O3C 4

Online Civic Culture Centre

Beyond Quick Fixes

How Users Make Sense of Misinformation Warnings on Personal Messaging

PUBLIC REPORT

Natalie-Anne Hall

Brendan T Lawson

Cristian Vaccari

Andrew Chadwick

Loughborough University

O3C

LEVERHULME TRUST

Table of Contents

About the Authors

Natalie-Anne Hall

Postdoctoral Research Associate, Everyday Misinformation Project, Online Civic Culture Centre, Department of Communication and Media, Loughborough University.

n.hall@lboro.ac.uk www.natalieannehall.com

Brendan T Lawson

Postdoctoral Research Associate, Everyday Misinformation Project, Online Civic Culture Centre, Department of Communication and Media, Loughborough University.

b.b.lawson@lboro.ac.uk

Cristian Vaccari

Professor of Political Communication, Department of Communication and Media, Loughborough University.

c.vaccari@lboro.ac.uk www.cristianvaccari.com

Andrew Chadwick

Professor of Political Communication and Director, Online Civic Culture Centre,, Department of Communication and Media, Loughborough University.

a.chadwicki@lboro.ac.uk www.andrewchadwick.com

About the Everyday Misinformation Project

Based in the Online Civic Culture Centre (O3C) and the Centre for Research in Communication and Culture (CRCC) at Loughborough University, the Everyday Misinformation Project is a three-year study funded by the Leverhulme Trust. The project's aim is to develop a better-contextualised understanding of why people share and correct misinformation online. The project has a unique focus on personal messaging, or what are sometimes called private social media or encrypted messaging apps. These services, particularly WhatsApp and Facebook Messenger, are hugely popular in the UK, but their role in the spread of misinformation is not well understood. In part, this is because, due to their nature, these services are difficult to research. Unlike public social media, they do not have public online archives and they feature end-to-end encryption.

Crucially, however, communication on personal messaging is never entirely defined by its privacy. Rather, these services are best understood as hybrid public-interpersonal communication environments. They weave constant, often emotionally intimate, connection into the fabric of everyday life and are used mainly to maintain relationships with strong ties, such as family, friends, parents, co-workers, and local communities. Yet often the information shared on these services comes from media and information sources in the public worlds of news, politics, science, and entertainment, before it then cascades across private groups, often losing markers of provenance along the way. Personal messaging involves private, interpersonal, and public communication in a variety of subtle, complex, and constantly shifting ways. Understanding how this shapes the spread and the correction of misinformation requires sensitivity to unique affordances and patterns of use. This is our project.

* * *

Funding for the Everyday Misinformation Project was applied for in May 2019 and received in March 2020. Following a delay due to the Covid pandemic, work began in March 2021. The Principal Investigator is Professor Andrew Chadwick, the Co-Investigator is Professor Cristian Vaccari; Dr Natalie-Anne Hall and Dr Brendan T Lawson are the Postdoctoral Research Associates.

The fieldwork has three strands:

- Longitudinal in-depth qualitative interviews with 102 members of the public based in three regions of the UK, recruited to roughly reflect the diversity of British society in terms of age, gender, ethnicity, educational attainment, and a basic indicator of digital literacy.
- Analysis of personal messaging content the participants voluntarily upload to personal online diaries via a mobile smartphone app.
- Multi-wave nationally representative panel surveys and experiments, to be designed based on findings from the first two strands of fieldwork.

This is the second public-facing report from the project. It presents findings based on the first and second strands of the fieldwork. Visit https://everyday-mis.info for more information.

This Report

In 2018, there were several high-profile, violent events linked to forwarded messages on WhatsApp. In response, WhatsApp introduced two key measures: a limit on forwarding messages to up to five chats at one time, and tags to clearly indicate which messages had been forwarded and forwarded many times. But we know very little about the effectiveness of these tags for stemming the spread of misinformation.

In this report, we highlight guiding principles to gauge the effectiveness of these tags and other misinformation warnings on encrypted personal messaging in the UK. We do so by shedding light on how WhatsApp's "forwarded" and "forwarded many times" tags are understood by people, in the context of their real, everyday experiences of personal messaging. We demonstrate the importance of looking beyond the technical function of the feature itself to understand the way such warnings work in practice. Personal messaging is a social arena where relationships and social norms play a key role. The outcomes of measures for mitigating the spread of misinformation depend on the interaction between the technology and its users within particular contexts. Therefore, we root our evaluation of the effectiveness of misinformation warnings within user experiences, knowledge, and social norms.

This report explores two questions:

1. How do UK personal messaging users understand the "forwarded" and "forwarded many times" tags, in the context of their personal messaging use and attitudes towards (mis)information?
2. What barriers and opportunities do these understandings pose for the ability of these and other misinformation warnings to help stem the spread of misinformation on personal messaging?

The evidence presented here is drawn from a larger study covering people's everyday experiences of, and attitudes towards, misinformation on personal messaging. We used a detailed qualitative and interpretive method based on in-depth semi-structured interviews with 102 members of the public in three regions of the UK: London, the East Midlands, and the North East of England. We recruited participants using Opinium Research's national panel of over 40,000 people. Those taking part roughly reflect the diversity of the UK population on age, gender, ethnicity, educational attainment, and a basic indicator of digital literacy.

Building on our findings, we put forward five principles for designing effective misinformation warnings on personal messaging platforms. The principles are sensitive to the unique nature of these communication spaces, where end-to-end encryption of messages makes automated fact-checking impossible. We argue that corporate design choices, which are often aimed at reducing user friction and avoiding negative associations between a platform and the spread of misinformation, inhibit the effectiveness of these misinformation warnings. To overcome the kinds of ambiguous user interpretations of the forwarded tags we reveal in this report, the connection between forwards and the potential spread of misinformation ought to be made more explicit for users of personal messaging.

All interview material used in this report has been anonymized through removal or replacement of any identifying details. All names used are pseudonyms assigned by the researchers.

Summary of Key Findings

Associating Tags with Misinformation

- **Association:** the effectiveness of the "forwarded" and "forwarded many times" tags relies on an association between forwards and misinformation in the minds of users.
- **Context-dependent:** this association may not be strong in all contexts. It may be particularly weak for users who perceive they do not usually receive forwards containing misinformation.
- **News exposure:** news media coverage can help form clear awareness of the reason why the tags were introduced to help users make sense of these measures.

Unintended Associations

- **Unwelcome jokes:** some users associate forwards with viral and unwelcome jokes because they routinely receive large amounts of this forwarded content. This may prompt dismissal of content tagged "forwarded" or "forwarded many times", but not the intended critical engagement with its veracity or origin.
- **Valuable:** other personal messaging users associate the status of "forwarded" with more desirable characteristics. These include content that is socially valuable or contains important or useful information. A minority of participants attributed quality or importance to content marked "forwarded many times".
- **Unaware:** other users may be unaware of or indifferent to the tags. This reflects a lack of appreciation of the significance of forwards for misinformation. To these users, other indicators, such as the content of the message itself, are more important in decisions about trustworthy information.

Five Principles for the Design of Effective Misinformation Warnings

1. **Don't rely on description alone:** simply indicating that a feature such as forwarding has been used is not enough. Misinformation warnings should clearly indicate the potential for misinformation.
2. **Introduce user friction:** misinformation warnings may be ignored or overlooked unless they incorporate more intrusive designs that force the user to stop and reflect.
3. **Gain media exposure:** platforms should engage in publicity campaigns to spread the word about the intended purpose of misinformation warnings.
4. **Consider the context:** understanding the different ways in which personal messaging platforms are used across contexts is crucial to the design of misinformation warnings that are relevant and useful. These contexts are dynamic, as they are shaped by social norms as well as people's relationships with others.
5. **Think beyond platforms:** technological features need to be combined with socially-oriented anti-misinformation interventions, focusing particularly on building social capacities to empower people to challenge misinformation and work together to use personal messaging platforms in ways that help reduce misinformation.

