworthiness of a source and its persuasive ability may be oversimplifying several important determinants of belief formation. While an individual may perceive a strong tie as being more trustworthy, this trust may not necessarily translate into agreement with these ties' evaluation of the veracity of information, especially when it is politically charged and concordant with existing (misinformed) beliefs, as was in our study. Other characteristics of the source, in addition to relational closeness, such as perceived intelligence, domain expertise (Metzger et al., 2010), political literacy, and awareness of worldly matters can also influence information reception (Wyer and Albarracín, 2005), and thus may dampen the persuasiveness of ties. This was reflected in our qualitative findings, with participants mentioning more knowledgeable

sources as more persuasive, even if they were relationally distant. This finding contradicts prior research which advocates that trustworthiness may be more influential than expertise when attempting to correct beliefs (Guillory and Geraci, 2013). We posit that expertise and knowledgeability of the source of a credibility dispute should not be perceived separate from trustworthiness, but rather as drivers of enough trustworthiness in another individual's judgement, which can influence one's beliefs.

Moreover, in our qualitative analysis we observed interesting evidence of the inverse — trust in weak ties sometimes manifested *temporarily* during our study, when the displayed credibility labels aligned with participants' own assessment of the headlines. This finding further suggests that relational closeness may not always be an indicator of trustworthiness, and we argue that trust in a tie alone may not be sufficient to revert deeply held beliefs. Future work is necessary to disentangle the persuasiveness impacts of these personal characteristics, and explore how they may be leveraged to modulate peer-supplied credibility disputes to be more effective on social media platforms.

Nevertheless, our qualitative analysis indicates that the presence of credibility labels from strong ties encouraged even those with firm beliefs in misinformation to conduct independent research to assess the veracity of the information they encountered. On social media platforms, such interventions that promote fact-checking, especially amongst those who may not be willing to do so themselves in the first place, should not be ruled out as completely ineffective. If social media platforms could highlight the expertise and knowledgeability of the peer raising a credibility dispute, belief in misinformation could be fact-checked even by relationally distant contacts. Future work can examine if more complex and collaborative peer-debunking efforts, such as those arising from multiple relationally close contacts, with varying domain expertise, may be more effective than individual disputes.

5.2. The role of attitudinal homophily in fighting misinformation

Social influence literature posits that messages received by members of one's in-group are more likely to be systematically processed (Chaiken, 1987). Thus, when providing credibility disputes, source-receiver attitudinal homophily – common ideologies from having a shared political identity (one's *in-group*) – could bring about a reduction in misperceptions (Mackie et al., 1990). Our results provide empirical evidence to this notion, as we observed a significant difference between high and low levels of political agreement in causing belief change **(RQ1)**. More specifically, when credibility was disputed by a member of the same political party, participants were more likely to re-evaluate their belief in the misinformation, irrespective of the level of relationship closeness. Even though participants were deliberately shown misinformation that reinforced their partisan beliefs, they stated that the source of the credibility label having values and morals similar to their own motivated them to re-evaluate their beliefs. Therefore, these results are not only in accordance with prior literature, but also extend it by demonstrating that credibility labels from in-group members can overcome motivated reasoning and reduce selection bias.

Furthermore, our study presented credibility-disputing labels by an individual's co-partisan, to false news that should be concordant with the ideologies and biases of both the individual and their co-partisan who forms the source of the label. Such an unexpected scenario may be generating more interest during information evaluation, when an individual sees a co-partisan *"speak against their self-interests"* (Pasquetto et al., 2022; Baum and Groeling, 2009) and dispute the accuracy of (mis)information which they both should agree with. This phenomenon is reflected in our results with an overwhelming difference in belief reduction caused by ties with high political agreement (63.5% of participants) compared to those with low political agreement (15.6% of participants), as shown in Table 2.

This finding presents several important implications for platforms to combat misinformation. Social media echo chambers, especially those with partisan undercurrents, either created explicitly through groups such as on Facebook, or induced implicitly by selectively engaging with like-minded individuals, can benefit substantially from these findings. When members of such politically-charged subnetworks share their credibility evaluation with others of the same group, it may cause a greater reduction in belief in misinformation as compared to when credibility is disputed by someone from outside the group. Therefore, fact-checks on social networking sites may be more believable when shared by connections who also happen to be co-partisans. Further, the finding that high political agreement persuades belief change irrespective of tie strength also generates implications for such online political groups. These groups represent a venue where users may share a high level of political agreement with each other, but may not necessarily exhibit relational closeness with one another, thus fostering relationships with both strong and weak ties. When individuals whose beliefs coincide on most topics present a fact-check to one another on social media, irrespective of relational closeness, they may be more persuaded to re-evaluate their beliefs in online misinformation.

It is also worthwhile to acknowledge the real-world challenges of co-partisan fact-checking: individuals may be more likely to fact-check counter-partisans than co-partisans, as observed during analyses of user behaviours on Twitter's crowd sourced fact-checking program, Birdwatch (Allen et al., 2022). However, unlike Birdwatch where interactions occur amongst internet strangers, our study examined the dynamics of fact-checking within existing social relationships. We analysed how belief in misinformation is influenced within social networks where individuals know the source raising a credibility dispute and have relational ties with them, whether close or distant. We believe that the factors causing peer disputes to be taken seriously by our participants may also be the very factors that encourage individuals to raise such disputes in the first place. Factors such as a desire for goodwill (Lupia and McCubbins, 1998; McCroskey and Teven, 1999) can play a pivotal role in making individuals more receptive to fact-checks within their social circles, and possibly also more willing to offer them as a form of *social responsibility*. This notion is supported by the finding that users also take upon content moderation and "flag" problematic content due to non-partisan motivations, such as protecting oneself and one's peers from the harms stemming from non-credible content (Zhang et al., 2023).

