Regular Issue: Original Article

Journalism & Mass Communication Quarterly
1–30

DOI: 10.1177/10776990251350563
http://journals.sagepub.com/home/jmq


Credibility as a Double-Edged Sword: The Effects of Deceptive Source Misattribution on Disinformation Discernment on Personal Messaging

Cristian Vaccari[1], Andrew Chadwick[2], Natalie-Anne Hall[3], and Brendan Lawson[2]

Abstract

Disinformation often features reputable sources to boost false information's credibility, but does this *deceptive source misattribution* shape its spread on personal messaging? In a preregistered between-subjects survey experiment on U.K. WhatsApp users ($N = 2{,}580$), we showed participants WhatsApp messages containing true or false news attributed to either British Broadcasting Corporation (BBC) News or no source. Attribution to BBC News significantly increased message credibility. Importantly, however, participants' responses to false messages attributed to BBC News were statistically indistinguishable from their responses to true messages. On personal messaging, source credibility can boost the spread of accurate news but can also be used deceptively to propagate falsehoods.

Keywords

personal messaging, WhatsApp, experiment, source effects, disinformation

[1]School of Social and Political Science, University of Edinburgh, UK
[2]Loughborough University, UK
[3]Cardiff University, UK

Corresponding Author:
Cristian Vaccari, School of Social and Political Science, University of Edinburgh, Office 2.13C, Chrystal Macmillan Building, 15a George Square, Edinburgh EH8 9LD, UK.
Email: cvaccari@ed.ac.uk

In 2018, a video reporting the supposed start of a nuclear war between Russia and the North Atlantic Treaty Organization (NATO) in the Baltic Sea went viral on WhatsApp, the world's most popular personal messaging platform. The video featured a logo and graphics typically seen in news from the British Broadcasting Corporation (BBC), the United Kingdom's main public service media organization and one of the most popular and trusted news outlets globally. In the clip, an actor impersonating a journalist moved across a broadcasting studio closely resembling one of the BBC's.[1] In fact, the video had not originally been created to spread false information. It had been posted on YouTube by a company that wanted to use it to assess its clients' reactions to a potential disaster. The BBC logo, graphics, and studio had been reconstructed to make the video realistic, but the original clip bore a disclaimer that it was a "fictional dramatization," and its description on YouTube stated: "This is a work of fiction and is not happening in fact. Don't panic." However, these caveats were absent in the viral clips, which spread so fast on WhatsApp that the BBC was forced to issue an official clarification.[2] Four years later, amid tensions caused by Russia's invasion of Ukraine, the same clip spread again.[3]

This episode was far from isolated, as highlighted by many reports of similar incidents by reputable fact-checkers. In August 2024, soon after the eruption of racist street violence following a mass stabbing of young girls in Southport (UK), Elon Musk, primary owner of X (formerly Twitter), used the platform to amplify a misleading post by the coleader of the U.K. far-right party Britain First. The post featured a fabricated screenshot constructed to resemble a headline published by British newspaper *The Daily Telegraph* and falsely attributed to one of its reporters.[4] In September 2024, the U.S. government seized multiple internet domains employed by Russian agencies to spread disinformation across Western democracies in what came to be known as the "doppelganger operation." This involved creating realistic "clones" of Western news websites, such as the BBC, to spread disinformation about the Russian invasion of Ukraine.[5] These and many other examples (see our Supplemental Material file[6] for an extensive illustrative list) highlight the important—but surprisingly underresearched—practice of deceptively attributing false information to reputable sources in today's media environment.

A significant strand of communication research has shown that most people struggle to assess the credibility of the messages they encounter and often rely on the credibility of their source as a heuristic (Flanagin & Metzger, 2007; Hocevar et al., 2017; Hovland, Janis, & Kelley, 1953; Hovland & Weiss, 1951; Kang et al., 2011; Metzger et al., 2020; Sundar & Nass, 2001; Van Der Heide & Lim, 2016). Deciding what messages are credible is particularly challenging in conditions of information abundance that characterize contemporary media environments (Metzger & Flanagin, 2013). When people must make many quick decisions on what to believe, the credibility of established news organizations can help them orient themselves (Strömbäck et al., 2020). However, online, malicious actors can easily post material that appears to come from a reputable source but is fabricated, altered, or taken out of context—a practice we term *deceptive source misattribution*.

Defining Deceptive Source Misattribution

Deceptive source misattribution is fundamentally a form of *disinformation*, that is, intentionally spreading false information. However, deceptive source misattribution's effectiveness as a form of disinformation can also involve what most researchers have, in recent years, defined as *misinformation*, that is, unintentionally spreading false information (Broda & Strömbäck, 2024). The original end goal of deceptive source misattribution is to spread false information by exploiting the increased likelihood that people will believe and share false information presented as originating from a credible source. Yet people might, of course, believe and share such false information in good faith and without malicious intent, precisely because they mistakenly believe they are circulating accurate information from a trusted source. In the conventional terminology that has evolved in this research field, this latter discrete component of deceptive source misattribution would be termed misinformation. To keep our terminology faithful to the goal underlying the original deceptive misattribution of the source—intentionally attributing false information to a credible source to increase its acceptance and circulation—throughout this article, we employ the term disinformation. However, we want to clarify at the outset that such activity can also subsume an element of misinformation.[7]

Attributing false information to reputable news sources in an online post requires little effort and few resources, and both organized actors and lone individuals are known to use this misleading tactic. For instance, Wardle (2018, p. 953) discusses "imposter content" as one of the seven types of information disorder: "Journalists often see their bylines alongside articles they did not write, and organizations' logos are used in video and images they did not create." Kreiss and McGregor (2019) unveil how a U.S. gubernatorial campaign posted Facebook advertisements featuring a newspaper headline that had been misleadingly edited by campaign staff. In a large-scale analysis of COVID-19 misinformation, Brennen et al. (2021) found that a substantial quantity of messages used visual cues that impersonated authoritative sources in attempts to increase the credibility of false claims. As these and the multiple other recent examples we catalog in our Supplemental Material file show, popular and trusted news organizations such as the BBC are prime and recurrent targets of deceptive source misattribution.

Little is known about how source credibility works on personal messaging, how it may be misappropriated to spread disinformation, and what the downstream effects might be. Today, messaging platforms are hugely popular around the world—WhatsApp alone had 2.7 billion users worldwide in 2023[8]—and their users often encounter news shared by various actors across private, semipublic, and public domains (Kligler-Vilenchik, 2022; Masip et al., 2021). Since the most popular platforms are encrypted and lack public archives, disinformation cannot be automatically moderated, markers of provenance are difficult to identify, and fact-checking is more complex than on public social media. Therefore, personal messaging may be particularly susceptible to disinformation via deceptive source misattribution. To date, however, no research has addressed its effects in these particular and important

communication contexts. To fill this gap, we designed a theoretically informed, novel, rigorous, between-subjects survey experiment to ethically test the impact of deceptive source misattribution on a sample of users of WhatsApp, which is the most popular personal messaging platform in the United Kingdom, with three-quarters of the population using it regularly. Overall, we found that on personal messaging, the credibility of a reputable news source can be a double-edged sword: it affects users' belief in, and intended response to, both false and true messages containing news, and it does so in similar ways. Our study contributes to the literature on source credibility by both showing its enduring relevance in personal messaging and highlighting the impact of its misappropriation to spread disinformation.

Source Credibility and Disinformation Discernment on Personal Messaging

We aim to understand how users attribute credibility to the messages they see on personal messaging platforms that contain news. Following Appelman and Sundar (2016, p. 63), we define message credibility as "an individual's judgment of the veracity of the content of communication." To shed light on these evaluations, our theoretical framework integrates theories on source credibility, the mechanisms underlying citizens' ability to discern between true and false information, and the affordances and relational uses of personal messaging.

Source Credibility as a Heuristic

In high-choice, fragmented, and saturated media environments, users must constantly choose what, or whom, to pay attention to and what, or whom, to believe despite limited time, attention, and cognitive resources. Theories of persuasion and information processing suggest two ideal-typical ways in which individuals evaluate information depending on their motivations and abilities. There are various definitions and theories of these two approaches, the most influential being the distinctions between the peripheral and central routes to persuasion (Petty & Cacioppo, 1984), heuristic and systematic processing (Chaiken, 1980; Metzger et al., 2010) and system 1 and system 2 thinking (Kahneman, 2011; Tversky & Kahneman, 1974, 1983). In summary, in peripheral, heuristic, and system 1 approaches, people engage with limited amounts of information based on cognitively efficient heuristics, which enable quick decisions with little effort. By contrast, in central, systematic, and system 2 approaches, people evaluate larger amounts of information more closely and comprehensively, and this requires greater cognitive resources.