1. Introduction

Encrypted personal messaging is extremely popular. WhatsApp is the leading personal messaging platform worldwide, with more than two billion global users.[1] It is also the leading social media platform in the UK, used by 60 percent of the adult population and outstripping all the public social media platforms, such as Facebook and Instagram. Facebook Messenger is the second most popular in the UK with 18.2 million adult users.[2]

Personal messaging platforms weave together social and personal interactions with discussion of information from the public arenas of news, politics, science, and health. We have therefore dubbed them 'hybrid public-interpersonal' communication settings.[3] This hybrid nature means that misinformation can make its way into people's everyday exchanges on personal messaging.[4] Social norms and interpersonal trust play key roles in the sharing, reception and correction of misinformation within these exchanges.[5]

In recent years we have seen evidence that personal messaging contributes to the spread of misinformation, including financial scams, far-right conspiracy theories, sectarian religious myths, fake remedies, and hate speech.[6] This is a difficult problem for these platforms to tackle, as one of their key selling points is end-to-end encryption of messages. Encryption means that automated moderation, fact-checking, and content removal are not possible on these platforms in the way they are on more public social media. This reduces the scope for technological fixes for misinformation.

Personal message encryption has become a major policy issue in the UK since the proposal of the Online Safety Bill. This bill seeks to force social media providers to take responsibility for harmful content published on their platforms, including misinformation. However, for encrypted apps like WhatsApp this would mean breaking their privacy promise to users in order to monitor and censor the messages that they send. Adamant that they will not compromise the basis of their international platforms, these apps have threatened to withdraw from the UK market if the bill is passed in its current form.[7]

This highlights the fundamental tension between the privacy that users rightly enjoy on encrypted personal messaging, and the ability to spread harmful misinformation on these apps unchecked. This is an extremely difficult tension to resolve. Privacy of peer-to-peer communications is a fundamental right within Western democracies. Ending message encryption, whilst potentially helping to mitigate criminal activity and misinformation, comes at considerable cost to citizens. Preserving encryption, on the other hand, means that the burden is primarily on users to identify, challenge and correct misinformation within their interactions on personal messaging. In the face of this challenge, some platforms have been experimenting with features that might raise awareness of the potential for misinformation, without compromising encryption. This includes the misinformation warnings that are the subject of this report.

[1] Dixon (2022).
[2] OFCOM (2021).
[3] Chadwick, Vaccari & Hall (2023).
[4] Nobre, Ferreira & Almeida (2022); Rand & Andrey (2021).
[5] Chadwick, Vaccari & Hall (2023); Chadwick, Hall & Vaccari (2023); Hall, Chadwick & Vaccari (forthcoming).
[6] Saurwein & Spencer-Smith (2020).
[7] Hern & Milmo (2023).

2. Forwarding and Misinformation on Personal Messaging

One function of personal messaging that has been particularly associated with the spread of misinformation is forwarding. This feature enables the sharing of a message received on a personal messaging app with other contacts or groups of contacts within the same app. Importantly, it is impossible to know how many forwarded messages contain misinformation. However, there are two main reasons that this feature could be conducive to misinformation's spread.

First, the fact that messages can be forwarded to multiple users at once enables the exponential diffusion and thus amplification of misinformation.[8] This is said to be exacerbated by the low-cost, low-risk nature of forwarding. The look and feel of the "forward" button may encourage habituated practices of forwarding without verifying.[9] Furthermore, on encrypted personal messaging, senders enjoy relative protection from repercussions from platform moderators or law enforcement.[10]

The second reason forwarding can be conducive to misinformation's spread is that, at least on WhatsApp, forwarded messages lose markers of provenance. Unlike content "shared" to WhatsApp from external mobile applications (e.g. web browsers, news applications or other social media platforms), forwarding within WhatsApp does not come with metadata about the direct source of the content. In other words, the original message is forwarded, but no information about which user the sender received it from, or when, reaches the recipient.[11] This may enable the spread of misinformation because information on the source of a message can help make judgments about credibility. The issue is particularly pronounced where the forwarded message is made up solely of text, with no links to outside sources.[12] Further complicating this is the fact that personal messages are received directly to one's mobile phone and usually exchanged between strong-ties (that is, intimate connections with those we trust, such as family members and close friends).[13] These messages thus carry a sense of being personal as well as immediate,[14] meaning we feel the need to engage with or respond to them, and to do so in real-time.

3. WhatsApp's Measures Against Misinformation

The potential dangers of the forwarding function came to the fore when it was allegedly used to disseminate rumours that fuelled a spate of deadly mob violence against minority group

[8] Baulch, Matamoros-Fernandez & Johns (2020); Lu et al (2022); Melo et al (2019).
[9] Johns & Cheong (2021).
[10] ibid.
[11] Melo et al (2019); Tandoc et al (2022).
[12] Metzger & Flanagin (2013); Sundar (2008).
[13] Masip et al (2021); Swart, Peters & Broersma (2019).
[14] Melo et al (2019).

members in India.[15] Adding to this were similar incidents of mob violence in Mexico,[16] the spread of vaccine misinformation in Brazil,[17] and political disinformation in the lead-up to Brazil's 2018 and India's 2019 elections.[18] These are complex events rooted in much broader social problems. Nonetheless, WhatsApp's strong presence in the digital communications markets of these countries meant the platform and its parent company, Facebook (now known as Meta), faced pressure from the media, the public, and the Indian government to more decisively tackle the spread of misinformation on its app.[19]

Facebook responded to these challenges by publicly expressing its concern over the use of WhatsApp to spread misinformation. In 2018, it promoted a short newspaper advertising campaign in India seeking to encourage public awareness of the risk posed by misinformation.[20] At the same time, two new features were introduced to the WhatsApp platform. First, limits were placed on the number of individuals and groups a user could forward a message to at one time. Second, WhatsApp introduced tags that indicated when a message had been "forwarded."[21] This was followed in 2019 by a tag marking messages that had been forwarded through a chain of five or more chats as "forwarded many times" (originally "frequently forwarded.")[22]

The first feature places a clear constraint on viral forwarding. A six-month evaluation by WhatsApp found that the implementation of forwarding limits resulted in a significant reduction of forwarded messages and bulk messages around the world[23] (although there is no indication of whether this also resulted in a reduction in the total amount of misinformation circulating on the platform). The effects of the second feature—the tags—are less clear cut, as we outline below.

Timeline of WhatsApp's Misinformation Measures[24]

2017-2018	July 2018	January 2019	August 2019
Multiple incidents of mob violence linked to rumours spread on WhatsApp and leading to 30+ deaths	"Forwarded" tag introduced Forwarding limits trialled in India	Internal evaluation suggests forwarding limits are effective Forwarding limits rolled out globally	"Frequently forwarded" tag introduced, later renamed "forwarded many times"

3.1. "Forwarded" and "Forwarded Many Times" as Misinformation Warnings

The "forwarded" and "forwarded many times" tags can be considered a variety of misinformation warning.[25] These particular tags are a light-touch intervention whose impact is difficult to measure. A survey experiment in Singapore provides some evidence that the "forwarded" and "forwarded many times" tags reduce perceived credibility of messages.[26] Strictly speaking, however, this is not the intended outcome of the tags.

When introducing the tags, WhatsApp explained that "This extra content […] helps you determine if your friend or relative wrote the message they sent or if it originally came from someone else.

[15] See Vasudeva & Barkdull (2020).
[16] Martínez (2018).
[17] Molteni (2018).
[18] Avelar (2019); Kazemi et al (2022); Sahoo (2022).
[19] Vasudeva & Barkdull (2020).
[20] Choudhary (2018).
[21] WhatsApp. (n.d.-a); WhatsApp Blog (2018).
[22] Bhushan (2019); WhatsApp. (n.d.-b); Sharma (2018).
[23] WhatsApp Blog (2019).
[24] Rajput et al (2018); Vasudeva & Barkdull (2020); WhatsApp Blog (2018); WhatsApp Blog (2019).
[25] Pennycook et al (2020); Xie et al (2022); sometimes called a misinformation warning label; Sharevski et al (2021).
[26] Tandoc et al (2022).

[…] WhatsApp cares deeply about your safety. We encourage you to think before sharing messages that were forwarded."[27] In other words, the aim of the "forwarded" and "forwarded many times" tags is to promote critical reflection on the source and veracity of message content. As of March 2023, the tags continue to feature at the top of WhatsApp's Help Centre entry "How to prevent the spread of misinformation" with the recommendation, "If you're not sure who wrote the original message, double check the facts."[28] This is followed by "Check your biases," "Fact check information with other sources," and "Look out for messages that look different."

The "forwarded" and "forwarded many times" tags are a light-touch intervention whose impact is difficult to measure.