This sense of *community cleaning* and relational familiarity may enhance the willingness of individuals to fact-check and engage in constructive dialogue, especially within close relationships and possibly across partisan lines, much more than what is observed on politically-driven fact-checking programs like Birdwatch. For example, it is plausible that one would feel more inclined to fact-check one's distant relative with differing political beliefs, than fact-check a complete stranger, particularly when this desire to fact-check stems from a place of goodwill and social responsibility rather than political agendas.

Overall, our findings highlight the potential efficacy of co-partisan fact-checking in mitigating belief in misinformation. Although we acknowledge the challenges of co-partisan fact-checking, the significance of our findings lies in the fact that the sociotechnical infrastructures underlying social media platforms *can* and *should* be harnessed to counteract the spread of misinformation. It is thus an important avenue for future research to investigate such interventions in the wild, as well as design platform affordances and mechanisms that encourage individuals, including co-partisans, to fact-check one another.

5.3. Peer-supplied credibility labels have little effect on sharing intentions

It is interesting to note that while credibility labels supplied by co-partisans lowered participants' belief in misinformation, we did not find them to influence participants' intentions to share articles with their social media networks. This observation is supported by recent

research which reports a disconnect between accuracy perceptions and sharing intentions (Sirlin et al., 2021; Epstein et al., 2023), with individuals disregarding the perceived accuracy of a headline when making decisions about sharing the headline (Pennycook et al., 2020). One possible explanation is rooted in the *attention-based account* proposed by Pennycook et al. (2021b). According to this framework, individuals possess a general inclination to avoid spreading inaccurate information. However, when considering sharing articles in a social media context, the influence of the platform may redirect their attention away from accuracy-related considerations, emphasising factors such as partisan alignment (Pennycook et al., 2020), and validation and reinforcement from their social media peers (Brady et al., 2020a; Crockett, 2017).

In our study, credibility labels successfully captured participants' attention and facilitated a reassessment of headline accuracy, thus explaining the observed influence on their belief in misinformation. However, participants may nevertheless have been willing to share these articles primarily to fulfil social needs rather than informational ones, perceiving sharing articles as a mode of self-expression (Chen et al., 2015; Jahanbakhsh et al., 2022) and as a means of soliciting others' opinions on the subject and expressing their own (Chen and Sin, 2013), overshadowing the influence of our credibility labels. Future work can explore the effectiveness of peer-supplied indicators tailored to also highlight the potential consequences of sharing misinformation, and examine whether they can more efficiently mitigate the sharing of non-credible content.

5.4. Belief change in a direction unintended: The backfire effect

Despite being a challenging phenomenon to reproduce (Weeks, 2015; Wood and Porter, 2019), our study manifested a situation in which the backfire effect was observed, following other studies that elicited this effect (Nyhan and Reifler, 2010). We provided participants with misinformation that would be more likely to align with their beliefs, based on their partisanship, and presented credibility disputes by both members of the same and opposite party. For some participants, being exposed to credibility disputes that challenged their worldview caused them to increase their beliefs in the misinformation presented. Our results indicate that credibility labels supplied by members of the opposite party, one's *out-group*, caused this effect more than members of one's own party **(RQ3)**. More specifically, 22.9% of individuals who saw credibility labels by members of the opposite party reported an increased belief in misinformation, while only 3.1% of those receiving labels from co-partisans reported reinforced belief (Table 2).

Participants cited their tie's perceived *lack of expertise* – either through their knowledge being limited or through poor fact-checking habits – as one of the reasons for strongly embracing their (wrong) belief when challenged by one's ties. This rationale seems sensible – credibility evaluation performed by seemingly uninformed individuals may not be trusted, and may promote further belief in the misinformation. More importantly, participants also reported the *hyper-partisan* nature of their named tie as a reason to discard their credibility evaluation, and enforce their belief in (mis)information. In other words, participants first seeing misinformation that aligns with their belief system, and then seeing it be challenged by hyper-polarised members of the opposite party, had their (incorrect) perceptions bolstered. This resonates with prior literature on the backfire effect, which hypothesises that in certain situations, fact-checking messages that challenge one's deeply held beliefs can be perceived as a threat to one's intellectual autonomy, and to the very truths that ground one's identity (Brehm, 1966; Byrne and Hart, 2009), thus influencing belief change in an unintended direction.

Interestingly, Nyhan and Reifler (Nyhan and Reifler, 2010) found the backfire effect to exist when corrections were presented to hyper-partisan individuals (who were more inclined to reject them). Our study does not make this distinction, and instead presents evidence of the backfire effect in relation to the *source* being hyperpartisan, not the *receiver*. Future work is required to fully understand how peer-supplied credibility disputes can cause unintended consequences based on the dynamics of deeply held belief systems and the partisan nature of both the receiver and the source. Nevertheless, platforms should not evade correcting misinformation fearing a backfire effect.

5.5. Limitations and future work

We highlight the following limitations of our study. First, we recognise that the validity of our results relies on our respondents' ability to elicit names that match the tie strength (TS) and political agreement (PA) required by their assigned experimental condition. It is possible that when asked to name a specific tie, participants may have chosen a person that did not fully fit within the given description. To account for non-compliance with the allocated treatment, we posed a set of questions towards the end of the survey that enquired about participants' perceived strength of the two manipulations: relationship closeness and level of political agreement with the nominated tie. We inferred the actual TS and PA by analysing participants' responses to this post-survey questionnaire. Although we rejected responses with a mismatch between the assigned and actual measurement of TS and PA, and continued recruiting more participants until we reached a set of valid 96 responses, there is still a small possibility that the self-reported relationship measurements at the end of the survey may have been unrepresentative of the truth. Moreover, due to the subjective nature of what constitutes a weak or strong tie (Granovetter, 1973), it is plausible that different participants may have had different baselines against which to evaluate tie strength, leading to inconsistencies. Nevertheless, this approach was essential to our overarching goals — it enabled us to test the effectiveness of credibility indicators in a more naturalistic sense, where participants' own peers, varying in relationship closeness and political agreement, were alerting them to the veracity of misinformation.