When engaging with streams of complex, layered, often multimodal content in their everyday lives online, people regularly rely on heuristics to make efficient decisions (Hilligoss & Rieh, 2008). In this complex context, *source credibility* often helps users assess message credibility. Source credibility is defined as "the attitude of the audience toward the communicator" (Hovland & Weiss, 1951, p. 632) and is

articulated into two key dimensions: *expertise*, "the extent to which a communicator is perceived to be a source of valid assertions," and *trustworthiness*, "the degree of confidence in the communicator's intent to communicate the assertions he [sic] considers the most valid" (Hovland et al., 1953, p. 21). Audiences' perceptions of sources' credibility, in turn, affect their evaluations of the messages attributed to these sources and the persuasiveness of such messages (Hocevar et al., 2017). According to Hilligoss and Rieh (2008), users evaluating online information rely on source-related heuristics that attribute greater credibility to familiar than unfamiliar sources and to primary (i.e., official) than secondary (i.e., unofficial) sources (see also Metzger et al., 2010).

Extensive evidence exists that source credibility enhances message credibility across various communication domains (Ou & Ho, 2024; Wilson & Sherrell, 1993). In the context of news, several studies show that people are more likely to perceive reporting as credible when attributed to reputable news organizations. News attributed to established news organizations is perceived as more credible and shareable than news attributed to fictitious sources (Bauer & Clemm von Hohenberg, 2021; Hameleers et al., 2023; Nekmat, 2020; Sterrett et al., 2019). People are more likely to believe public health messages attributed to established news organizations than messages attributed to friends (Van Der Meer & Jin, 2020) or to news organizations that lack credibility (Edgerly et al., 2020; Oeldorf-Hirsch & DeVoss, 2020). Links to authoritative sources can render misinformation corrections more effective (Vraga & Bode, 2018). Granted, users may not systematically remember the sources of the news they see when they access it indirectly via search engines and social media (Kalogeropoulos & Newman, 2017; Pearson, 2021) and may employ other cues such as social endorsements (Messing & Westwood, 2014), the proximity of the platform where the news appeared (Kang et al., 2011), or the trustworthiness of the person sharing a story (Sterrett et al., 2019). Still, there is a strong consensus in source credibility research that reputable news brands boost the credibility of news online. In this study, we advance global disinformation research by examining how the deceptive use of source credibility can affect message credibility and intended behaviors that are crucial to the spread of falsehoods on personal messaging platforms.

Source Credibility and Audience Discernment Between Accurate and Misleading Information

Deceptive source misattribution employed by malicious actors constitutes a threat to the quality of public discourse—particularly on personal messaging, given the specific features of these platforms, discussed below. However, current knowledge is limited because most studies of source credibility online have solely focused on either entirely accurate or inaccurate news, thus precluding any assessment of how attribution to credible sources influences citizens' ability to discern between truth and falsehood.

Understanding citizens' responses to disinformation requires estimating whether they can accurately distinguish between true and false news, while taking into account their overall tendency to deem news as credible or not (Batailler et al., 2022). For

example, an experiment examining reactions to falsehoods around vaccines might reveal that most participants accurately label false claims as untrue, suggesting a general proficiency in recognizing such falsehoods. However, this outcome could also indicate a generalized skepticism toward all vaccine-related information, regardless of its truthfulness. A more comprehensive approach, as suggested by Guay et al. (2023) and adopted here, compares reactions to both accurate and misleading content and enables us to gauge citizens' discernment between truth and falsehood.

Although widely used in disinformation research, this design has not yet been adopted in studies of source credibility. To date, experimental research manipulating both source credibility and the factual accuracy of messages is lacking. The study that comes closest to an exception exposed participants to mocked-up Facebook news feeds that included true and false headlines from different news outlets, whose *visibility*—not credibility—was experimentally manipulated (Dias et al., 2020). The authors found that participants rated the headlines as equally plausible and shareable regardless of how visible the sources were in the posts. Importantly, however, the headlines had actually been published by the outlets presented in the experiment, and thus their content could have influenced participants above and beyond their source. Indeed, results showed that the perceived credibility of the headlines was strongly correlated with participants' trust in the outlets, and only when there was a mismatch between the plausibility of the news and the reputation of the outlet (for instance, *when a reputable outlet published misleading news*) were participants influenced by the visibility of the outlet. That finding hints at the risks arising when disinformation is presented as if it originated from a credible news organization—the scenario we test in this study. Since Dias et al. (2020) did not experimentally manipulate whether the content published by more and less credible sources was true or not, their study cannot adjudicate whether source credibility helps or hinders discrimination between true and false news.

Testing the effects of true and false messages also enables us to contribute to the debate on users' ability to correctly identify truth and falsehood encountered on personal messaging. Despite widespread concern (Flynn et al., 2017), studies focusing on social media show that most citizens can discern between true and false information on these platforms and that they treat true messages as more credible and shareable than false ones (Allcott & Gentzkow, 2017; Luo et al., 2022; Traberg & Van Der Linden, 2022; Tu et al., 2023; Vaccari et al., 2023). Relatedly, other studies find that people are more likely to try to verify false headlines than true headlines (Edgerly et al., 2020) and that only a minority of social media users deliberately shares misleading information (Chadwick et al., 2025; Grinberg et al., 2019; Guess et al., 2019).

Since most users are more likely to share online content they consider true (Pennycook & Rand, 2019; Vaccari et al., 2023), establishing whether source credibility affects discernment between true and false messages is crucial in assessing the downstream impact of false messages that are deceptively misattributed to credible sources. To assess this threat, we developed a research design that enables us to disentangle the implications of seeing true and false messages with and without attribution to a credible source.

Personal Messaging as a Distinctive Context

Research on source credibility online has mainly investigated social media rather than personal messaging and has rarely engaged with the problem of disinformation. Addressing these gaps requires a distinctive conceptualization of personal messaging and key outcomes that matter for the credibility and spread of false information therein.

We conceptualize personal messaging as *hybrid public–interpersonal* communication environments, which combine personal interactions about everyday life with public content shared in private or semiprivate conversations between individuals or among groups of different sizes. Personal messaging users often encounter news while interacting with family, friends, neighbors, work colleagues, and people with whom they share interests (Masip et al., 2021). In this context, deceptive source misattribution can occur in several ways. People may forward a news story they received from someone else, but the identity of the person or organization that originally posted it might not be visible in the forwarded message (Chadwick et al., 2024; Valenzuela et al., 2021). Users might copy and paste content from a news website without including the link and the source. Or they might inaccurately report a different source from the one where they originally saw the news. Importantly, all of these behaviors are vulnerable to malevolent actors who might post fabricated news while suggesting it was published by a credible news organization, as discussed above. Crucially, considering the centrality of multimodal communication on personal messaging (Hagedoorn et al., 2023; Sundar et al., 2021), visual elements of a message, such as logos, videos, and animations may combine with text to deceptively increase the credibility of false information. Research designs ought to properly consider these factors.

The specific features of personal messaging mean that their users play a more prominent role in preventing or reducing the spread of disinformation than on public social media, which makes understanding their behavior all the more important. On personal messaging, end-to-end encryption prevents automated anti-disinformation techniques available to social media platforms, such as content moderation, prioritization, and removal (Rossini, 2023). In this context, users are crucial in challenging the information shared by others. Capturing these dynamics requires a novel conceptualization of how users respond to news encountered on personal messaging (see next section).

To our knowledge, only two studies have investigated the effects of source credibility on news credibility within personal messaging platforms, but neither assessed whether these effects differ between true and false news, and both focused solely on widely studied outcomes not specific to these platforms. The first (Munger et al., 2024) showed participants in Columbia and Mexico mock-ups of WhatsApp messages containing false news and varied whether or not the message included a link to the news outlet that had originally published it. Participants were then asked how credible they perceived the news to be and how likely they would be to read and share it. The findings showed that including such a link bolstered the perceived credibility of the false news stories and increased participants' likelihood of reading and sharing them. Thus, attributing false news to the sources who published them can enhance their

credibility on personal messaging. However, the study's use of only false news stories prevents an assessment of participants' ability to distinguish between true and false information. Moreover, the study attributed news to the source that had actually published it, and thus it could not explore the critically important scenario where false information is misleadingly attributed to a credible news source. The second study (Tsang, 2021) exposed Hong Kong-based participants to mocked-up WhatsApp posts sharing news on the role of Chinese police in suppressing anti-extradition bill protests. The research manipulated whether the story was attributed to a pro-opposition legacy news outlet, to an online forum, or to no source, and measured perceived news credibility as the outcome. Source attribution did not affect perceptions that the news story was false across the whole sample, but it did among pro-extradition participants who saw the news attributed to a pro-opposition outlet. Because the study varied only the source of the news without altering the content, it cannot determine whether participants would react differently to true versus false messages attributed to various sources.