The focus on individual responsibility here alludes to the limitations for regulation on encrypted personal messaging platforms such as WhatsApp. As noted above, because the content of messages on WhatsApp and other personal messaging platforms is end-to-end encrypted and private, the platform itself is not able to engage in moderation of content. In this context, the responsibility for protecting oneself and others from misinformation shifts onto the individual user. Faced with these limitations, the "forwarded" and "forwarded many times" tags do not aim to remove, block or intercept misinformation, but to encourage individual user responsibility. This separates misinformation warnings on encrypted personal messaging from those on public social media, which are used to mark individual content that has been fact-checked and found to be misinformation.

Example of a message with a "forwarded" tag on an anonymised contribution submitted by one of our research participants

Importantly, the "forwarded" and "forwarded many times" tags are descriptive labels that merely indicate the fact that a message has been forwarded or forwarded many times. They are not accompanied by any explicit warning or behavioural prompt. Studies conducted into misinformation warnings on public social media platforms such as Facebook and Twitter have found that "contextual" warnings that do not interrupt the user experience, such as the "forwarded" and "forwarded many times" tags, are more likely to be ignored than "interstitial" warnings that compel a user response, such as covers that require a user to click to reveal content.[29] However, introducing this kind of friction in the user experience is seen as harder to justify when no fact checking has taken place.

Besides potentially being ignored or going unnoticed, the "forwarded" and "forwarded many times" tags may even be entirely misunderstood. Because the tags do not include any additional information, it is the user's responsibility to make the association between a message being forwarded (or forwarded many times) and the potential need to more carefully consider its veracity. In the same Singapore-based study mentioned above—the only one thus far that

[27] WhatsApp Blog (2018).
[28] WhatsApp (n.d.-b).
[29] Kaiser et al (2021); see also Sharevski et al (2021).

specifically examined the effectiveness of the tags—interviews and focus groups with users produced mixed results. Some individuals associated the tags with trustworthiness, others with untrustworthiness; some individuals associated the tags with insincere communication, others with a sense of belonging or being included.[30]

While end-to-end encryption precludes automated fact-checking, it does not prevent WhatsApp from incorporating text in the tags that more explicitly encourages users to take action. The tags could read, for example, "Forwarded: double check the facts and think before sharing." In fact, the minimalist design of the tags is a choice. It is reasonable to assume that WhatsApp would not seek to draw excessive attention to the association between messages sent on its platform and poor quality or harmful content, as this could harm user retention and engagement. The same goes for the creation of user friction through interstitial warnings. Thus, WhatsApp has a delicate balance to strike between its corporate interests and its ethical duty (and reputational imperative) to prevent harm, all within the constraints of end-to-end encryption. The design of the "forwarded" and "forwarded many times" tags represents an attempt to strike this balance. The question remains, however, whether the tags are able to fulfil their original aim in their current form.

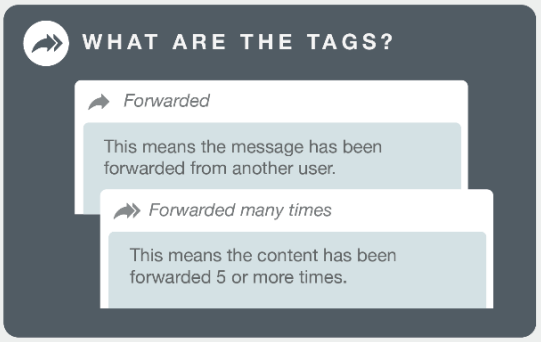
The minimalist design of the "forwarded" tags is a deliberate choice. WhatsApp does not want to draw excessive attention to the association between messages sent on its platform and poor quality or harmful content, as this could harm user retention and engagement. The same goes for the creation of user friction through interstitial warnings. WhatsApp has a delicate balance to strike between its corporate interests and its ethical duty (and reputational imperative) to prevent harm, all within the constraints of end-to-end encryption.

Answering this question requires understanding how the "forwarded" and "forwarded many times" tags function in practice, with careful attention to the contexts in which people actually use personal messaging. This can shed light on how these and other misinformation warnings can be made most effective. In the following section, we outline an approach that facilitates this type of enquiry.

[30] Tandoc et al (2022).

What's up with WhatsApp's forwarded tags?

How a user should respond when they see the tag



Desirable response

Undesirable response

4. Affordances and Contexts

The affordances of the "forwarded" and "forwarded many times" tags emerge from a combination of technological features, the people who use them, and the contexts in which this takes place.

The "forwarded" and "forwarded many times" tags have been dubbed "electronically relayed information cues (ERICs)."[31] When seen by the user, these cues are said to be able to trigger internal rules and shortcuts (called "heuristics" in psychology) for making judgments. As noted above, a forwarded message does not contain information about its source. In this context of information scarcity, the "forwarded" and "forwarded many times" tags may help users assess the trustworthiness of content.[32]

However, whether the tags function as credibility cues in practice is unclear. As noted above, they can be interpreted by recipients in different ways. Understanding their effectiveness requires shifting attention to the contexts in which personal messaging is used. Under what conditions do the tags promote critical engagement with the content's source and veracity? What aspects of the contexts of personal messaging use might impede or promote this? To answer these questions, it is useful to understand the tags as a feature whose affordances are not predetermined but arise within the relationship between tags and users.[33] This relationship is coloured by the complexity of the individual motivations, personal relationships, everyday experiences, knowledge, and social norms in which it is embedded on personal messaging.

Features such as the "forwarded" and "forwarded many times" tags are static and binary. That is, the tags always take the same form and they either appear or they do not. In contrast, the affordances of communication technology are dynamic and variable. Affordances emerge from a combination of technological features, the people who use them, and the contexts in which this takes place.[34] This means they have multiple potential outcomes, even contradictory and unintended ones.[35] For example, while forwarding can be used for the viral spread of misinformation, it can also be used for the viral spread of misinformation debunking.[36]

Our broader research in the Everyday Misinformation Project is continuing to demonstrate the importance of recognising situations of use and user experiences. Through qualitative work with 102 UK personal messaging users, we have been able to reveal, for example, how determining the trustworthiness of statistics and other numbers on personal messaging is a complex, ongoing process undertaken by individuals. The strategies people employ include verification in the public sphere, but also appraisal of relevant peers' motivations and expertise. We have also shown how a norm of conflict avoidance acts as a powerful obstacle to individuals' capacity to correct misinformation posted by their friends, family and acquaintances on personal messaging.[37] In this context too, individuals develop strategies such as scaling down to smaller groups who are known to share

[31] ibid; see also Sundar (2008).
[32] Tandoc et al (2022).
[33] Leonardi (2013).
[34] Evans et al (2017).
[35] Majchrzak et al (2013).
[36] Pasquetto et al (2022).
[37] Chadwick, Vaccari & Hall (2022, 2023).

similar views, or deliberately placing certain topics off limits in interactions with others who are misinformed. Some communities of users take it upon themselves to devise "group rules" to try to soften platform affordances they believe cause misinformation to spread or are harmful to the group's members.[38] Social norms again come into play in people's propensity to engage with anti-misinformation measures. We found that some people position themselves as inherently responsible, critical information consumers, and as a result they may consider themselves immune to misinformation and be less likely to pay attention to interventions.[39]

The same emphasis on situations of use and user experiences is necessary when evaluating the "forwarded" and "forwarded many times" tags. User motivations, abilities, and the social contexts surrounding these are just some of the variables that determine whether the tags give rise to critical reflection and, ultimately, misinformation identification and mitigation. To shed light on these issues, below we discuss evidence from our in-depth qualitative fieldwork with personal messaging users in the UK. The attitudes and perceptions of these people, contextualised within their broader experiences and social environment, allow us to identify potential barriers and opportunities for the effectiveness of the "forwarded" and "forwarded many times" tags as misinformation warnings.

5. Associating the Tags with Potential Misinformation

5.1. The Role of Personal Experiences

We found that people's experience and knowledge of specific contexts on personal messaging are key to the effectiveness of the "forwarded" and "forwarded many times" tags. Among our participants, those who associate these tags with potentially untrustworthy content either have experience in contexts where forwarded misinformation on personal messaging is common or are explicitly aware of these contexts.

For instance, Akram[40] (28, M, London) is an avid news consumer who frequently corrects people he thinks have posted misinformation. In his interviews, Akram described a number of instances of receiving or coming across misinformation online, and he was able to contribute to our smartphone app multiple examples of misleading content he encountered. He said that the "forwarded" and "forwarded many times" tags,

[38] Chadwick, Vaccari & Hall (2023).
[39] Hall, Chadwick & Vaccari (forthcoming).