Second, rather than examining the sole impact of political orientations, we adopted a more holistic concept of alignment between an individual's political orientation and that of the peer providing the credibility dispute. This wider perspective of congruence allows for greater applicability of our findings across various political stances, beyond just Republicans and Democrats. Despite this, our study may not fully capture the attitudes and behaviours of more diverse partisans, or non-partisans. Further research is needed to examine the generalisability of our findings to participants with more diverse political identities.

Third, our labels were designed and evaluated in a platform-agnostic way, and do not replicate a real-world social media platform. Such a platform-agnostic approach allowed us to avoid any moderating influences that may be exerted by platforms and their affordances. We deemed this necessary to enable us to isolate the effect of our credibility labels on belief in news headlines. Moreover, previous research has found no significant difference in the effectiveness of social corrections offered on Twitter versus Facebook, even if the influence of including sources in social corrections did differ between the two platforms (Vraga and Bode, 2018). Nevertheless, we acknowledge the potential role of the platform and its affordances in influencing the acceptance of credibility disputes from peers, and further research is needed to comprehensively investigate this.

Further, our study examined a specific phrasing for peer-supplied credibility disputes. Alternate formulations, such as naming multiple known online peers, could influence the persuasiveness of credibility labels. Future research should explore diverse wordings of credibility disputes to better understand their effects on belief change.

While it was necessary to present participants with attitude-congruent misinformation spanning diverse topics to ensure methodological rigour, the choice of topics could potentially impact the study outcomes. Moreover, different types of misinformation (e.g., health, policy, electoral, statistical, etc.) may interact with participants' prior

Table B.3
Participant demographic data.

Demographic data	Participant distribution
Age	8–24 years old (n = 22), 25–34 years old ($\boldsymbol{n}$ = **31**), 35–44 years old ($\boldsymbol{n}$ = **15**), 45–54 years old ($\boldsymbol{n}$ = **15**), 55–64 years old ($\boldsymbol{n}$ = **10**), 65+ years old ($\boldsymbol{n}$ = **3**)
Gender	Male ($\boldsymbol{n}$ = **40**), Female ($\boldsymbol{n}$ = **55**), Non-binary ($\boldsymbol{n}$ = **1**), Prefer not to say ($\boldsymbol{n}$ = **0**)
Highest Education	Some high school or less ($\boldsymbol{n}$ = **2**), High school diploma or GED ($\boldsymbol{n}$ = **15**), Some college but no degree ($\boldsymbol{n}$ = **24**), Associates or technical degree ($\boldsymbol{n}$ = **11**), Bachelor's degree ($\boldsymbol{n}$ = **33**), Graduate or professional degree ($\boldsymbol{n}$ = **11**), Prefer not to say ($\boldsymbol{n}$ = **0**)
Employment	Employed full-time ($\boldsymbol{n}$ = **40**), Employed part-time ($\boldsymbol{n}$ = **15**), Self-employed ($\boldsymbol{n}$ = **10**), Unemployed but looking for a job ($\boldsymbol{n}$ = **10**), Unemployed and not looking for a job ($\boldsymbol{n}$ = **4**), Full-time parent/homemaker ($\boldsymbol{n}$ = **5**), Retired ($\boldsymbol{n}$ = **3**), Student ($\boldsymbol{n}$ = **9**), Military ($\boldsymbol{n}$ = **0**)

beliefs differently, and may elicit varying magnitudes of belief change, perhaps also across partisan lines. Although we presented participants with attitude-congruent misinformation, future work is necessary to investigate how the effectiveness of corrective efforts can be influenced by different themes of misinformation.

Lastly, we recognise that our study trialled a best-case scenario for corrective practices — as misinformation was corrected immediately after it was presented to participants, albeit after a short distractor task. In a real-life context, this may not be the case due to the fragmented and engagement-driven nature of social media, as it is possible that some individuals may come across a post containing misinformation before it has been disputed by their peers.

6. Conclusion

Misinformation can spread like wildfire on social media platforms and continue to circulate unchecked through news feeds, often driven by individual biases and partisan predilections. Nevertheless, the unique affordances of social media platforms can make them a suitable venue for corrective operations. In this paper, we investigate whether the social capital offered by such platforms, operationalised through relationship closeness and level of political agreement between social media contacts, can be leveraged to design effective credibility labels that tap into deeply held (misinformed) beliefs, and persuade change. We selected news headlines containing misinformation more likely to be believed based on the partisanship of respondents, and displayed credibility labels which we designed to look as if they had been supplied by the respondents' named peer. We tested the effectiveness of these labels in four different experimental conditions by manipulating two variables: tie strength (*weak* vs. *strong*) and political agreement (*low* vs. *high*) with the peer. We found evidence that when individuals come across credibility disputes by members of the same political party as them, they felt persuaded to significantly reduce their belief in even attitude-congruent misinformation, which is perhaps the most difficult to debunk, owing to our tendency to maintain existing beliefs. Although we did not find a significant influence of tie strength on belief change, we identified that credibility labels by close peers did reduce confidence in one's evaluation of the headline veracity, and that perceived source expertise and knowledgability may trump relational closeness. We compare our results with prior research on the persuasiveness of political homophily and relationship closeness on influencing attitudes, and discuss how our findings can guide the design of social networking sites to combat misinformation. Lastly, we emphasise that research to combat misinformation in online environments which explores social connections and the persuasiveness of interpersonal relationships remains in its infancy. More work is needed to disentangle the various determinants of it as identified in this study, and to generate a better understanding of how they influence belief formation and revision in the fast-paced nature of information consumption in social media feeds.

CRediT authorship contribution statement

Saumya Pareek: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jorge Goncalves:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. News articles

A.1. For Republican respondents

1. "Three doctors from the same hospital 'die suddenly' in the same week," after the hospital mandated a fourth COVID-19 vaccine for employees.
2. 2020 Election fraud in Pennsylvania: "Pennsylvania sent out 1, 823,148 mail-in ballots. They got back 2,589,242."
3. New ID cards for illegal immigrants: "Biden wants to give illegals ID cards so they can start collecting American benefits."
4. "Democrats' new army of 87,000 IRS agents will be coming for you --- with 710,000 new audits for Americans who earn less than $75k."