Conceptualizing User Responses to Disinformation on Personal Messaging

Research on disinformation on social media has mainly centered on three sets of outcomes: *ratings of the veracity* of content, the *intention to share* content online, and, less frequently, the *intention to verify* the accuracy of content online (Allcott & Gentzkow, 2017; Edgerly et al., 2020; Sterrett et al., 2019; Vaccari et al., 2023; Walter et al., 2021). We include those outcomes in this study but we augment them with three further outcomes that recent research, discussed below, suggests might be especially relevant for personal messaging: the *intention to stay silent* after seeing a message, the *intention to ask for additional information*, and the *intention to provide additional information*.

It is important to stress two points here. Firstly, little prior research has examined people's motivations for spreading or correcting disinformation on personal messaging, and what exists has mostly been exploratory. Secondly, while we build on that prior research in this part of our framework, we did not aim to assess all possible psychological states—whether cognitive or affective—that might result from being exposed to disinformation. Rather, we deliberately restricted our design to an assessment of the impact of deceptive source misattribution on key *intended behavioral responses* that will have a significant impact on whether false and misleading information spreads on personal messaging.

Staying silent and not intervening after seeing a message leaves a message unchallenged. Recent in-depth interpretive research on personal messaging users has revealed that a reluctance to engage in potentially conflictual interactions when other users post controversial content can leave false information uncorrected (Chadwick et al., 2024). However, other work has shown that the combination of everyday sociality and political talk on these platforms empowers users to speak out (Kligler-Vilenchik, 2021). Given our focus on assessing behavioral responses, and not psychological states, we

treat this as an empirical matter and test the hypothesis that individuals are more inclined to remain silent when they encounter a message they consider as credible, due to its content or source, and, axiomatically, are more likely to respond when they perceive a message as not credible.

Conversely, asking for, or providing, additional information in response to a message posted by another person are behaviors that involve intervening to manage or regulate the discussion and potentially hold others accountable for the content they share, perhaps to prevent harm to others in the interaction (Chadwick et al., 2025). Thus, we reason that personal messaging users should be more likely to ask about, or provide additional information about, content they do not consider credible.

We fuse these ideas to empirically assess six key outcomes. Three outcomes capture user responses that are likely to augment, or at least not prevent, the spread of disinformation on personal messaging: perceiving the message as credible, sharing the message, and staying silent after seeing the message. In contrast, three outcomes capture user responses that are likely to contextualize, problematize, or open a discussion about such information: asking for additional information about the message, providing additional information about the message, and verifying the accuracy of the message. Identifying to what extent (mis)attribution to a credible source yields these outcomes enables us to better understand why false information spreads on personal messaging.

In summary, prior research suggests that, when news shared online is attributed to a credible news organization, people will be more likely to consider the message credible as well. However, this phenomenon has not been extensively studied in the context of personal messaging, where it is likely to be relevant but challenging to tackle without users' awareness and involvement. Prior research only tested the effects of source credibility, without manipulating whether the news was factually accurate or false—meaning user discernment could not be gauged—and only focused on widely studied outcomes which, while relevant, do not capture the broader spectrum of user behaviors affecting the spread of information on personal messaging. To address these gaps, we developed an experimental design that resembles the experience of personal messaging as realistically as possible in a controlled setting, we exposed WhatsApp users to both true and false news, and we conceptualized a wide range of responses that go beyond the outcomes typically studied by social media-focused research. Our approach, therefore, captures the important ways platform environment, message characteristics, and users' responses shape the spread of disinformation that uses deceptive source misattribution on personal messaging.

Hypotheses

We test three preregistered hypotheses derived from our integrated theoretical framework. These focus on the presence of a credible source (H1), the factual accuracy of the message (H2), and the differential effects of the presence of a credible news source on true and false messages (H3). Each hypothesis includes all six outcomes discussed above.

Our first hypothesis (H1) builds on extensive evidence that source credibility enhances message credibility and promotes behaviors that facilitate the dissemination of such messages: *Participants who see a message containing an attributed credible news source will be more likely to perceive the message as credible (H1a), share or forward the message (H1b), and stay silent (H1c), but less likely to ask for additional information (H1d), provide additional information (H1e), and verify the accuracy of the message (H1f) than participants who see a message posted without an attributed credible news source.*

Our second hypothesis centers on the factual accuracy of a message. Since most people do not routinely believe or share falsehoods on social media, true messages should stand a better chance of being treated as credible, and false messages should be more likely to be challenged on personal messaging. We therefore advance our second set of hypotheses (H2): *Participants who see a message containing news that is factually accurate will be more likely to perceive the message as credible (H2a), share or forward the message (H2b), and stay silent (H2c), but less likely to ask for additional information (H2d), provide additional information (H2e), and verify the accuracy of the message (H2f) than participants who see a message containing news that is false*.[9]

Our third hypothesis focuses on the differential effects of source credibility across true and false messages. We posit that the deceptive inclusion of a credible news organization as the source of a message featuring news might weaken users' discernment, disproportionately boosting the credibility of false messages over true ones. All else being equal, false information should start from a lower credibility baseline than true information. Hence, the deceptive inclusion of a credible news source is likely to give false information a comparatively stronger credibility "boost" than attaching said news source to true information, which is more likely to be believed in the first place.[10] Our third set of hypotheses (H3) thus stipulates that *attribution to a credible news source will moderate the positive relationships between factual accuracy of the message and the likelihood that participants will perceive the message as credible (H3a), share or forward the message (H3b), and stay silent (H3c), and the negative relationships between factual accuracy of the message and the likelihood that participants will ask for additional information (H3d), provide additional information (H3e), and verify the accuracy of the message (H3f), so that these relationships will be weaker for messages attributed to a credible news source than for messages not attributed to a credible news source.*

Research Design, Data, and Method

We conducted a between-subjects full factorial experiment embedded in an online survey of 2,580 U.K. WhatsApp users. We designed various brief vignettes that featured realistic graphical mock-ups of WhatsApp messages and showed these to randomly assigned subgroups of participants, whom we thoroughly debriefed after the experiment. The study was preregistered and received ethical approval from Loughborough University prior to data collection.[11]

Design

Our full factorial design is 2 (topic: climate change vs. voter ID laws) × 2 (credible source attribution: present vs. absent) × 2 (factual accuracy of the message: true vs. false) × 3 (tie strength: a family member vs. a friend vs. an acquaintance from work or the local area) × 3 (group size: one-on-one chat vs. small group of up to 5 people vs. large group of more than 10 people). Due to space constraints, here we focus on two factors essential to our hypotheses: the presence of an attributed credible news source and the accuracy of the news in the message. Consistent with our preregistration, we do not subset our data based on topic, but the Supplemental Material file reports separate analyses by topic. We do not discuss tie strength and group size, but robustness checks confirm our results were not affected by these manipulations (see below and Supplemental Material file).

Vignette Treatments

To maximize ecological validity in a survey-based experiment, we followed best practices to ensure that our treatments were as realistic as possible (Pennycook, Binnendyk, Newton, & Rand, 2021). To this end, we selected both false and true news that users would be likely to see in their everyday lives at the time of the experiment, and we embedded this news content in realistic renditions of informal WhatsApp exchanges.

Each participant was randomly assigned to see a vignette treatment, which was presented on its own on a separate page. Treatments comprised a brief introductory text that invited participants to imagine that a person had posted a message on WhatsApp, immediately followed by a realistic mock-up of two messages in the same WhatsApp mobile interface window.[12] To avoid introducing confounding factors, the mock-ups were blurred in the parts where, in real life, the profile image and name of a sender would appear. The timestamp indicated the messages had been posted "Today" at 8:03 a.m.—a realistic timeframe for participants taking the survey at most hours during the day.

The mock-ups all featured a message at the top stating: "Came across this just now. Worth thinking about." We chose this colloquial style to simulate the informality of news sharing on personal messaging. A substantive message then featured immediately underneath, on a randomly assigned news topic: climate change or voter identification. We chose these topics because they were salient at the time of our data collection, further enhancing the study's ecological validity. Climate change is a regular feature of political debate in the United Kingdom, and new voter identification laws came into effect a few weeks before our experiment. Accordingly, our monitoring of reports by reputable U.K.-based fact-checkers indicated that misinformation about these issues was circulating widely in the weeks prior to data collection.[13] For each topic, participants were randomly assigned to a treatment where the second message contained either factually accurate or false information. For instance, the false news on climate change claimed, with a brief explanation, that the Earth's temperatures will drop by 2°C during this century as part of the Sun's natural cycle, whereas the accurate

message claimed that temperatures will drop by .1° to .2°. To ensure treatments were realistic, we sourced this content from the posts that originally spread this misinformation and the fact-checkers that debunked it, respectively.