[40] All of the participant names given in this report are pseudonyms. Please see Appendix A for details.

> *can be useful if it sort of warns you about the implications of that, because then one can choose to sort of decide whether it's likely to be reliable or not, because not everything that's shared across many people will be reliable.*

Akram's interpretation is close to the intended outcome of the tags—he uses them as a prompt for critical reflection. Effat (29, F, North East) similarly associates the tags with potentially untrustworthy content. She has recently immigrated to the UK from Pakistan with her family, and bases her interpretation on her experience receiving forwarded misinformation:

> *Most forward messages, I don't trust them, because most of the time they are not true and it's just fake information . […] Anyone can create anything on WhatsApp, anyone can forward it, so there's no point of trusting that.*

Effat's approach could be seen as healthy scepticism. However, she has found herself giving up on forwards altogether, deciding "there's no point" trusting them. While understandable given her negative past experiences, a blanket dismissal of forwards may not be the most effective at fostering the critical media literacy required for combating misinformation. It may also mean that people dismiss valuable and accurate content shared using this feature on personal messaging.

"Anyone can create anything on WhatsApp, anyone can forward it, so there's no point of trusting that."

Often these individual experiences are set within a broader context. Those in social milieus where viral misinformation forwards are a well-known problem use this to inform their interpretations of the "forwarded" and "forwarded many times" tags. Let us consider Anish (38, M, London), a doctoral researcher who immigrated to London from India just a few years ago. Anish is well aware of the dangers of trusting unverified information, and links this to forwards specifically. He differentiates between two types of forwards and says he responds to each type differently: "I don't pay any attention to forwards unless they're forwards to links from trusted organisations. […] I usually don't bother with forwards that are just random text messages saying someone's random opinion." Anish divulged that these sorts of unsubstantiated forwards sometimes appear in his family WhatsApp groups. These include misleading home remedies "that'll save you from Covid" and political "forwards about this party or this government." He says these are the sorts of forwards he would "usually just ignore." According to Anish, this makes the "forwarded" and "forwarded many times" tags "definitely" useful,

> *because then, in a sense, it gives you a rough idea of […] "this is their own personal opinion," or if this is something else they've been forwarding on from wherever. […] I think even, I pre-judge whether something is worth looking at based on whether it's got that "forwarded" tag on it.*

Anish also made explicit reference to his diaspora background. He is aware that the prominence of forwards in his family WhatsApp groups is partly the product of the specificities of WhatsApp use in India, where most of his relatives live:

> *I think WhatsApp has become a major source of information for people in India, because it's at your fingertips. […] My folks, for example, they're not exactly tech savvy people. They didn't know how to use a mobile […] for a long time, but now once they've discovered WhatsApp, my mum's online all the time. […] She's always scrolling through messages or sending forwards. […] They think "this is reliable or useful," so they forward it.*

The affordances of a particular platform differ between contexts.[41] In India, as Anish points out, WhatsApp plays a significant role not only in social life but in many people's information diets. In this context, habitual, everyday practices of WhatsApp forwarding have come to contribute to the spread of misinformation.[42] According to WhatsApp, India is the top forwarding nation in the world.[43] The high profile of the serious misinformation events discussed earlier in this report have fostered a keen public awareness of the link between forwards and misinformation. As yet, there is no evidence that WhatsApp forwarding plays the same role and enjoys the same popularity in the UK context. This difference has an impact on individual awareness, and thus acts as a barrier to the effectiveness of the "forwarded" and "forwarded many times" tags.

5.2. Learned Knowledge

But direct experience of receiving forwarded misinformation is not necessary for developing effective interpretations. Other people emphasise their knowledge of the circumstances surrounding the introduction of the tags, gleaned through news reporting. Kiri (25, F, London), for example, is a middle-class professional with an economics degree who describes herself as a very sceptical media consumer. She says she has seen the "forwarded" tag only once or twice and never the "forwarded many times" tag. Yet, she attests that if she did see it on a message,

> *I think I would immediately think it's probably not believable, or shouldn't be trusted too much. Mainly because the only times you hear about being forwarded loads and loads of times, I think the last time I really heard about that was in the news. Think it was a BBC article about how it was the Indian communities generally or something. They're having trouble with that because messages are just being forwarded on WhatsApp.*

Oscar (32, M, North East) is another tech-savvy, critical media consumer who says he has "sometimes […] read about, you know, false information being spread on WhatsApp and social media." He claims "the 'forwarded many times' caption added by WhatsApp always makes me sceptical of the content," and expressed an interpretation of the tag that matched its original stated aim:

> *I think that's probably something WhatsApp has actually used to try and slightly, gently warn people, you know, "this isn't something your friends thought of themselves. […] This is something that is kind of going viral on WhatsApp and maybe think about it before deciding." […] That's what it would flag in my mind.*

When asked whether he thought there was a relationship between forwarded content and misinformation, Oscar said his "initial gut reaction would be yes." This is despite the fact that, by his own admission, he has only ever seen the tags on harmless entertainment content, and cannot recall receiving forwarded misinformation. In other words, Oscar's interpretation of the tags is underpinned by his explicit knowledge of the intention behind them and what he had read about the issue of misinformation spreading through forwards on WhatsApp.

"I think the last time I really heard about that was in the news."

These accounts point to the potential for the "forwarded" and "forwarded many times" tags to be effective when their intended purpose is well-communicated. However, they also allude to the barriers to such communication in contexts where practices of personal messaging use differ

[41] Costa (2018).
[42] Banaji et al (2019); Maddox & Kanthawala (2022); Mukherjee (2020); Sarkar (2022).
[43] Sircar (2018).

greatly from those underpinning the tags. In the next section, we outline how those who lack knowledge about the association between forwards and misinformation fill in the tags with their own meanings which are not necessarily in line with their original purpose.

6. Imbuing the Tags with Other Meanings

As noted above, the "forwarded" and "forwarded many times" tags have been designed as descriptive labels that simply mark a message as having been forwarded. For them to prompt critical reflection on the source and veracity of messages, there needs to be an association between forwarding and potentially untrustworthy information in the minds of users. However, this is not always the case. In this section, we outline three other associations that UK personal messaging users make with forwards that affect their interpretation of the tags, and that are not in line with the tags' original purpose.

6.1. No Associated Meaning

We found evidence that the "forwarded" and "forwarded many times" tags are not important or meaningful for some people. These people are not aware of the tags' purpose and do not associate them with the marking of potential misinformation. Instead, they prioritise other factors when making decisions about the trustworthiness of messages they receive on personal messaging, particularly the content of the message itself.

When asked whether knowing that something had been forwarded or forwarded many times was useful, Maisie (68, F, North East) said that it "doesn't matter." Brian (46, M, East Midlands) expressed a similar view and elaborated:

> *Because I'd look at it whatever [it is], so it's not, you know, if it's forwarded, fine, but it wouldn't, it don't make you think, you know, "oh he's forwarded that to whoever." It doesn't, no. It doesn't bother me, that.*

Brian's comment that he would "look at it whatever" asserts his intention to judge each message on its own merits. Because of this, he dismisses as irrelevant the information that a message has been forwarded. Similarly, Joyce (62, F, North East) said she would not pay much attention to the tags because "It depends on what the context of the thing was." Barry (43, M, London), when asked whether the "forwarded" and "forwarded many times" tags bear any relationship to the accuracy of the content, replied "No, not necessarily. […] It's something that I will decide once I read it."

"[The fact that it has been forwarded] doesn't bother me in the slightest."