A.2. For democratic respondents

1. Biden: "The typical elementary school teacher pays $7239 in federal income taxes, and firefighters pay $5328, while Donald Trump paid $750."
2. "If every person on earth just recycled, stopped using plastic straws, and drove an electric car, 100 corporations would still produce 70% of total global emissions."
3. "Trump had to confess in writing, in court, to illegally diverting charitable contributions that were supposed to go to veterans."
4. "50% of the guns sold in Texas, because of the loopholes, do not pass through a background check."

Appendix B. Participant demographic data

See Table B.3.

Table C.4
Text prompt displayed to participants at the beginning of the survey to elicit the name of a peer, with specific characteristics of their relationship outlined based on participants' assigned condition.

Political agreement	Tie strength	Prompt to participant
High	Strong	Please enter the first name of a person that shares the **same political beliefs as you**, and is **close to you**. This could be a family member, a friend, or a coworker, for example. This individual must be someone you know personally, not a politician.
Low	Strong	Please enter the first name of a person that shares **different political beliefs than you**, and is **close to you**. This could be a family member, a friend, or a coworker, for example. This individual must be someone you know personally, not a politician.
High	Weak	Please enter the first name of a person that shares the **same political beliefs as you**, and is **not close to you**. This could be a **distant** family member, a friend, or a coworker, for example. This individual must be someone you know personally, not a politician.
Low	Weak	Please enter the first name of a person that shares **different political beliefs than you**, and is **not close to you**. This could be a **distant** family member, a friend, or a coworker, for example. This individual must be someone you know personally, not a politician.

Appendix C. Survey details

Link to the survey for `high Political Agreement, strong Tie Strength` condition: https://melbourneuni.au1.qualtrics.com/jfe/form/SV_8iBsddlYznCxnU2. The survey format for the remaining three conditions is identical, except for changes in the characteristics of the participants' relationship with the peer whose name we elicit. The prompts presented to participants to obtain these names are presented in Table C.4.

References

Acerbi, A., 2019. Cognitive attraction and online misinformation. Palgrave Commun. 5 (1), 1–7. http://dx.doi.org/10.1057/s41599-019-0224-y, URL https://www.nature.com/articles/s41599-019-0224-y.

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. J. Econ. Perspect. 31 (2), 211–236. http://dx.doi.org/10.1257/jep.31.2.211, URL https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

Allen, J., Martel, C., Rand, D.G., 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's birdwatch crowdsourced fact-checking program. In: CHI Conference on Human Factors in Computing Systems. CHI '22, ACM, http://dx.doi.org/10.1145/3491102.3502040.

Amazeen, M., Krishna, A., 2020. Correcting Vaccine Misinformation: Recognition and Effects of Source Type on Misinformation via Perceived Motivations and Credibility. Rochester, NY, http://dx.doi.org/10.2139/ssrn.3698102, URL https://papers.ssrn.com/abstract=3698102.

Au, C.H., Ho, K.K.W., Chiu, D.K., 2022. The role of online misinformation and fake news in ideological polarization: Barriers, catalysts, and implications. Inf. Syst. Front. 24 (4), 1331–1354. http://dx.doi.org/10.1007/s10796-021-10133-9.

Axt, J.R., Landau, M.J., Kay, A.C., 2020. Fake news attributions as a source of nonspecific structure. In: The Psychology of Fake News. Routledge, pp. 220–234. http://dx.doi.org/10.4324/9780429295379-15.

Balmas, M., 2014. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. Commun. Res. 41 (3), 430–454. http://dx.doi.org/10.1177/0093650212453600.

Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R., 2015. Tweeting from left to right: Is online political communication more than an echo chamber? Psychol. Sci. 26 (10), 1531–1542. http://dx.doi.org/10.1177/0956797615594620.

Barrera, O., Guriev, S., Henry, E., Zhuravskaya, E., 2020. Facts, alternative facts, and fact checking in times of post-truth politics. J. Public Econ. 182, 104123. http://dx.doi.org/10.1016/j.jpubeco.2019.104123.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models Usinglme4. J. Stat. Softw. 67 (1), http://dx.doi.org/10.18637/jss.v067.i01.

Baum, M.A., Groeling, T., 2009. Shot by the messenger: Partisan cues and public opinion regarding national security and war. Political Behav. 31 (2), 157–186. http://dx.doi.org/10.1007/s11109-008-9074-9.

Brady, W.J., Crockett, M.J., Van Bavel, J.J., 2020a. The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. Perspect. Psychol. Sci. 15 (4), 978–1010. http://dx.doi.org/10.1177/1745691620917336.

Brady, W.J., Gantman, A.P., Van Bavel, J.J., 2020b. Attentional capture helps explain why moral and emotional content go viral. J. Exp. Psychol. [Gen.] 149 (4), 746–756. http://dx.doi.org/10.1037/xge0000673.

Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., Van Bavel, J.J., 2017. Emotion shapes the diffusion of moralized content in social networks. Proc. Natl. Acad. Sci. USA 114 (28), 7313–7318. http://dx.doi.org/10.1073/pnas.1618923114.

Braun, V., Clarke, V., 2006. Using thematic analysis in psychology. Qual. Res. Psychol. 3 (2), 77–101. http://dx.doi.org/10.1191/1478088706qp063oa, URL https://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa.

Brehm, J.W., 1966. A Theory of Psychological Reactance. Academic Press, Oxford, England.

Brown, J., 1958. Some tests of the decay theory of immediate memory. Q. J. Exp. Psychol. 10 (1), 12–21. http://dx.doi.org/10.1080/17470215808416249.

Budak, C., Agrawal, D., El Abbadi, A., 2011. Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web. WWW '11, Association for Computing Machinery, New York, NY, USA, pp. 665–674. http://dx.doi.org/10.1145/1963405.1963499.