Another randomly assigned feature of the mocked-up WhatsApp messages was the presence or absence of an attributed credible news source. We chose to present the public service broadcaster BBC News as the source because it is the most credible news organization in the United Kingdom. In 2023, 61% of the U.K. public trusted the BBC and 45% used it at least weekly for news—by far the highest levels among news organizations in the country.[14] A 2020 survey of U.K. adults by Ipsos found that, among those who follow any news, 62% considered the BBC as both "the one source you are most likely to turn to if you want accurate coverage" and "the one source you are most likely to turn to for news you trust the most."[15] No other news organization was chosen by more than 9% in either category, showing that the vast majority of the British public attributes to the BBC high levels of both expertise and trustworthiness—the two key components of source credibility (Hocevar et al., 2017; Hovland et al., 1953). Moreover, comparative research shows the BBC enjoys wide-ranging and cross-cutting appeal among the U.K. public, attracting large numbers of viewers, listeners, and users irrespective of their ideological preferences (Fletcher et al., 2020).

For these reasons, our study is a highly rigorous and realistic test of the impact of attribution to a credible source and of deceptive source misattribution. It tests attribution to a credible source because messaging users are accustomed to encountering news from the BBC shared by their contacts and consider it a credible source endowed with expertise and trustworthiness. It tests deceptive source misattribution because there are multiple, previously documented attempts to misleadingly appropriate the highly credible BBC News brand to spread falsehoods, as we discussed earlier. To increase the prominence of the source and resemble how WhatsApp presented posts with links to external websites at the time of the experiment, the message mock-ups with a source included an image with the BBC News logo, the address of the BBC website, and a link.[16]

Measurement of Dependent Variables

Our hypotheses focus on six outcomes, all measured after exposure to the treatments. On a new page in the survey interface, participants saw the following introductory text: "In the situation we asked you to imagine, someone shared a message on WhatsApp. We will now ask you a few questions about your views on this message." This was followed by a question measuring perceived message credibility and five questions, presented in random order, which measured various intended behavioral outcomes. Table 1 summarizes the question wording, response modes, and descriptive statistics for all dependent variables. A correlation matrix is available in the Supplemental Material file.

Participants

We used the software G*Power 3.1 (Faul et al., 2009) to conduct an *ex ante* power analysis. Our goal was to obtain .95 power to detect a small effect size of .15 at the standard .05 alpha error probability (See Supplemental Material file). We expected

Table 1. Measurement of Dependent Variables.

Outcome	Question	Response modes (values)	Mean	*SD*
Perceived credibility of the message	"What do you think about the accuracy of the message?"	"It is definitely true" (4); "It is probably true" (3); "I am not sure if it is true or false" (2); "It is probably false" (1); "It is definitely false" (0)	1.84	1.15
Likelihood of sharing or forwarding the message	"Would you share or forward the message on WhatsApp or other messaging apps?"	"Definitely" (4); "Probably" (3); "Possibly" (2); "Probably not" (1); "Definitely not" (0)	.81	.96
Likelihood of staying silent	"Would you stay silent and not reply to the message?"	Same as above	2.20	1.17
Likelihood of asking for additional information	"Would you reply, to ask the person who sent this message to provide more information about it?"	Same as above	1.67	1.19
Likelihood of providing additional information	"Would you send a message giving additional information about the content of the message?"	Same as above	1.44	1.13
Likelihood of verifying the accuracy of the message	"Would you search for information online to check whether the message is accurate?"	Same as above	.96	2.28

Note. *SD* = standard deviation.

small effect sizes because these are prevalent in communication research (Rains et al., 2018), including on source effects (Wilson & Sherrell, 1993). Based on the results of the power analysis, we targeted a sample size of 2,500. We recruited participants via the U.K.-based research company Opinium Research and employed quotas so that our sample matched the U.K. adult population based on gender, age, education, ethnicity, and region of residence. The Supplemental Material file shows that our sample closely resembles our target population on these characteristics. To ensure participants would find the experiment realistic, we only included respondents who said they used WhatsApp at least once a month. We timed our data collection for after the English local elections on the 4th of May 2023. Opinium Research invited 7,800 members of their online panel to participate in our study. We excluded 826 participants who were over quota, 324 who did not complete the survey, and 270 whose responses did not meet the quality criteria specified in our preregistration, i.e., they either failed to commit to providing their best answers in response to a question asked at the start of the survey, or they failed an attentiveness check question asked halfway through the questionnaire, before random assignment to the treatments. Our final sample comprises 2,580 respondents, whose responses were collected between the 19th of May and the 21st of June 2023 (participation rate 33%).

Procedure

After some screening questions and measures of the sampling quotas, participants were asked questions measuring their knowledge of issues in the news, media use, political attitudes, attitudes toward the issues featured in the treatments, digital news literacy, and use of WhatsApp for news. They were then shown the treatment to which they had been randomly assigned, presented on a separate page from which they could not depart for at least 30 s.[17] This method of presenting the treatments resembles the personal messaging users' experience. Unlike social media platforms, which encourage rapid scrolling through heterogeneous news feeds combining multiple posts from different senders, personal messaging organizes content by sender or group. Users focus on one sender or group at a time, rather than seamlessly moving from one post to the next. Our experimental design replicated this experience by presenting posts from a single thread and requiring participants to engage with this thread individually, separate from other content on the platform.

Once they clicked through the page showing the treatment, participants were then asked the questions listed in Table 1, plus other questions (not discussed here due to space constraints), measuring trust in news seen on WhatsApp and preferences for different policies aimed at reducing the spread of misinformation. After manipulation check questions, the survey ended with an extensive, easily printable debriefing note that revealed the study's goals, highlighted that the treatments may have included false information, and provided clear corrections from reputable fact-checkers. The median overall completion time was 10 min and 13 s.

Results

Following our preregistered analysis plan, we ran two sets of Analysis of Variance (ANOVA) to test our two hypotheses focused on main effects. The key independent variables differentiated between participants exposed to a message posted with and without an attributed credible news source (H1), and between participants exposed to a factually accurate and a false message (H2). To test H3, we ran two-way ANOVAs with these two independent variables and an interaction term between them. We present the results in Table 2. In interpreting our findings, we apply Bonferroni corrections to account for multiple comparisons.[18] Since we conducted a total of 18 comparisons, we only reject null hypotheses for *p*-values below .003 (.05/18).

Seeing news in the WhatsApp message attributed to BBC News increased participants' perceptions that the message was credible and the likelihood that they would stay silent after seeing it, while it decreased the likelihood that they would ask for additional information about it, when compared with messages that did not attribute news to any source. All these effects were statistically significant. We did not detect any significant effects of the BBC News source on the likelihood that participants would provide additional information, verify the accuracy of the message, and, once a Bonferroni correction was applied, share or forward the message. However, as we show below, the patterns are in the direction we expected.

To illustrate our findings, Figure 1 plots the mean values and 95% confidence intervals of our six dependent variables among respondents randomly assigned to WhatsApp message mock-ups that either contained or did not contain a BBC News source. As H1 predicted, participants who saw messages with a BBC News source perceived the messages as more credible (a .25 difference on a 0–4 scale, or one-sixth of a standard deviation) and stated they were more likely to stay silent after seeing them and more likely to share them. Participants exposed to mock-ups containing a BBC News source also stated they were less likely to ask for additional information (a .24 difference on a 0–4 scale, or one-fifth of a standard deviation), to provide additional information, and to verify the accuracy of the message. Overall, the results provide some support for H1 because we detected statistically significant effects of credible source attribution for three of the six outcomes we measured. However, these effects are notably small, with the presence of a credible source explaining only 1% or less of the variance in each outcome, as reflected by the η^2 coefficients in Table 2.

The relationships tested in H2 are even weaker, as revealed by the small η^2 coefficients (Table 2). When we compared participants exposed to factually accurate messages and participants exposed to false messages, there were significant differences only for perceived message credibility. As Figure 2 shows, participants responded to the treatments in ways consistent with our predictions: they rated the messages containing true statements as significantly more credible than the messages containing false statements (a difference of .16 on a 0–4 scale, or one-seventh of a standard deviation) and were also marginally more likely to share the messages containing true statements and to stay silent in response to them. Conversely, participants were also marginally less likely to both ask for and provide additional information when they

Table 2. Effects of the Presence of a Credible News Source, Factual Accuracy of the Message, and Interaction Between News Source and Accuracy on Perceived Credibility of and Intended Behavior Toward WhatsApp Messages.