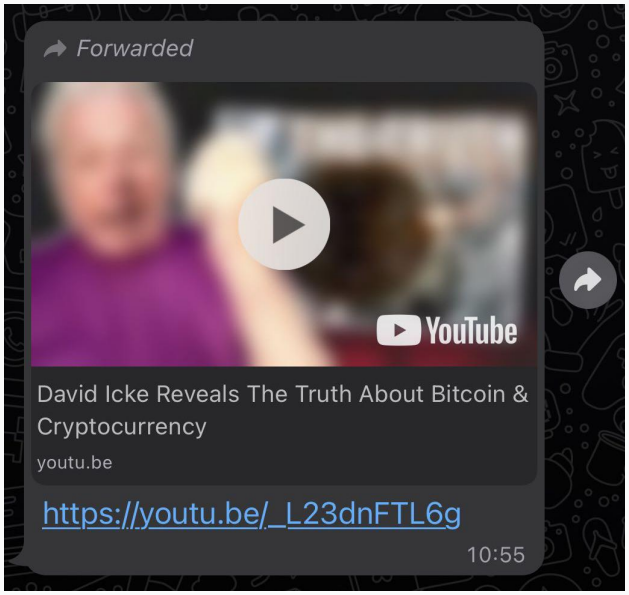
The examples contributed by our participants via our project smartphone app enable us to situate these assessments within participants' concrete experiences. Evan (51, M, East Midlands) contributed a screenshot of a video that was forwarded to him by a friend on WhatsApp. The video is of infamous conspiracy theorist David Icke and carried the "forwarded" tag. When prompted to add a comment to elaborate on his upload, Evan wrote to us via the app:

> *Mr Icke's views are difficult to believe and comprehend. He also doesn't give any*

evidence of how he comes to his conclusions. Using a search engine I can't find any evidence to backup [sic] his claims so I'm putting this in the false category.

Evan did not mention the fact that the message had been forwarded, and according to him it is other information and cues on which he bases his dismissal of the content. In his second interview, when asked directly what he thought about the fact that this video was forwarded, Evan downplayed the significance of this:

> *Yeah, if somebody sends me something, I might look at it, I might not. […] I can look at it and decide for myself what I think of the article or the video. If I don't particularly think it's accurate, I'll just dismiss it. So, no, [the fact that it has been forwarded] doesn't bother me in the slightest.*



The forwarded video Evan received from a friend on WhatsApp

These comments indicate not only a lack of reflection on the significance of the tags, but a lack of recognition of the implications of a message having been forwarded. Just as in Brian's, Joyce's and Barry's examples, Evan claims to want to decide for himself whether a message is trustworthy, regardless of the form it comes in.

"I can look at it and decide for myself what I think of the article or the video."

The emphasis on this could be underpinned by social norms around the importance of critical media literacy skills. In the UK and many other countries today, individuals are increasingly expected to take responsibility for their own critical media literacy, and being a critical media consumer is seen as socially desirable.[44] But in today's media environment, we all use signs and signals to make shortcut decisions about what to trust, dismiss, or verify. Even if users are vigilant enough not to need misinformation warning tags on personal messaging, their dismissal of such warnings does little to promote vigilance among others. Tackling misinformation is a collective effort in which we all have a stake and need to be involved. Given the potential for forwards to facilitate viral misinformation described in section 2 of this report, more not less promotion of awareness of the tags' meaning is needed.

Another interesting example is Elsie (48, F, East Midlands). Elsie said she had "no idea" what the purpose of the "forwarded many times" tag might be beyond a guess that it was "to say that it's been shared by numerous people?" When asked whether she had ever noticed the tags, Elsie scrolled back through five months' worth of a WhatsApp group chat to a video posted by a friend on Christmas Day and exclaimed,

> *Oh! "Forwarded many times"—I've not noticed this! […] I must admit I've never noticed it until I clicked on it then. It wasn't something that was apparent to me and I don't, you know, if somebody sends me a video, I click it and watch it and then probably delete it. Again, wouldn't notice where it had come from*

[44] Hall, Chadwick & Vaccari (forthcoming).

or how many times it had been forwarded, if I'm honest.

In addition to Elsie's indifference to the fact that a message has been forwarded, her comment reflects the potential for the tags to go unnoticed. As noted above, misinformation warnings on public social media platforms that do not interrupt the flow of users' experience have been found to be more likely to be ignored.[45]

"I've never noticed it until I clicked on it then. It wasn't something that was apparent to me."

Asked whether she thought it was useful, in hindsight, to know that this message had been "forwarded many times," Elsie echoed some of the testimonies above: "Not necessarily. Don't know—it depends where she's forwarded it from." Elsie's comment acknowledges the importance of the source of the information—something on which the tags aim to prompt recipients' reflection. Yet, she sees little use for the tags themselves and makes no association between forwards and a greater need to consider provenance. According to Elsie's descriptions across her two interviews, she is not a heavy user of WhatsApp for exchanging news—she and her friends prefer to use WhatsApp to organise meeting up, and then have their discussions in person. The specific use Elsie makes of personal messaging potentially contributes to her indifference to forwards, and thus to her lack of awareness of the "forwarded" and "forwarded many times" tags' intended meaning. Her example demonstrates how the ways features like forwarding are used, and the uses people make of the "forwarded" and "forwarded many times" tags, arise not only out of individual user motivations, but also their relationships with others they communicate with online.

[45] Kaiser et al (2021).

6.2. Viral and Unwelcome Jokes

Some people *do* see the "forwarded" and "forwarded many times" tags as meaningful, but not in the intended way. Unaware of the tags' original meaning, they misunderstand what they are meant to convey. In fact, in the minds of one subset of our participants, there is a strong association between forwards and inappropriate or unwelcome humorous content. This connection is based on people's experiences of receiving this type of forwarded content. In this subset of users we spoke to, there is no acknowledgment that forwards could be associated with the more serious threat of misinformation. They dismiss content tagged as "forwarded" or "forwarded many times" out of disinterest, rather than critically engaging with its veracity or origin.

Information exchanged on personal messaging is intermingled with entertainment content that often combines text, images, and audio. Jokes forwarded from user to user and shared in groups might be valued for being entertaining and can build collective solidarity or make difficult times bearable.[46] The people we spoke to, however, describe forwarded jokes as superfluous or bothersome. David (31, M, Northeast), for example, said that the "forwarded" and "forwarded many times" tags are

normally on stupid videos or sort of like pictures of like […] jokes and stuff like that. […] It's like the old chain emails and texts [you] used to get. […] You'd get the same joke off ten people when I was like, sort of 18, 19, and now they're just forward[ed] round WhatsApp.

[46] Cancelas-Ouviña (2021); Cruz-Moya & Sánchez-Moya (2021).

David's cynicism hints at how this sort of viral social media content quickly loses its entertainment value, as one may receive the same piece of content from multiple contacts within a short space of time. Maisie (68, F, Northeast) echoed this:

> *A topical joke […] [has] usually been forwarded so many times. […] I might think "that's quite funny" and I'll pass it on to somebody else and I usually get the "oh, I've already seen that. It's been doing the rounds."*

Josephine (50, F, North East) has a similar view. She said she has seen the "forwarded" and "forwarded many times" tags on "jokes and stuff like that, […] rude jokes. […] I look at it, and I think 'Oh, this has been forwarded many times, it's one: a joke, and two: it's gonna be something probably rude.'" Josephine is no stranger to the use of social media for entertainment, describing looking at fun and cute content about dogs on TikTok as her "favourite waste of time." However, she does not appreciate the jokes that friends forward on to her via personal messaging:

> *I have a lot of male friends who feel the need to send us [slang: me] rude jokes that have been forwarded many times, and why send them to me? I don't know if they think I've been born in the wrong body or what, […] but they're filth, and they send us them, so, yeah "forwarded many times," […] I mean, why? […] I'm not really a person that you would send a joke to.*

For Josephine, lack of due consideration for one's audience is clearly part of this problem. Forwards are seen as impersonal because the same content is sent indiscriminately to multiple contacts. However, Josephine also disapproves of participating in the chain at all. She said, "I wouldn't send something that's been sent many times, I know that." In Josephine's case this is not the result of an association between forwarded content and the irresponsible dissemination of (mis)information—in our discussion, she did not once link a content's status as "forwarded" or "forwarded many times" to the potential for it to be untrustworthy. Her disapproval seems to stem from her distaste for the inappropriate content with which she associates forwarded messages.