Byrne, S., Hart, P.S., 2009. The boomerang effect a synthesis of findings and a preliminary theoretical framework. Ann. Int. Commun. Assoc. 33 (1), 3–37. http://dx.doi.org/10.1080/23808985.2009.11679083.

Campbell, K.E., Marsden, P.V., Hurlbert, J.S., 1986. Social resources and socioeconomic status. Social Networks 8 (1), 97–117. http://dx.doi.org/10.1016/S0378-8733(86)80017-X, URL https://www.sciencedirect.com/science/article/pii/S037887338680017X.

Chaiken, S., 1987. The heuristic model of persuasion. In: Social Influence: the Ontario Symposium, Vol. 5. In: Ontario symposium on personality and social psychology, Lawrence Erlbaum Associates Inc, Hillsdale, NJ, US, pp. 3–39.

Chakrabarti, S., Stengel, L., Solanki, S., 2018. Duty, identity, credibility: Fake news and the ordinary citizen in India. BBC World Service Audiences Research.

Chen, X., Sin, S.-C.J., 2013. 'Misinformation? What of it?': Motivations and individual differences in misinformation sharing on social media. In: Proceedings of the 76th ASIS&T Annual Meeting: beyond the Cloud: Rethinking Information Boundaries. ASIST '13, American Society for Information Science, USA, pp. 1–4.

Chen, X., Sin, S.-C.J., Theng, Y.-L., Lee, C.S., 2015. Why do social media users share misinformation? In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '15, Association for Computing Machinery, New York, NY, USA, pp. 111–114. http://dx.doi.org/10.1145/2756406.2756941.

Cohen, J., 1992. A power primer. Psychol. Bull. 112 (1), 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155.

Crockett, M.J., 2017. Moral outrage in the digital age. Nat. Hum. Behav. 1 (11), 769–771. http://dx.doi.org/10.1038/s41562-017-0213-3.

Dias, N., Pennycook, G., Rand, D.G., 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. Harvard Kennedy Sch. Misinf. Rev. 1 (1), http://dx.doi.org/10.37016/mr-2020-001, URL https://misinforeview.hks.harvard.edu/article/emphasizing-publishers-does-not-reduce-misinformation/.

Dibble, J.L., Levine, T.R., Park, H.S., 2012. The unidimensional relationship closeness scale (URCS): Reliability and validity evidence for a new measure of relationship closeness. Psychol. Assess. 24 (3), 565–572. http://dx.doi.org/10.1037/a0026265.

Ecker, U.K.H., Ang, L.C., 2019. Political attitudes and the processing of misinformation corrections. Political Psychol. 40 (2), 241–260. http://dx.doi.org/10.1111/pops.12494, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12494.

Ecker, U.K.H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L.K., Brashier, N., Kendeou, P., Vraga, E.K., Amazeen, M.A., 2022. The psychological drivers of misinformation belief and its resistance to correction. Nat. Rev. Psychol. 1 (1), 13–29. http://dx.doi.org/10.1038/s44159-021-00006-y, URL https://www.nature.com/articles/s44159-021-00006-y.

Ecker, U.K.H., Lewandowsky, S., Swire, B., Chang, D., 2011. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. Psychon. Bull. Rev. 18 (3), 570–578. http://dx.doi.org/10.3758/s13423-011-0065-1.

Epstein, Z., Pennycook, G., Rand, D., 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–11. http://dx.doi.org/10.1145/3313831.3376232.

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., Rand, D., 2023. The social media context interferes with truth discernment. Sci. Adv. 9 (9), http://dx.doi.org/10.1126/sciadv.abo6169.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191. http://dx.doi.org/10.3758/bf03193146.

Festinger, L., 1962. A Theory of Cognitive Dissonance. Stanford University Press.

Flintham, M., Karner, C., Bachour, K., Creswick, H., Gupta, N., Moran, S., 2018. Falling for fake news: Investigating the consumption of news via social media. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, Association for Computing Machinery, New York, NY, USA, pp. 1–10. http://dx.doi.org/10.1145/3173574.3173950.

Friggeri, A., Adamic, L., Eckles, D., Cheng, J., 2014. Rumor cascades. In: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 8, No. 1. pp. 101–110. http://dx.doi.org/10.1609/icwsm.v8i1.14559, URL https://ojs.aaai.org/index.php/ICWSM/article/view/14559.

Fulton, M., Edge, C., Sattar, J., 2022. Robot communication via motion: A study on modalities for robot-to-human communication in the field. ACM Trans. Human-Robot Interaction 11 (2), 1–40. http://dx.doi.org/10.1145/3495245.

Gao, M., Xiao, Z., Karahalios, K., Fu, W.-T., 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. Proc. ACM Human-Comput. Interaction 2 (CSCW), 55:1–55:16. http://dx.doi.org/10.1145/3274324.

Garrett, R.K., Nisbet, E.C., Lynch, E.K., 2013. Undermining the corrective effects of media-based political fact checking? The role of contextual cues and Naïve theory. J. Commun. 63 (4), 617–637. http://dx.doi.org/10.1111/jcom.12038.

Garrett, R.K., Weeks, B.E., 2013. The promise and peril of real-time corrections to political misperceptions. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. CSCW '13, ACM, http://dx.doi.org/10.1145/2441776.2441895.

Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., Roy, D., 2018. Me, my echo chamber, and I: Introspection on social media polarization. In: Proceedings of the 2018 World Wide Web Conference. WWW '18, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 823–831. http://dx.doi.org/10.1145/3178876.3186130.

Granovetter, M.S., 1973. The strength of weak ties. Am. J. Sociol. 78 (6), 1360–1380, URL https://www.jstor.org/stable/2776392.

Guillory, J.J., Geraci, L., 2013. Correcting erroneous inferences in memory: The role of source credibility. J. Appl. Res. Memory Cogn. 2 (4), 201–209. http://dx.doi.org/10.1016/j.jarmac.2013.10.001, URL https://www.sciencedirect.com/science/article/pii/S2211368113000752.

Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A., 2013. Faking sandy: characterizing and identifying fake images on Twitter during hurricane sandy. In: Proceedings of the 22nd International Conference on World Wide Web. In: WWW '13 Companion, Association for Computing Machinery, New York, NY, USA, pp. 729–736. http://dx.doi.org/10.1145/2487788.2488033.

Haque, M.M., Yousuf, M., Alam, A.S., Saha, P., Ahmed, S.I., Hassan, N., 2020. Combating misinformation in Bangladesh: Roles and responsibilities as perceived by journalists, fact-checkers, and users. Proc. ACM Human-Comput. Interaction 4 (CSCW2), 130:1–130:32. http://dx.doi.org/10.1145/3415201.

Henry, E., Zhuravskaya, E., Guriev, S., 2020. Checking and sharing alt-facts. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.3597191.

Hornsey, M.J., Fielding, K.S., 2017. Attitude roots and jiu jitsu persuasion: Understanding and overcoming the motivated rejection of science. Am. Psychol. 72 (5), 459–473. http://dx.doi.org/10.1037/a0040437.

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., DeShon, R.P., 2012. Detecting and deterring insufficient effort responding to surveys. J. Bus. Psychol. 27 (1), 99–114. http://dx.doi.org/10.1007/s10869-011-9231-8.

Jahanbakhsh, F., Katsis, Y., Wang, D., Popa, L., Muller, M., 2023. Exploring the use of personalized AI for identifying misinformation on social media. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, ACM, http://dx.doi.org/10.1145/3544548.3581219.

Jahanbakhsh, F., Zhang, A.X., Karger, D.R., 2022. Leveraging structured trusted-peer assessments to combat misinformation. Proc. ACM Human-Comput. Interaction 6 (CSCW2), 524:1–524:40. http://dx.doi.org/10.1145/3555637.

Jakesch, M., Koren, M., Evtushenko, A., Naaman, M., 2018. The role of source, headline and expressive responding in political news evaluation. SSRN Electron. J. http://dx.doi.org/10.2139/ssrn.3306403.

Kang, H., Bae, K., Zhang, S., Sundar, S.S., 2011. Source cues in online news: Is the proximate source more powerful than distal sources? Journalism Mass Commun. Q. 88 (4), 719–736. http://dx.doi.org/10.1177/107769901108800403.

Karlova, N., Fisher, K., 2013. A social diffusion model of misinformation and disinformation for understanding human information behaviour. Inf. Res. 18.

Kemp, S., 2022. Digital 2022 July Global Statshot Report. DataReportal – Global Digital Insights, URL https://datareportal.com/reports/digital-2022-july-global-statshot.

Kraft, P.W., Lodge, M., Taber, C.S., 2015. Why people "don't trust the evidence": Motivated reasoning and scientific beliefs. Ann. Am. Acad. Political Soc. Sci. 658 (1), 121–133. http://dx.doi.org/10.1177/0002716214554758.

LaPaglia, J.A., Wilford, M.M., Rivard, J.R., Chan, J.C.K., Fisher, R.P., 2013. Misleading suggestions can alter later memory reports even following a cognitive interview. Appl. Cogn. Psychol. 28 (1), 1–9. http://dx.doi.org/10.1002/acp.2950.

Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L., 2018. The science of fake news. Science 359 (6380), 1094–1096. http://dx.doi.org/10.1126/science.aao2998, URL https://www.science.org/doi/10.1126/science.aao2998.

Lewandowsky, S., Ecker, U.K.H., Seifert, C.M., Schwarz, N., Cook, J., 2012. Misinformation and its correction: Continued influence and successful debiasing. Psychol. Sci. Public Interest 13 (3), 106–131. http://dx.doi.org/10.1177/1529100612451018.

Lewandowsky, S., Oberauer, K., 2016. Motivated rejection of science. Curr. Dir. Psychol. Sci. 25 (4), 217–222. http://dx.doi.org/10.1177/0963721416654436.

Lupia, A., McCubbins, M.D., 1998. The democratic dilemma: Can citizens learn what they need to know?. In: Political economy of institutions and decisions, Cambridge University Press, Cambridge, U.K. ; New York.

Mackie, D.M., Worth, L.T., Asuncion, A.G., 1990. Processing of persuasive in-group messages. J. Personal. Soc. Psychol. 58, 812–822. http://dx.doi.org/10.1037/0022-3514.58.5.812.

Marsden, P.V., Campbell, K.E., 1984. Measuring tie strength. Social Forces 63 (2), 482–501. http://dx.doi.org/10.1093/sf/63.2.482.

McCroskey, J.C., Teven, J.J., 1999. Goodwill: A reexamination of the construct and its measurement. Commun. Monogr. 66 (1), 90–103. http://dx.doi.org/10.1080/03637759909376464.

McGinnies, E., Ward, C.D., 1980. Better liked than right: Trustworthiness and expertise as factors in credibility. Pers. Soc. Psychol. Bull. 6 (3), 467–472. http://dx.doi.org/10.1177/014616728063023.

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. http://dx.doi.org/10.1002/poi3.214, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.214.

Messing, S., Westwood, S.J., 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. Commun. Res. 41 (8), 1042–1063. http://dx.doi.org/10.1177/0093650212466406.

Metzger, M.J., Flanagin, A.J., Medders, R.B., 2010. Social and heuristic approaches to credibility evaluation online. J. Commun. 60 (3), 413–439. http://dx.doi.org/10.1111/j.1460-2466.2010.01488.x.

Morozov, E., 2009. Swine flu: Twitter's power to misinform. Foreign Policy URL https://foreignpolicy.com/2009/04/25/swine-flu-twitters-power-to-misinform/.

Morrow, G., Swire-Thompson, B., Polny, J.M., Kopec, M., Wihbey, J.P., 2022. The emerging science of content labeling: Contextualizing social media content moderation. Journal of the Association for Information Science and Technology 73 (10), 1365–1386. http://dx.doi.org/10.1002/asi.24637, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24637.