H1 (source)	Source	No source			F	p	η^2
Perceived message credibility	1.951 [1.887, 2.014]	1.746 [1.684, 1.808]			2.506	<.001	.008
Likelihood of sharing/forwarding the message	.857 [.803, .911]	.772 [.721, .823]			5.013	.025	.002
Likelihood of staying silent	2.273 [2.211, 2.335]	2.126 [2.06, 2.192]			1.033	.002	.004
Likelihood of asking for additional information	1.546 [1.482, 1.61]	1.784 [1.719, 1.85]			26.028	<.001	.010
Likelihood of providing additional information	1.41 [1.349, 1.471]	1.467 [1.404, 1.529]			1.606	.205	.001
Likelihood of verifying the accuracy of the message	2.247 [2.178, 2.317]	2.312 [2.243, 2.381]			1.677	.195	.001
H2 (factual accuracy)	True	False			F	p	η^2
Perceived message credibility	1.925 [1.867, 1.984]	1.766 [1.699, 1.833]			12.323	<.001	.005
Likelihood of sharing/forwarding the message	.831 [.779, .883]	.796 [.742, .85]			.845	.358	.000
Likelihood of staying silent	2.245 [2.182, 2.309]	2.15 [2.085, 2.215]			4.214	.040	.002
Likelihood of asking for additional information	1.64 [1.577, 1.704]	1.694 [1.627, 1.761]			1.312	.252	.001
Likelihood of providing additional information	1.414 [1.354, 1.474]	1.464 [1.4, 1.528]			1.258	.262	.000
Likelihood of verifying the accuracy of the message	2.307 [2.239, 2.376]	2.253 [2.182, 2.323]			1.191	.275	.000
H3 (source*accuracy)	Source and true	Source and false	No source and true	No source and false	F	p	η^2
Perceived message credibility	2.012 [1.928, 2.097]	1.887 [1.793, 1.981]	1.84 [1.76, 1.921]	1.649 [1.555, 1.743]	.527	.468	.000
Likelihood of sharing/forwarding the message	.848 [.773, .923]	.866 [.788, .944]	.814 [.743, .886]	.728 [.655, .802]	1.873	.171	.001
Likelihood of staying silent	2.319 [2.232, 2.406]	2.225 [2.136, 2.314]	2.173 [2.081, 2.265]	2.078 [1.983, 2.172]	.000	.984	.000
Likelihood of asking for additional information	1.486 [1.398, 1.574]	1.608 [1.514, 1.701]	1.79 [1.7, 1.881]	1.778 [1.683, 1.873]	2.058	.152	.001
Likelihood of providing additional information	1.368 [1.283, 1.453]	1.453 [1.364, 1.542]	1.459 [1.373, 1.544]	1.475 [1.383, 1.567]	.576	.448	.000
Likelihood of verifying the accuracy of the message	2.26 [2.162, 2.358]	2.234 [2.136, 2.333]	2.353 [2.257, 2.449]	2.27 [2.171, 2.37]	.327	.568	.000

Note. The table reports mean values with 95% confidence intervals in square brackets. The model statistics in the last three columns come from one-way (H1 and H2) and two-way (H3) ANOVAs. Bonferroni-adjusted p-value for null hypothesis rejection = .003.

Figure 1. Perceived Credibility of and Intended Behavior Toward WhatsApp Messages Among Participants Exposed to Messages that Included and Did Not Include a Credible News Source (H1).

Note. The jittered gray dots represent participants, the black dots represent mean values, and the error lines represent 95% confidence intervals. See Table 2 for means and confidence intervals of all distributions.

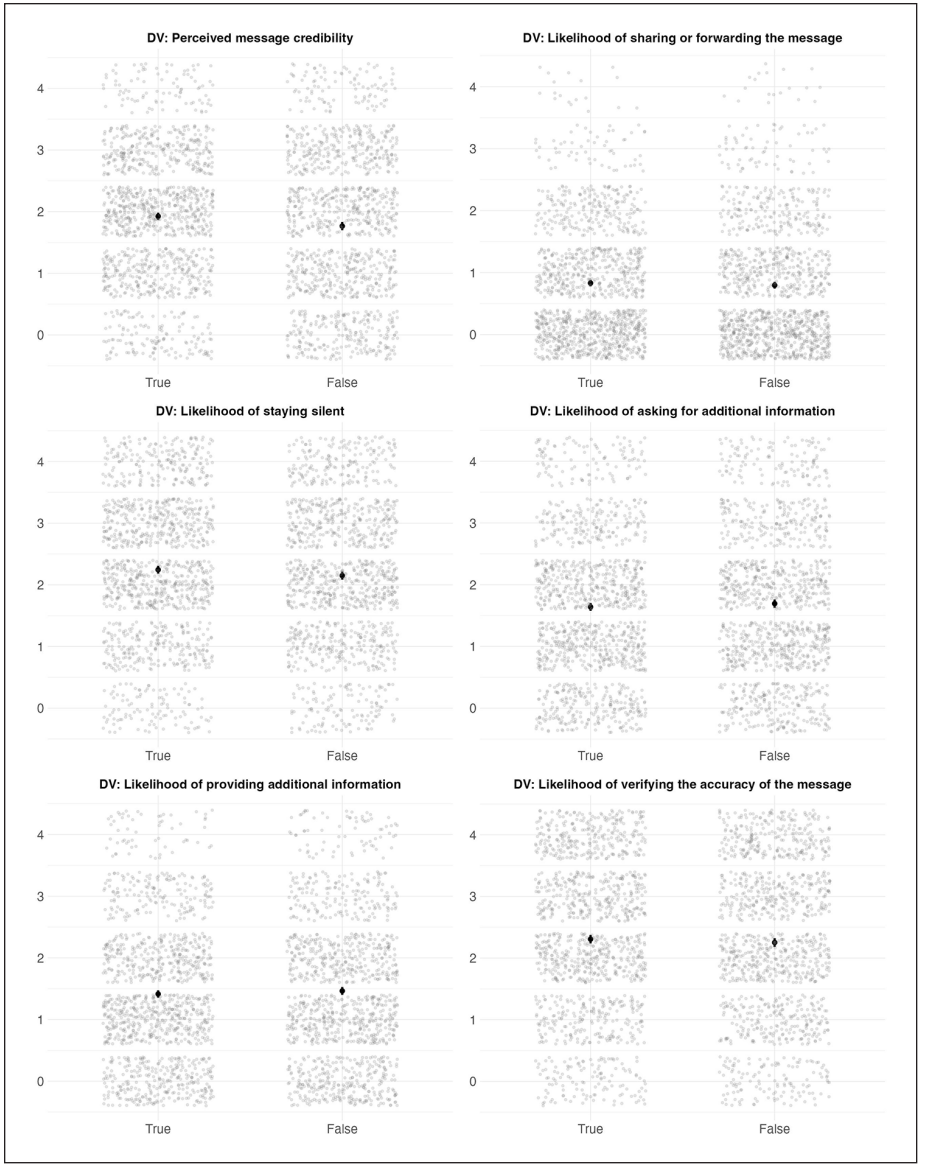

Figure 2. Perceived Credibility of and Intended Behavior Toward WhatsApp Messages Among Participants Exposed to Factually Accurate and False Messages (H2).
Note. The jittered gray dots represent participants, the black dots represent mean values, and the error lines represent 95% confidence intervals. See Table 2 for means and confidence intervals of all distributions.

saw a true message than when they saw a false one. However, none of the coefficients measuring these effects reached conventional significance levels. Finally, contrary to our expectations, participants were slightly more likely to say they would check the accuracy of a true rather than a false message, although, again, these effects were not statistically significant. Overall, the data fail to provide sufficient support for H2, except for perceived message credibility, where, however, the factual accuracy of the message only explains .5% of the variance.

H3 predicted that the presence of a credible news source would weaken the relationships between the factual accuracy of the message and participants' perceptions of its credibility and intended behaviors toward it. As shown in Table 2, none of the six 2-way ANOVAs we ran returned significant coefficients for the interaction term between the presence of a credible news source and factual accuracy of the message. Thus, *we found no evidence that participants reacted differently when a credible source featured alongside either true or false news. Importantly, attributing news to the BBC in the WhatsApp message produced the same effects for both true and false messages.* Including a credible source boosted the perceived credibility of both true and false messages and encouraged participants to state intended behaviors consistent with that assessment.

Although these results do not support H3, the patterns emerging from Figure 3 are striking. When they saw false news attributed to a credible source, participants reacted in essentially the same way as when they saw true news. Consider the top-left pane, which shows how credible participants thought the message was. Overall, participants rated the true messages as more credible than the false messages, irrespective of whether they contained a source. However, on average, participants who saw a *false* message attributed to BBC News rated it equally as credible ($M = 1.887$; 95% CI [1.793, 1.981]) as those who saw a *true* message attributed to BBC News ($M = 2.012$; 95% CI [1.928, 2.097]) and equally as credible as those who saw a *true message without a source* ($M = 1.84$; 95% CI [1.76, 1.921]).[19] As shown in Table 2 and Figure 3, this pattern holds for all the outcomes we measured.

Additional Exploratory Analyses and Robustness Checks

We conducted various preregistered exploratory analyses to estimate potential heterogeneous treatment effects for both attribution to a credible news source and factual accuracy. We did not detect any evidence of differential effects based on education, interest in the news, the perception that the news is complicated, news diets, ideology, attitudes toward and perceived importance of the topic of the news reported in the treatment, frequency of WhatsApp use, frequency of posting news on WhatsApp, experience of encountering inaccurate news on WhatsApp, confidence in one's ability to judge the accuracy of information on WhatsApp, or the experience of correcting inaccurate information encountered on WhatsApp. However, we found some evidence of differential effects of attribution to a credible news source by news knowledge and media literacy. The Supplemental Material file reports and discusses these analyses.