> **"I look at it, and I think [...] it's one: a joke, and two: it's gonna be something probably rude."**

This reflects a sense among a handful of our participants that forwards are associated with supposedly humorous content that feels irrelevant and at times even unwelcome. Eve (43, F, Northeast), who notices the "forwarded" and "forwarded many times" tags on "really silly" videos her partner forwards on from his male colleagues, said she "wouldn't miss [the videos] if they weren't there." She showed the interviewer one of these videos, which appeared to be based on a sexist joke, and added "sometimes [my partner] sends them to me and I think he forgets that I'm not one of the guys in the bus depot." Renee (58, F, East Midlands) is another participant who is frustrated with the inappropriate forwards she receives on WhatsApp. Renee said she has seen the "forwarded" and "forwarded many times" tags on jokes, and spoke of a contact who was "always sending me stuff. […] ridiculous stuff." She said,

> *I stopped conversing in the end because they send me a lot of rubbish, keep forwarding stuff onto other people. […] Don't send me this stuff. […] Meme-, meemees or something? […] I don't do it. Got better things to do.*

The association between forwards and impersonal, inappropriate or irrelevant content means that some people routinely dismiss messages that bear a "forwarded" or "forwarded many times" tag. From the perspective of protecting users from misinformation, this may seem like a positive outcome. However, it does

not constitute the critical engagement with forwarded content that would stimulate awareness of the dangers of misinformation and promote digital media literacy. In their interviews, Josephine, Eve and Renee did not demonstrate any awareness that the "forwarded" and "forwarded many times" tags were meant to mark potential misinformation. It appears they only saw the tags as relevant for marking unwanted humorous content.

This is important because, while unwanted humorous content and misinformation are both varieties of low quality content, the two have different social and political consequences, and necessitate different responses. Users need to see these as distinct if we are to promote the kinds of critical media literacy that are required to combat misinformation. In their current form, the tags do not facilitate this. This is further complicated by the fact that, as we shall see below, not all forwards contain low-quality content worthy of dismissal.

6.3. Important Content

In contrast to the interpretations discussed above, we also found evidence that a subset of UK personal messaging users associate the "forwarded" and "forwarded many times" tags with positive meanings. These include content that is socially valuable or contains important or useful information. Such associations can lead some people to prioritise information contained in these messages, which is contrary to the tags' intention. However, this association between forwards and importance is not necessarily related to misinformation. Rather, these people consider forwarding a useful way to quickly share socially important and practical information with peers.

For instance, David (31, M, North East) described his interpretation of the "forwarded" and "forwarded many times" tags thus:

> *If something's sort of happened in the social space, something like something viral in a football match or a boxing match, […] that tends to be the ones, the things you'd see forwarded many times. It's just been shared about the groups. […] Probably [the tag tells you] just that it's gone viral, [or] it's going to go viral very quickly. I don't think it really changes what's the content at all.*

David's testimony, which he based on his own experiences with forwards, alludes to the way in which forwarding affords virality, whose outcomes can be positive or negative. Interestingly, the last part of David's comment indicates that, in this context, the fact that something has been forwarded many times has no bearing on his assessment of the content itself.

"If something's sort of happened in the social space [...] that tends to be the ones, the things you'd see forwarded many times."

David's comment also demonstrates how forwarding affords real-time social participation. Forwarding on personal messaging helps create embodied experiences of participation in social events and turns them into shared social moments, as is common on more public social media platforms.[47] A number of people we spoke to seem to have experienced this feeling of event-based community. Henna (57, F, East Midlands), for example, is a film buff whose active social life revolves around going to the cinema and to pubs and cafes with fellow film society members. These habits were severely curtailed during the Covid-19 pandemic. As restrictions relaxed, Henna appreciated the way forwarding afforded her and her friends the ability to keep each other up to date and share in the embodied excitement of these events:

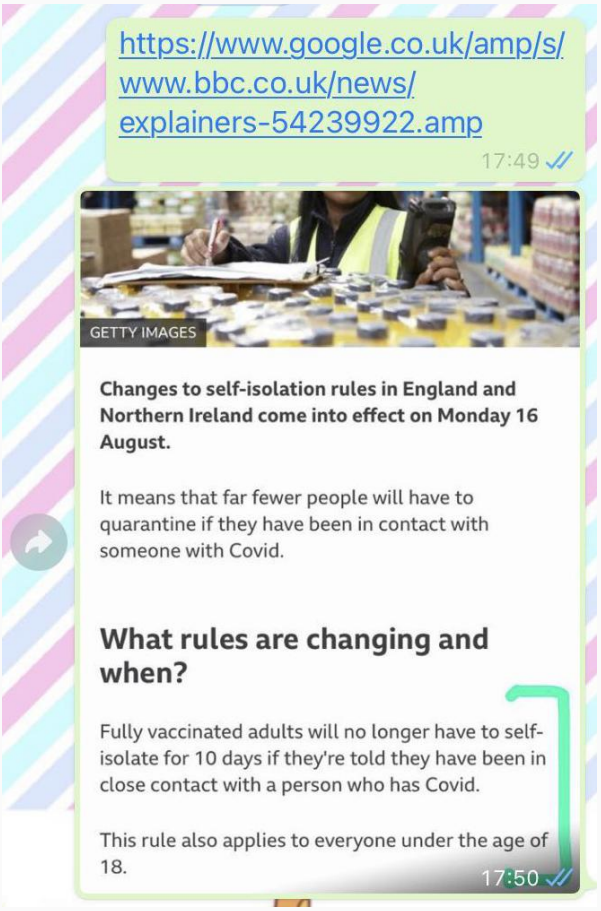
[47] Chadwick (2017); Papacharissi (2015).

> *Now that the places are opening, I have received a lot of messages and links to "We are open. […] Would you like to book a table," […] and then I forward them to my friends and go "oh my God, look, it's open! We need to book a table, let's go!"*

Some of our participants associate forwards not only with social value, but also practical value. We found evidence that forwarding is considered valuable for sharing crucial information, including around public health. Penelope (20, F, East Midlands) is an interesting example. She has plenty of experience of receiving misinformation on personal messaging and even went so far as to say that the unreliable information being posted into her WhatsApp groups comprises "mostly the forwarded things." She said that in chats with older family members on WhatsApp, there tends to be a lot of "chain" forwards, some of which have "been sent on to scare people." She is aware that "with forwarded ones, you don't know where it's come from. It could have gone through multiple people before coming to you and it gets like Chinese whispers." However, Penelope went on to say,

> *And then there are other things which are useful, such as in [Penelope's local area] we've had a mass vaccine rollout and there was some forwardeds [sic] saying where you can get it, what you need to get it, and how to apply, and that was quite useful because it got sent around to a lot of people, and a lot of people got vaccinated.*

The situation in the UK during the first two years of the pandemic was extremely fluid. Restrictions were changing frequently in response to infection rates and other political pressures, at times varying by region or even by local area. Vaccine availability conditions were also in constant flux. In this context, forwarding's affordance for passing on useful information became meaningful to some personal messaging users. The value placed on forwards during the Covid-19 pandemic is also evidenced by an example Penelope uploaded to our mobile app: a link to a BBC online article explaining new self-isolation rules. This was in the context of the UK's relaxing of Covid-19 restrictions.



The useful content Penelope shared with family members, which they then forwarded on to others

Penelope added this written comment:

> *[…] I shared this to help older family members to better understand the rules and clear any confusion due to our family having to isolate at the time during which the rules would be changing. […] the article was informative and helpful and was then forwarded onto others by the person I shared it with.*

The contrast that Penelope aptly identifies between "chain" messages of questionable provenance "sent on to scare people" and forwards like this one that are "informative and

helpful" demonstrates the dual functioning of forwards and the plurality of experiences of them. All this complexity comes to bear on the variegated associations that in turn inform people's interpretations of the "forwarded" and "forwarded many times" tags.

"That was quite useful because it got sent around to a lot of people, and a lot of people got vaccinated."

Even when participants are not aware of the existence of some tags, their reflections provide useful illustrations of how they would react if they saw them. For instance, Barry (43, M, London) said he has noticed the "forwarded" tag but could not remember ever having seen "forwarded many times." What inference would he draw from the latter if he were to come across it?

> *I suppose it's just that it has been passed around via lots of people. Therefore, it's something, not necessarily important but popular, I suppose. I'm not saying popular as in good popular. As in, popular, as in lots of people have read it.*

Barry's deliberately neutral description of the tags hits on the key barrier we identify here. Because the tags offer no explicit reference to the problem of misinformation, it is up to recipients of messages to understand the purpose of the tags. Without awareness of the implied association between forwarding and misinformation, "forwarded many times" can simply mean that the content has been shared by a large number of people, for whatever reason. Although Barry does not recall encountering tagged content that conflicts with the intended interpretation, he still does not make the association between content "forwarded many times" and potential viral misinformation. In this case, the fact that the tags do not explicitly mention the potential for misinformation is a clear barrier to their effectiveness. In fact, an association between popularity and content "forwarded many times" can even lead some people to prioritise content marked as such, and assume more rather than less trustworthiness. In other words, far from carrying negative connotations, for some users "forwarded many times" triggers a "bandwagon heuristic,"[48] meaning it signals that something is valuable and credible because many others have engaged with it. For example, Jack (23, M, North East) interprets the "forwarded" and "forwarded many times" tags as follows:

> *It makes us [slang: me] think that the message is probably of high importance, especially if it's in sort of a group chat where there's loads of messages flowing through. […] Yeah, it does make us think it's important.*

Even after the interviewer explained the intended purpose of the tags, Jack reiterated, "Again, I would probably think it was highly important." Jack is university educated and working in software engineering. He is aware of the problem of misinformation, and describes experiences of receiving dubious content from friends on personal messaging as well as dealing with family members who believe in unfounded conspiracy theories. Yet, in Jack's mind, viral forwarding bears no particular relation to this issue.