Nadarevic, L., Reber, R., Helmecke, A.J., Köse, D., 2020. Perceived truth of statements and simulated social media postings: An experimental investigation of source credibility, repeated exposure, and presentation format. Cogn. Res. Princ. Implic. 5 (1), 56. http://dx.doi.org/10.1186/s41235-020-00251-4.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C.T., Nielsen, R.K., 2021. Reuters Institute Digital News Report 2021. Reuters Institute for the Study of Journalism, URL https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021.

Nyhan, B., 2021. Why the backfire effect does not explain the durability of political misperceptions. Proc. Natl. Acad. Sci. 118 (15), http://dx.doi.org/10.1073/pnas.1912440117.

Nyhan, B., Reifler, J., 2010. When corrections fail: The persistence of political misperceptions. Political Behav. 32 (2), 303–330. http://dx.doi.org/10.1007/s11109-010-9112-2.

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., Boyle, M.P., 2020. The ineffectiveness of fact-checking labels on news memes and articles. Mass Commun. Soc. 23 (5), 682–704. http://dx.doi.org/10.1080/15205436.2020.1733613.

Pan, C.A., Yakhmi, S., Iyer, T.P., Strasnick, E., Zhang, A.X., Bernstein, M.S., 2022. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. Proc. ACM Human-Comput. Interaction 6 (CSCW1), 82:1–82:31. http://dx.doi.org/10.1145/3512929.

Park, S., Park, J.Y., Chin, H., Kang, J.-h., Cha, M., 2021. An experimental study to understand user experience and perception bias occurred by fact-checking messages. In: Proceedings of the Web Conference 2021. WWW '21, Association for Computing Machinery, New York, NY, USA, pp. 2769–2780. http://dx.doi.org/10.1145/3442381.3450121.

Pasquetto, I.V., Jahani, E., Atreja, S., Baum, M., 2022. Social debunking of misinformation on WhatsApp: The case for strong and in-group ties. Proc. ACM Human-Comput. Interaction 6 (CSCW1), 1–35. http://dx.doi.org/10.1145/3512964.

Pennycook, G., Binnendyk, J., Newton, C., Rand, D.G., 2021a. A practical guide to doing behavioral research on fake news and misinformation. Collabra: Psychol. 7 (1), http://dx.doi.org/10.1525/collabra.25293.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., Rand, D.G., 2021b. Shifting attention to accuracy can reduce misinformation online. Nature 592 (7855), 590–595. http://dx.doi.org/10.1038/s41586-021-03344-2.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D.G., 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychol. Sci. 31 (7), 770–780. http://dx.doi.org/10.1177/0956797620939054.

Petty, R.E., Cacioppo, J.T., 1979. Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. J. Pers. Soc. Psychol. 37 (10), 1915.

Pew Research Center, 2014. Political Polarization in the American Public. Pew Research Center - U.S. Politics & Policy, URL https://www.pewresearch.org/politics/2014/06/12/section-4-political-compromise-and-divisive-policy-debates/.

Pew Research Center, 2019. In a Politically Polarized Era, Sharp Divides in Both Partisan Coalitions. Pew Research Center - U.S. Politics & Policy, URL https://www.pewresearch.org/politics/2019/12/17/7-domestic-policy-taxes-environment-health-care/.

Pew Research Center, 2020a. America is exceptional in the nature of its political divide. URL https://www.pewresearch.org/short-reads/2020/11/13/america-is-exceptional-in-the-nature-of-its-political-divide/.

Pew Research Center, 2020b. As economic concerns recede, environmental protection rises on the public's policy agenda. URL https://www.pewresearch.org/politics/2020/02/13/as-economic-concerns-recede-environmental-protection-rises-on-the-publics-policy-agenda/.

Pornpitakpan, C., 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. J. Appl. Soc. Psychol. 34 (2), 243–281. http://dx.doi.org/10.1111/j.1559-1816.2004.tb02547.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1559-1816.2004.tb02547.x.

Prasad, M., Perrin, A.J., Bezila, K., Hoffman, S.G., Kindleberger, K., Manturuk, K., Powers, A.S., 2009. "There must be a reason": Osama, saddam, and inferred justification. Sociol. Inquiry 79 (2), 142–162. http://dx.doi.org/10.1111/j.1475-682X.2009.00280.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-682X.2009.00280.x.

Revell, T., 2017. How to turn facebook into a weaponised AI propaganda machine. New Sci. URL https://www.newscientist.com/article/2142072-how-to-turn-facebook-into-a-weaponised-ai-propaganda-machine/.

Saltz, E., Barari, S., Leibowicz, C., Wardle, C., 2021a. Misinformation Interventions are Common, Divisive, and Poorly Understood. Harvard Kennedy Sch. Misinf. Rev. http://dx.doi.org/10.37016/mr-2020-81, URL https://misinforeview.hks.harvard.edu/article/misinformation-interventions-are-common-divisive-and-poorly-understood/.

Saltz, E., Leibowicz, C.R., Wardle, C., 2021b. Encounters with visual misinformation and labels across platforms: An interview and diary study to inform ecosystem approaches to misinformation interventions. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. In: CHI EA '21, Association for Computing Machinery, New York, NY, USA, pp. 1–6. http://dx.doi.org/10.1145/3411763.3451807.

Schaewitz, L., Flanagin, A.J., Hoss, T., Kölmel, L., Metzger, M.J., Winter, S., Krämer, N.C., 2022. Social sharing of political disinformation: Effects of tie strength, message valence, and corrective information on evaluations of political figures. West. J. Commun. 1–23. http://dx.doi.org/10.1080/10570314.2022.2100471.

Sellers, R., 2013. How sliders bias survey data. MRA's Alert 53 (3), 56–57, URL https://greymatterresearch.com/wp-content/uploads/2019/09/Alert-Sliders-2013.pdf.

Shahid, F., Mare, S., Vashistha, A., 2022. Examining source effects on perceptions of fake news in rural India. Proc. ACM Human-Comput. Interaction 6 (CSCW1), 89:1–89:29. http://dx.doi.org/10.1145/3512936.