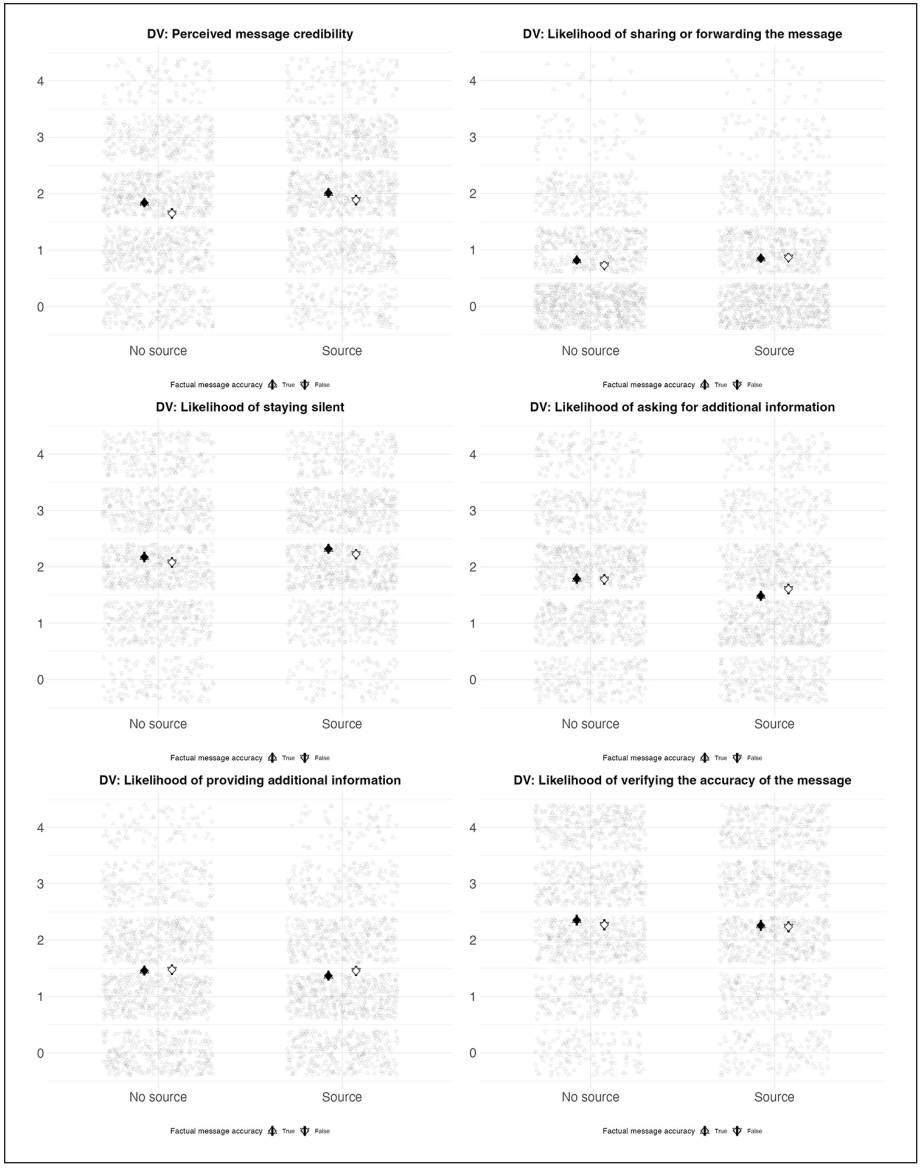

Figure 3. Perceived Credibility of and Intended Behavior Toward WhatsApp Messages Among Participants Exposed to Factually Accurate and False Messages that Included and Did Not Include a Credible News Source (H3).

Note. The jittered gray dots represent participants, the triangles represent mean values, and the error lines represent 95% confidence intervals. See Table 2 for means and confidence intervals of all distributions.

Besides the presence of a credible news source, factual accuracy of the message, and message topic, our preregistered experiment included two more factors, not discussed here due to space constraints: tie strength and group size. To ensure our results are robust to the inclusion of these factors, we supplemented the ANOVA testing H1 and H2 with interaction terms between source and factual accuracy, respectively, and both tie strength and group size. We also ran multivariate regression analyses that included all our experimental factors, as well as a range of individual-level predictors measured before exposure to treatments. We further specified these models with interaction terms between the presence of a credible news source and factual accuracy of the message and tie strength and group size (See Supplemental Material file). The results regarding our hypotheses are consistent with those of the ANOVAs reported here. Thus, our key findings on the effects of deceptive source misattribution were not affected by our other experimental manipulations, nor by our preregistered analytical choices.

Discussion

Online disinformation often features credible sources to boost false information's credibility, a practice we term deceptive source misattribution. We theorized that, due to their distinctive characteristics, personal messaging platforms could be vulnerable to this strategy. Our results advance both theoretical and empirical knowledge on source credibility as a vehicle for disinformation on personal messaging.

Our first key finding is that, when participants saw realistic mock-ups of WhatsApp messages reporting news attributed to the BBC, they rated the message as significantly more credible and saw themselves as significantly more likely to stay silent after seeing it, as well as less likely to ask for additional information about it. This finding expands knowledge on the role of source credibility as a key heuristic in information processing (Hovland & Weiss, 1951; Metzger et al., 2020; Ou & Ho, 2024; Wilson & Sherrell, 1993) in at least four ways. First, by examining the under-researched yet widely popular context of personal messaging, our study extends beyond existing research focused primarily on source credibility on social media (Bauer & Clemm von Hohenberg, 2021; Edgerly et al., 2020; Hameleers et al., 2023; Nekmat, 2020; Oeldorf-Hirsch & DeVoss, 2020; Sterrett et al., 2019; Van Der Meer & Jin, 2020). Secondly, we demonstrate that credible news sources can affect personal messaging users even in digital environments where incessant information flows may hamper source awareness (Kalogeropoulos & Newman, 2017; Pearson, 2021). Thirdly, we establish that attribution to a credible news source enhances the credibility of *both* factually accurate *and* false messages, expanding hitherto limited research on source credibility and disinformation discernment on personal messaging (Munger et al., 2024; Tsang, 2021). Finally, we show that the effects of source credibility are not limited to cognitive appraisals of the veracity of the information but extend to downstream behaviors that can affect its further spread, particularly on personal messaging. These findings highlight the value of bridging the literatures on source credibility and disinformation, as well as of studying the highly relevant, distinctive, but under-researched personal messaging environments.

The importance of source credibility emerges even more clearly in light of our second key finding. Consistent with previous research (Allcott & Gentzkow, 2017; Luo et al., 2022; Traberg & Van Der Linden, 2022), our participants rated true messages as significantly more credible than false ones. However, in contrast with other studies (Tu et al., 2023; Vaccari et al., 2023), they did not report significantly different behavioral intentions toward true versus false messages. Personal messaging users do not seem to be willing to treat factually accurate and misleading content differently in their interactions on these platforms. These null findings may depend on some aspects of our experimental design, but they may also reveal some distinctive features of personal messaging as a social context. Consistent with previous in-depth interpretive studies of personal messaging users (Chadwick et al., 2024; Kligler-Vilenchik, 2022; Pearce & Malhotra, 2022), many people might not be prepared to act on these platforms to endorse news they consider credible and to challenge news they see as misleading. Considering the centrality of personal messaging users, rather than platforms, journalists, or fact-checkers, to the spread and correction of disinformation in these contexts, our study highlights the need to better understand the mechanisms that lead users to take action to protect the integrity of their information environments.

This brings us to the third—and we believe the most important—contribution of this study. Participants' responses to false messages attributed to a credible news source were statistically indistinguishable from their responses to true messages. In other words, *attribution to a credible source elevated the credibility of true and false news alike, thus failing to enhance participants' discernment between true and false information*. Seeing a message that reported a news story attributed to the BBC substantially boosted the message's credibility, regardless of whether that message was true or false. We detected similar patterns for all the dependent variables we measured, suggesting that the presence of a credible source results in similar cognitive and behavioral outcomes irrespective of the veracity of the content. It is striking how participants' discernment between truth and falsehood was disrupted by the addition of a credible source—a tactic that is simple for malicious actors to employ yet relatively difficult to detect in the comparatively hidden world of personal messaging. As this was, to our knowledge, the first experimental study of deceptive source misattribution, and as we focused exclusively on personal messaging, we cannot generalize our results beyond the context of our research. However, as our findings on source credibility confirm those from studies of social media and expand decades of communication research, it is plausible that deceptive source misattribution may have similar effects on social media as those documented here for personal messaging.

Limitations

Our study has some limitations that need to be acknowledged.