Abeni (39, F, London) also said it is "quite useful" to know that something has been forwarded many times and elaborated:

> *That means maybe because the subject is interesting, is like really important, yeah, so, that makes me want to read it. […] Because sometimes I don't wanna waste my time too much. So, sometimes I don't [read certain messages], but if it says that it's been forwarded [or] I see like it's been forwarded so many times, so I think, maybe, that tells you that it's really important, you know, that you have to*

[48] Sundar (2008).

read it. So, yeah, […] I would be more interested to read that.

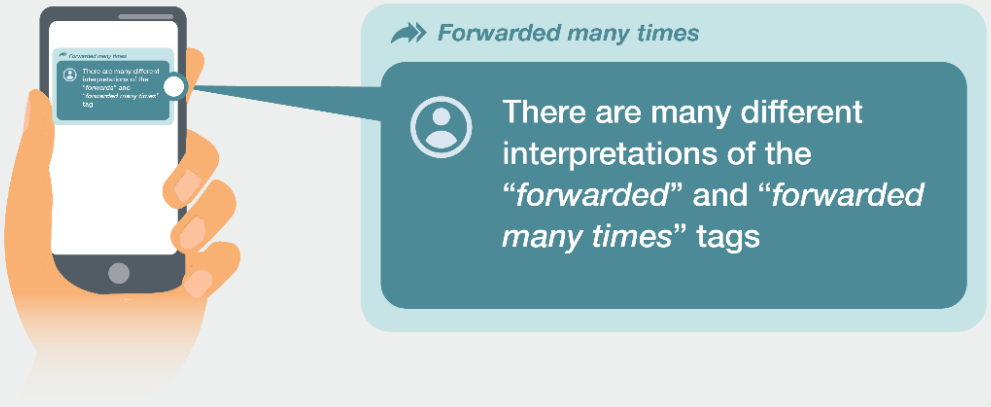
Discussion across our two interviews with Abeni revealed that some of the content she receives on WhatsApp contains misinformation from trusted peers. For instance, she disclosed how members of her family group on WhatsApp implored her daily not to take her Covid-19 vaccination, arguing it was unsafe and would cause her physical harm. In her second interview she reported that she ended up having her first dose of the vaccination, but far from convinced of the safety of vaccines, she remains too afraid to go for her second. This is despite Abeni and her children contracting Covid in the meantime and experiencing symptoms for about one month that she described as "really quite scary."

"I see like it's been forwarded so many times, so I think, maybe, that tells you that it's really important […] I would be more interested to read that."

At the end of her participation in our study, Abeni told us that she had always "wait[ed] for people to tell me" whether something was misinformation. Until participating in our study, she had "never tried before […] [to] investigate more to find out myself." Abeni's case raises the important question of whether the intended effect of technical solutions such as the "forwarded" and "forwarded many times" tags are reaching those who might need the most help identifying misinformation. An opportunity for Meta to use the tags to raise awareness about misinformation is missed, due to their deliberately vague wording.

Responses to receiving a forwarded WhatsApp message

Not everyone responds in the intended way



What people said...

These responses were based on our study...

Key

- Desirable response
- Undesirable response

It's misinformation — *I know this because I'm aware of the tag* → Content not shared

It's not welcome — *I think this is inappropriate content* → Content not shared

It's a normal message — *I am unaware of this tag* → No effect on user

It's important content — *I've seen the tag on socially-valuable, practical and important information* → More often shared

User thoughts — *User reasoning* → User action

7. Conclusion

WhatsApp's “forwarded” and “forwarded many times” tags are a light touch intervention to tackle misinformation. The tags are intended to prompt critical reflection on the source and the veracity of the message content, particularly before passing it on further. This process relies on people associating the tags with misinformation. Our in-depth research with 102 members of the public, however, emphasises that this association is not always made. Our participants demonstrated multiple conflicting interpretations of the tags—with some associations potentially increasing the spread of misinformation.

We found evidence that this reliance on users to make such associations acts as a key barrier to the effectiveness of the “forwarded” and “forwarded many times” tags. WhatsApp developed these tags in response to specific events in societies where the viral distribution of misinformation via forwards on this platform is a well-known social issue. In the absence of knowledge about the events surrounding the tags’ introduction, or experiences of receiving large volumes of misinformation-related forwards on personal messaging, people ignore the tags or imbue them with their own meanings. This gives rise to interpretations and behaviours that do not necessarily promote the kind of critical media literacy needed for people to be able to identify forwarded misinformation.

Based on their own experience of the forwards they receive, some UK personal messaging users associate forwards not with potential misinformation, but with viral and unwelcome jokes. This can prompt automatic dismissal of content bearing a “forwarded” and “forwarded many times” tag. Inappropriate jokes and misinformation can both be considered varieties of low-quality content, whose virality WhatsApp may wish to limit. However, the former, while unwelcome, is seen as harmless, while the latter carries real social, political, and public health dangers, and requires a different level of critical reflection by recipients.

Other UK personal messaging users even associate forwards with positive meanings. This reflects the variegated uses of the forward feature, particularly in times of public health crisis when it can be used to alert others of potentially life-saving information. However, where the “forwarded many times” tag in particular is seen to signal important, high quality or popular content, it runs the risk of being seen as a sign that content should be prioritised. This highlights the importance of fostering a clear understanding among people of the intended purpose of misinformation warnings.

7.1. Limitations of this Report

The findings from this report emerge from a qualitative analysis of in-depth interviews with 102 UK-based participants. Situating discussion of the “forwarded” and “forwarded many times” tags within broader attitudes towards forwards has allowed us to present a rich, context-sensitive evaluation of the role this feature does or does not play. These findings are, however, explorative. Studies based on larger samples that are fully representative of the UK population are required in order to draw conclusions about the prevalence of these practices across the UK population as a whole.

8. Implications

How can these insights about WhatsApp's "forwarded" and "forwarded many times" tags be applied to misinformation warnings on personal messaging platforms more broadly?

The vague nature of the tags is not an outcome of end-to-end encryption, but rather is a design choice made by Meta in order to avoid continuously prompting negative associations between WhatsApp and harmful content. Cases such as these show how such corporate decisions can get in the way of warning tags being useful measures against the spread of misinformation.

Fundamentally, we need to understand misinformation warnings as affordances. This means examining the relationship between app design features, users' individual attributes and abilities, and complex situations of use. The latter are partly made up of social norms and interpersonal relationships. This contextual approach is especially important when we consider that on personal messaging there are no algorithms that amplify or filter out misinformation, as in public social media platforms such as Facebook or Twitter. In the hybrid public-interpersonal communication environments constituted by personal messaging, combating misinformation is a matter not only of technology, but interpersonal exchanges, social norms, situations of use and individual experience.

On this basis, we put forward five principles that personal messaging platforms should consider when designing such measures. These principles can be implemented without compromising end-to-end encryption, which we anticipate and hope will continue. These are meant to serve as a foundation for further research, and we encourage researchers to build on the evidence presented here to continue to investigate how misinformation warnings can be made most effective in practice.

1. Don't Rely on Description Alone

Warnings that merely describe which functions have been used to send a message (such as that the message has been "forwarded") may not prompt an association with potential misinformation and, in turn, the desired critical reflection and due consideration before re-sharing. Users' understandings of features arise from different contexts of platform use and cannot necessarily be predicted. Therefore, misinformation warnings should include explicit wording about the risk of misinformation and the need for vigilance, or they run the risk of eliciting unintended and contradictory interpretations. This would mean a compromise on the part of personal messaging platforms. More vague tags may be in their corporate interest, as they avoid negative associations with their brand. But mitigating online harms should be the priority.