Sharma, M., Yadav, K., Yadav, N., Ferdinand, K.C., 2017. Zika virus pandemic—analysis of facebook as a social media health information platform. Am. J. Infect. Control 45 (3), 301–302. http://dx.doi.org/10.1016/j.ajic.2016.08.022, URL https://www.ajicjournal.org/article/S0196-6553(16)30918-X/fulltext.

Sirlin, N., Epstein, Z., Arechar, A.A., Rand, D.G., 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. Harvard Kennedy Sch. Misinf. Rev. http://dx.doi.org/10.37016/mr-2020-83.

Sterrett, D., Malato, D., Benz, J., Kantor, L., Tompson, T., Rosenstiel, T., Sonderman, J., Loker, K., 2019. Who shared it?: Deciding what news to trust on social media. Dig. Journalism 7 (6), 783–801. http://dx.doi.org/10.1080/21670811.2019.1623702.

Stieglitz, S., Dang-Xuan, L., 2013. Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior. J. Manage. Inf. Syst. 29 (4), 217–248. http://dx.doi.org/10.2753/MIS0742-1222290408.

Swire-Thompson, B., DeGutis, J., Lazer, D., 2020. Searching for the backfire effect: Measurement and design considerations. J. Appl. Res. Memory Cogn. 9 (3), 286–299. http://dx.doi.org/10.1016/j.jarmac.2020.06.006.

Taber, C.S., Lodge, M., 2006. Motivated skepticism in the evaluation of political beliefs. Am. J. Political Sci. 50 (3), 755–769. http://dx.doi.org/10.1111/j.1540-5907.2006.00214.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2006.00214.x.

Thorson, E., 2016. Belief echoes: The persistent effects of corrected misinformation. Political Commun. 33 (3), 460–480. http://dx.doi.org/10.1080/10584609.2015.1102187.

Trevors, G., Duffy, M.C., 2020. Correcting COVID-19 misconceptions requires caution. Educ. Res. 49 (7), 538–542. http://dx.doi.org/10.3102/0013189X20953825.

Van Noort, G., Antheunis, M.L., Van Reijmersdal, E.A., 2012. Social connections and the persuasiveness of viral campaigns in social network sites: Persuasive intent as the underlying mechanism. J. Mark. Commun. 18 (1), 39–53.

Vicario, M.D., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W., 2016. The spreading of misinformation online. Proc. Natl. Acad. Sci. 113 (3), 554–559. http://dx.doi.org/10.1073/pnas.1517441113, URL https://www.pnas.org/doi/abs/10.1073/pnas.1517441113.

Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science 359 (6380), 1146–1151. http://dx.doi.org/10.1126/science.aap9559, URL https://www.science.org/doi/10.1126/science.aap9559.

Vraga, E.K., Bode, L., 2018. I do not believe you: how providing a source corrects health misperceptions across social media platforms. Inf. Commun. Soc. 21 (10), 1337–1353. http://dx.doi.org/10.1080/1369118X.2017.1313883.

Vraga, E.K., Bode, L., 2020. Correction as a solution for health misinformation on social media. Am J Public Health 110 (Suppl 3), S278–S280. http://dx.doi.org/10.2105/AJPH.2020.305916, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7532323/.

Walter, N., Tukachinsky, R., 2020. A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? Commun. Res. 47 (2), 155–177. http://dx.doi.org/10.1177/0093650219854600.

Wardle, C., Derakhshan, H., 2018. Thinking about 'information disorder': Formats of misinformation, disinformation, and mal-information.. In: Ireton, C., Posetti, J. (Eds.), Journalism, 'Fake News' &Amp; Disinformation. Unesco Publishing, pp. 43–54.

Wardle, C., Singerman, E., 2021. Too little, too late: Social media companies' failure to tackle vaccine misinformation poses a real threat. BMJ 372, http://dx.doi.org/10.1136/bmj.n26, URL https://www.bmj.com/content/372/bmj.n26.

Weeks, B.E., 2015. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. J. Commun. 65 (4), 699–719. http://dx.doi.org/10.1111/jcom.12164.

Wijenayake, S., van Berkel, N., Kostakos, V., Goncalves, J., 2019. Measuring the effects of gender on online social conformity. Proc. ACM Human-Comput. Interaction 3 (CSCW), 1–24. http://dx.doi.org/10.1145/3359247.

Wijenayake, S., van Berkel, N., Kostakos, V., Goncalves, J., 2020. Quantifying the effect of social presence on online social conformity. Proc. ACM Human-Comput. Interaction 4 (CSCW1), 1–22. http://dx.doi.org/10.1145/3392863.

Wijenayake, S., Hettiachchi, D., Hosio, S., Kostakos, V., Goncalves, J., 2021. Effect of conformity on perceived trustworthiness of news in social media. IEEE Internet Comput. 25 (1), 12–19. http://dx.doi.org/10.1109/MIC.2020.3032410.

Wintersieck, A., Fridkin, K., Kenney, P., 2021. The message matters: The influence of fact-checking on evaluations of political messages. J. Political Mark. 20 (2), 93–120. http://dx.doi.org/10.1080/15377857.2018.1457591.

Wood, T., Porter, E., 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. Political Behav. 41 (1), 135–163. http://dx.doi.org/10.1007/s11109-018-9443-y.

Wyer, Jr., R.S., Albarracín, D., 2005. Belief formation, organization, and change: Cognitive and motivational influences. In: The Handbook of Attitudes. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 273–322.

Yaqub, W., Kakhidze, O., Brockman, M.L., Memon, N., Patil, S., 2020. Effects of credibility indicators on social media news sharing intent. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, Association for Computing Machinery, New York, NY, USA, pp. 1–14. http://dx.doi.org/10.1145/3313831.3376213.

Zhang, A.Q., Montague, K., Jhaver, S., 2023. Cleaning up the streets: Understanding motivations, mental models, and concerns of users flagging social media posts. http://dx.doi.org/10.48550/ARXIV.2309.06688, arXiv URL https://arxiv.org/abs/2309.06688.