All our treatments simulated a personal messaging environment, and as a result we cannot estimate whether participants' responses would be different if they had been exposed to treatments resembling social media or other digital spaces. Relatedly, we focused on WhatsApp, which is by far the most popular messaging app worldwide but has specific affordances and uses that may not universally apply to other platforms.

Our online experimental environment enabled us to tightly control key features of the content and the context participants engaged with and to take various measures to enhance realism, but users' everyday interactions on personal messaging could be substantially different from those we asked participants to imagine. The only solution to this threat to ecological validity is to conduct field experiments within personal messaging, which might be feasible in certain group settings if sufficient ethical safeguards are in place (e.g., Vermeer et al., 2021) but are impossible in one-to-one interactions that constitute the majority experience.

In addition, our study is situated in a single country, the United Kingdom, whose systemic features—particularly the popularity, reputation, and cross-cutting appeal of the BBC (Fletcher et al., 2020)—may limit the generalizability of our findings, though it is worth stating that, for the same reasons, our inclusion of the BBC was a particularly robust test of our hypotheses due to its demonstrably high credibility. Still, because we designed our experimental treatments to operationalize the construct of a credible source through visible markers of source provenance, in this case referring to BBC News, we did not measure respondents' perceptions of the BBC's expertise and trustworthiness before randomly assigning them to our treatments, nor did we expose participants to news attributed to different sources (Edgerly et al., 2020). Future experimental research could extend our study by assessing the perceived credibility of different sources prior to presenting treatments featuring those different sources to participants.

We also exposed participants to news about two specific topics—climate change and voter identification laws—and our findings may not seamlessly translate to other issues; see Clemm Von Hohenberg (2023) for a comprehensive approach.

Furthermore, we relied on self-reports of participants' intended behavior, and although research suggests self-reported measures correlate with real-world online activity (Mosleh et al., 2020), unobtrusive measures of behavior would be preferable, though we note, too, that they are impractical on encrypted messaging platforms.

We also used single-item measures for all our key outcomes to avoid overburdening respondents with an excessive number of questions at the end of the survey, which would have reduced response quality (Galesic & Bosnjak, 2009). Moreover, employing multi-item scales measuring message credibility (e.g., Appelman & Sundar, 2016) would have risked priming participants' accuracy motivations, as is the case when respondents are asked to evaluate the truthfulness of information (Epstein et al., 2021; Pennycook, Epstein, et al., 2021). This, in turn, would have artificially increased participants' discernment when answering the subsequent questions measuring our key behavioral outcomes.

Finally, we were able to detect some statistically significant effects based on an adequately powered sample, but the small effect sizes indicate the need for caution in interpreting them.

Conclusion

We conclude by suggesting some implications of our findings for digital platforms, policymakers, news organizations, and researchers.

Digital platforms should consider adopting policies that penalize the deceptive misattribution of false information to reputable news organizations. This is easier on public social media, where automated content moderation is feasible, than on end-to-end encrypted personal messaging. At a minimum, messaging providers should enhance their digital literacy initiatives to better equip users to recognize and deal with these risks, as well as support fact-checking programs.

For policymakers, our evidence that the BBC News source influenced our participants' responses suggests that public service news organizations can maintain an important role in orienting audiences. Yet the adverse effects that can occur when false information is misattributed suggest the need for policies that help protect news organizations from misuse of their reputations online. Our findings should also inform the design of media literacy campaigns that encourage citizens to check the sources of the information they encounter,[20] as well as initiatives that curate lists of reliable news organizations to aid in these assessments.[21] These approaches may be ineffective, or even counterproductive, if users are ill-equipped to protect themselves against deceptive source misattribution, as our study shows.

Our results suggest that news organizations' brands convey their credibility even in fragmented media environments where many users access news indirectly and accidentally (Toff & Nielsen, 2018). In the hybrid public–interpersonal context of personal messaging, information often flows from the public domain of news into and across private and semiprivate interactions (Chadwick et al., 2024). However, a news organization's reputation can be jeopardized if its brand is attached to false news, and the risks are even greater on encrypted personal messaging because this content cannot be removed or moderated.

Finally, by integrating the theories of source credibility, disinformation, and personal messaging, we have uncovered new and potentially problematic effects of source credibility in contemporary media environments. Overall, the credibility of a reputable news organization matters in the context of personal messaging platforms, but it turns out to be a double-edged sword. When source credibility is attached to legitimate news, it can help factually accurate information spread. But when source credibility is attached deceptively to false news, it can not only increase the acceptance of disinformation, but also stimulate behaviors that facilitate disinformation's spread.

Data Availability

The data underlying this article and the code necessary to replicate all our findings are available at https://osf.io/h4jg6/

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Research reported in this manuscript was funded by the

Leverhulme Trust as part of the project "Understanding the Everyday Sharing of Misinformation on Private Social Media" (RPG-2020-019), Principal Investigator: Professor Andrew Chadwick, Loughborough University.

ORCID iD

Cristian Vaccari https://orcid.org/0000-0003-0380-8921

Notes

1. See https://web.archive.org/web/20170425002349/https://www.youtube.com/watch?v=2VZ3LGfSMhA
2. See https://www.bbc.co.uk/news/blogs-trending-43822718
3. See https://www.reuters.com/article/idUSL1N2U41K4/
4. See https://www.telegraph.co.uk/news/2024/08/08/elon-musk-telegraph-article-fake-viral/
5. See https://www.disinfo.eu/doppelganger-operation/
6. The Supplemental Material file is available at https://osf.io/h4jg6/
7. We thank one of the anonymous peer reviewers for suggesting we include this explanation.
8. See https://www.statista.com/statistics/1306022/whatsapp-global-unique-users/
9. The literature on motivated reasoning (e.g., Osmundsen et al., 2021) suggests that accuracy may be less relevant than ideology when people are motivated by directional goals. In our exploratory analyses (discussed below), we found little evidence that ideology and attitudes toward the issues discussed in the treatments moderated the relationships addressed by H2.
10. In Holbert and Park's (2020) terminology, we hypothesize a *contributory* type of moderation, where there is a significant relationship between the presence of a credible news source and all our key outcomes for both true and false messages, but we reason that this relationship should be stronger for false messages.
11. The preregistration is available at https://doi.org/10.17605/OSF.IO/2MGNV. The three sets of hypotheses discussed here are numbered as H4, H5, and H11 in the preregistration. We have made some light edits to the original wording of some hypotheses to enhance clarity. The results of all preregistered analyses not discussed in this article due to space constraints are available in the Supplemental Material file.
12. See the example mock-ups in the Supplemental Material file.
13. See https://www.bbc.co.uk/news/science-environment-59251912 for climate change and https://www.channel4.com/news/factcheck/factcheck-why-will-voter-identification-be-required-for-elections-in-great-britain-and-what-id-will-polling-stations-accept-explained for voter identification
14. See https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/united-kingdom
15. See https://www.ipsos.com/sites/default/files/ct/news/documents/2020-05/trust-accuracy-impartiality-2020.pdf
16. The format of this link resembled that of a BBC news webpage, but the link did not point to an existing webpage. On our experiment interface, participants could neither click on this link nor copy it.
17. The median time participants spent viewing the treatments was 35 s. Of 2,580 participants, 1,316 (51%) saw a message about climate change and 1,264 (49%) saw a message about voter identification; 1,307 (50.7%) saw a message that included the BBC News source and 1,273 (49.3%) saw a message that did not include a source; finally, 1,309 (50.7%) participants saw a factually accurate message and 1,271 (49.3%) saw a false message.

18. This necessary provision was not included in our preregistered analysis plan.
19. These results are unlikely to be caused by ceiling or floor effects. As can be seen in Table 1, the variable measuring message credibility ranges from 0 to 4, and the mean value is slightly below 2, corresponding to a cautious response of "I am not sure if it is true or false." As shown in Table 2, average levels of perceived credibility were also lower than 2 among all participants who saw posts with a true news story. There was clearly substantial room for this variable to move both upward, toward higher perceived credibility, and downward, toward lower perceived credibility, as a result of the experimental manipulations. Similar considerations apply to the other outcomes we measured.
20. For instance, a cornerstone of the U.K. Government's 2021 Online Media Literacy Strategy is to help users "assess the reliability of a source of information." See https://assets.publishing.service.gov.uk/media/60f6a632d3bf7f56867df4e1/DCMS_Media_Literacy_Report_Roll_Out_Accessible_PDF.pdf
21. For example, the Journalism Trust Initiative, initiated by Reporters Without Borders, developed an accreditation mechanism for news organizations that aims to "enable consumers and citizens, regulators, investors, donors and the private sector [. . .] to identify and reward trustworthy journalism." See https://www.journalismtrustinitiative.org/

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–236.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, *93*, 59–79.

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, *17*, 78–98.

Bauer, P. C., & Clemm von Hohenberg, B. (2021). Believing and sharing information by fake sources: An experiment. *Political Communication*, *38*, 647–671.

Brennen, J. S., Simon, F. M., & Nielsen, R. K. (2021). Beyond (mis)representation: Visuals in COVID-19 misinformation. *The International Journal of Press/Politics*, *26*, 277–299.

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: Lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, *48*, 139–166.

Chadwick, A., Hall, N.-A., & Vaccari, C. (2025). Misinformation rules!? Could "group rules" reduce misinformation in online personal messaging? *New Media & Society*, *27*, 106–126.

Chadwick, A., Vaccari, C., & Hall, N.-A. (2024). What explains the spread of misinformation in online personal messaging networks? Exploring the role of conflict avoidance. *Digital Journalism*, *12*, 574–593.

Chadwick, A., Vaccari, C., & Kaiser, J. (2025). The amplification of exaggerated and false news on social media: The roles of platform use, motivations, affect, and ideology. *American Behavioral Scientist*, *69*, 113–130.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*, 752–766.

Clemm Von Hohenberg, B. (2023). Truth and bias, left and right: Testing ideological asymmetries with a realistic news supply. *Public Opinion Quarterly*, *87*, 267–292.

Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, *1*(1), 1–12. https://doi.org/10.37016/mr-2020-001

Edgerly, S., Mourão, R. R., Thorson, E., & Tham, S. M. (2020). When do audiences verify? How perceptions about message and source influence audience verification of news headlines. *Journalism & Mass Communication Quarterly*, *97*, 52–71.

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*, *2*, 1–12. https://doi.org/10.37016/mr-2020-71

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Flanagin, A. J., & Metzger, M. J. (2007). The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information. *New Media & Society*, *9*, 319–342.

Fletcher, R., Cornia, A., & Nielsen, R. K. (2020). How polarized are online and offline news audiences? A comparative analysis of twelve countries. *The International Journal of Press/Politics*, *25*, 169–195.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*, 127–150.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*, 349–360.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*, 374–378.

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour*, *7*, 1231–1233.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*, Article eaau4586.

Hagedoorn, B., Costa, E., & Esteve-del-Valle, M. (2023). Photographs, visual memes, and viral videos: Visual phatic news sharing on WhatsApp during the COVID-19 pandemic in Spain, Italy, and The Netherlands. *Digital Journalism*, *12*(5), 656–679.

Hameleers, M., Harff, D., & Schmuck, D. (2023). The alternative truth kept hidden from us: The effects of multimodal disinformation disseminated by ordinary citizens and alternative hyper-partisan media: Evidence from the US and India. *Digital Journalism*, 1–22.

Hilligoss, B., & Rieh, S. Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*, *44*, 1467–1484.

Hocevar, K. P., Metzger, M., & Flanagin, A. J. (2017). Source credibility, expertise, and trust in health and risk messaging. In K. P. Hocevar, M. Metzger, & A. J. Flanagin (Eds.), *Oxford research encyclopedia of communication*. Oxford University Press.

Holbert, R. L., & Park, E. (2020). Conceptualizing, organizing, and positing moderation in communication research. *Communication Theory*, *30*, 227–246.

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion: Psychological studies of opinion change*. Greenwood Press.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*, 635–650.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kalogeropoulos, A., & Newman, N. (2017). *"I saw the news on Facebook": Brand attribution when accessing news from distributed environments*. Reuters Institute for the Study of Journalism.

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 Communication Quarterly*, *88*, 719–736.

Kligler-Vilenchik, N. (2021). Friendship and politics don't mix? The role of sociability for online political talk. *Information, Communication & Society*, *24*, 118–133.

Kligler-Vilenchik, N. (2022). Collective social correction: Addressing misinformation through group practices of information verification on WhatsApp. *Digital Journalism*, *10*, 300–318.

Kreiss, D., & McGregor, S. C. (2019). The "arbiters of what our voters see": Facebook and Google's struggle with policy, process, and enforcement around political advertising. *Political Communication*, *36*, 499–522.

Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research*, *49*, 171–195.

Masip, P., Suau, J., Ruiz-Caballero, C., Capilla, P., & Zilles, K. (2021). News engagement on closed platforms. Human factors and technological affordances influencing exposure to news on WhatsApp. *Digital Journalism*, *9*, 1062–1084.

Messing, S., & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements Trump partisan source affiliation when selecting news online. *Communication Research*, *41*, 1042–1063.

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, *59*, 210–220.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*, 413–439.

Metzger, M. J., Hartsell, E. H., & Flanagin, A. J. (2020). Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research*, *47*, 3–28.

Mosleh, M., Pennycook, G., & Rand, D. G. (2020). Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *PLoS One*, *15*, Article e0228882.

Munger, K., Villegas-Cruz, A., Gallego, J., & Vásquez-Cortés, M. (2024). Reenviado Muchas Veces": How platform warnings affect WhatsApp users in Mexico and Colombia. *Political Communication*, *41*, 719–742.

Nekmat, E. (2020). Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media + Society*, *6*, Article 205630511989732.

Oeldorf-Hirsch, A., & DeVoss, C. L. (2020). Who posted that story? Processing layered sources in Facebook news posts. *Journalism & Mass Communication Quarterly*, *97*, 141–160.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., & Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, *115*, 999–1015.

Ou, M., & Ho, S. S. (2024). Factors associated with information credibility perceptions: A meta-analysis. *Journalism & Mass Communication Quarterly*, *101*, 346–372.

Pearce, K. E., & Malhotra, P. (2022). Inaccuracies and Izzat: Channel affordances for the consideration of face in misinformation correction. *Journal of Computer-Mediated Communication*, *27*, Article zmac004.

Pearson, G. (2021). Sources on social media: Information context collapse and volume of content as predictors of source blindness. *New Media & Society*, *23*, 1181–1199.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*, Article 25293.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*, 590–595.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Petty, R. E., & Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *Advances in Consumer Research*, *11*, 668–672.

Rains, S. A., Levine, T. R., & Weber, R. (2018). Sixty years of quantitative communication research summarized: Lessons from 149 meta-analyses. *Annals of the International Communication Association*, *42*, 105–124.

Rossini, P. (2023). Farewell to big data? Studying misinformation in mobile messaging applications. *Political Communication*, *40*, 361–366.

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. *Digital Journalism*, *7*, 783–801.

Strömbäck, J., Tsfati, Y., Boomgaarden, H., Damstra, A., Lindgren, E., Vliegenthart, R., & Lindholm, T. (2020). News media trust and its impact on media use: Toward a framework for future research. *Annals of the International Communication Association*, *44*, 139–156.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, *26*, 301–319.

Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication*, *51*, 52–72.

Toff, B., & Nielsen, R. K. (2018). "I just Google it": Folk theories of distributed discovery. *Journal of Communication*, *68*, 636–657.

Traberg, C. S., & Van Der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, *185*, Article 111269.

Tsang, S. J. (2021). Motivated fake news perception: The impact of news sources and policy support on audiences' assessment of news fakeness. *Journalism & Mass Communication Quarterly*, *98*, 1059–1077.

Tu, F., Pan, Z., & Jia, X. (2023). Facts are hard to come by: Discerning and sharing factual information on social media. *Journal of Computer-Mediated Communication*, *28*, Article zmad021.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*, 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Vaccari, C., Chadwick, A., & Kaiser, J. (2023). The campaign disinformation divide: Believing and sharing news in the 2019 UK general election. *Political Communication*, *40*, 4–23.

Valenzuela, S., Bachmann, I., & Bargsted, M. (2021). The personal is the political? What do WhatsApp users share and how it matters for news knowledge, polarization and participation in Chile. *Digital Journalism*, *9*, 155–175.

Van Der Heide, B., & Lim, Y. (2016). On the conditional cueing of credibility heuristics: The case of online influence. *Communication Research*, *43*, 672–693.

Van Der Meer, T. G. L. A., & Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health Communication*, *35*, 560–575.

Vermeer, S. A. M., Kruikemeier, S., Trilling, D., & de Vreese, C. H. (2021). WhatsApp with politics?! Examining the effects of interpersonal political discussion in instant messaging apps. *The International Journal of Press/Politics*, *26*, 410–437.

Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication & Society*, *21*, 1337–1353.

Walter, N., Edgerly, S., & Saucier, C. J. (2021). "Trust, then verify": When and why people fact-check partisan information. *International Journal of Communication*, *15*, 21.

Wardle, C. (2018). The need for smarter definitions and practical, timely empirical research on information disorder. *Digital Journalism*, *6*, 951–963.

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science*, *21*, 101–112.

Author Biographies

Cristian Vaccari is the chair of Future Governance, Public Policy, and Technology at the University of Edinburgh.

Andrew Chadwick is a professor of Political Communication in the Department of Communication and Media at Loughborough University.

Natalie-Anne Hall is a lecturer in Social Sciences at Cardiff University.

Brendan Lawson is a lecturer in Communication and Media in the Department of Communication and Media at Loughborough University.