2. Introduce User Friction

Misinformation warnings that do not compel a user response are more likely to be ignored. This poses a risk to their effectiveness, particularly when the warning is plain and unobtrusive. We found evidence that people may come across such warnings but not recall doing so. Introducing friction in the user experience can help draw attention to misinformation risks. Features that could help improve the effectiveness of warnings like the "forwarded" and "forwarded many times" tags include:

- Marking tagged messages with a different colour to make them stand out
- Covers that require users to click to reveal message content

- Asking users to confirm they have considered the veracity of the message and are sure they want to forward it on.

These designs will help ensure that people notice and engage with the warnings. But their intrusiveness will have to be balanced with the fact that in some contexts only a minority of forwarded content will be misinformation.

3. Gain Media Exposure

Publicity campaigns by personal messaging platforms and their parent companies—in this case Meta—can help spread the word about the intention of misinformation warnings and thus improve their effectiveness. These can involve traditional news media as well as online and social media outlets. Active efforts to raise awareness are particularly needed where warnings are vague and are open to different interpretations, or in contexts where awareness of the role of personal messaging in misinformation may be low. We found evidence that even among those who did not have direct experience with forwarded misinformation, learned knowledge about the context behind the "forwarded" and "forwarded many times" tags' introduction facilitated their effectiveness. Platforms have the resources and the responsibility to work with different media outlets to influence how people understand and react to their misinformation warnings.

4. Consider the Context

Understanding of the divergent ways in which personal messaging platforms are used across social contexts is crucial to the design of misinformation warnings that are relevant and useful. There are likely to be differences in the ways particular social groups use personal messaging, including different age groups, different local communities, or those with different degrees of interest in communication technology. Variations in patterns of use will affect who is more or less exposed to forwarded misinformation. Furthermore, for some groups, message characteristics other than being forwarded might be more salient markers of potential misinformation. Platforms need to consider in what ways misinformation spreads in specific contexts, and whether a broader variety of anti-misinformation measures are needed.

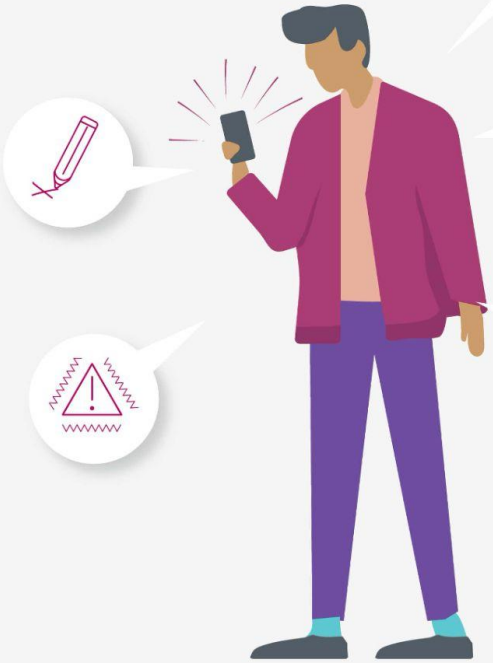
5. Think Beyond the Platforms

In addition to considering all of the above, it is important that personal messaging platforms recognise the limitations of within-platform measures for tackling misinformation. Misinformation is a complex social problem that cannot be wholly addressed through the introduction of new technical features alone. As personal messaging platforms are hybrid public-interpersonal communication environments, relationships and social norms are crucial.[49] Understanding the complexities of the interactions within which misinformation is shared, ignored, or challenged on personal messaging platforms is key. Technical features need to be combined with socially-oriented anti-misinformation interventions if they are to be successful at reducing the spread of misinformation on personal messaging. These should focus on building social capacities—that is, empowering people to talk about or challenge misinformation within their social networks and to work together to use personal messaging platforms in ways that help reduce misinformation.

[49] Chadwick, Vaccari & Hall (2022); Chadwick, Vaccari & Hall (2023); Chadwick, Hall, & Vaccari (2023); Hall, Chadwick, & Vaccari (forthcoming).

Principles for effective misinformation warnings

Advice for personal messaging platforms



1 Don't rely on description alone

Tags should include an explicit warning of the potential for misinformation.

2 Introduce user friction

Features that require active user confirmation may be necessary to ensure misinformation warnings are noticed and engaged with.

3 Gain media exposure

PR campaigns by the platforms themselves can help users recognise misinformation warnings and react appropriately.

4 Consider the context

Design of misinformation warnings must appreciate that users' experiences of platforms are set within local contexts.

5 Think beyond the platforms

There must be a recognition that misinformation is a social problem that cannot be wholly addressed by introducing new platform features.

References

Avelar, D. (2019, October 30). WhatsApp fake news during Brazil election 'favoured Bolsonaro'. *The Guardian*. https://www.theguardian.com/world/2019/oct/30/whatsapp-fake-news-brazil-election-favoured-jair-bolsonaro-analysis-suggests.

Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana Pravin, M. (2019). WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. Department of Media and Communications, London School of Economics and Political Science. http://www.lse.ac.uk/media-and-communications/research/research-projects/whatsapp-vigilantes.

Baulch, E., Matamoros-Fernandez, A., & Johns, A. (2020). Introduction: Ten years of WhatsApp: The role of chat apps in the formation and mobilization of online publics. *First Monday*, *25*(1), Article 1. https://eprints.qut.edu.au/135573/.

Bhushan, K. (2019, August 2). WhatsApp now tells users when a message has been forwarded many times. *HT Tech*. https://tech.hindustantimes.com/tech/news/whatsapp-now-tells-users-when-a-message-has-been-forwarded-many-times-story-IRFr4I0jLGFM0VNQZLrFPP.html.

Cancelas-Ouviña, L.-P. (2021). Humor in Times of COVID-19 in Spain: Viewing Coronavirus Through Memes Disseminated via WhatsApp. *Frontiers in Psychology, 12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.611788.

Chadwick, A. (2017). *The Hybrid Media System: Politics and Power*. 2nd edition. Oxford University Press.

Chadwick, A., Hall, N-A., & Vaccari, C. (2023). Misinformation Rules?! Could "Group Rules" Reduce Misinformation in Online Personal Messaging? *New Media and Society*, OnlineFirst, https://doi.org/10.1177/14614448231172964.

Chadwick, A., Vaccari, C., & Hall, N-A. (2023). What Explains the Spread of Misinformation in Online Personal Messaging Networks? Exploring the Role of Conflict Avoidance. *Digital Journalism*, SI: Meso News Spaces and Beyond, OnlineFirst, https://doi.org/10.1080/21670811.2023.2206038.

Chadwick, A., Vaccari, C., & Hall, N-A. (2022). Covid Vaccines and Online Personal Messaging: The Challenge of Challenging Misinformation. Online Civic Culture Centre, Loughborough University, https://www.lboro.ac.uk/media/media/research/o3c/pdf/Chadwick-Vaccari-Hall-Covid-Vaccines-and-Online-Personal-Messaging-2022.pdf.

Choudhary, V. (2018, July 10). 'Forwarded' label now live on WhatsApp to combat fake news. *HT Tech*. https://tech.hindustantimes.com/tech/news/forwarded-label-now-live-on-whatsapp-to-combat-fake-news-story-kw0pzLZpWB7LDyQkotWrXP.html.

Costa, E. (2018). Affordances-in-practice: An ethnographic critique of social media logic and context collapse. *New Media & Society, 20*(10), 3641–3656. https://doi.org/10.1177/1461444818756290.

Cruz-Moya, O., & Sánchez-Moya, A. (2021). Humour in multimodal times: Insights from online interactions among senior users of a WhatsApp group. *Internet Pragmatics, 4*(1), 52–86. https://doi.org/10.1075/ip.00061.cru.

Dixon, S. (2022, July 27). Most Popular Messaging Apps 2022. *Statista*. https://www.statista.com/statistics/258749/most-popular-global-mobile-messenger-apps/.

Evans, S. K., Pearce, K. E., Vitak, J., & Treem, J. W. (2017). Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research. *Journal of Computer-Mediated Communication, 22*(1), 35–52. https://doi.org/10.1111/jcc4.12180.

Hall, N-A., Chadwick, A. and Vaccari, C. (forthcoming). Online Misinformation and Everyday Ontological Narratives of Social Distinction. *Media, Culture & Society*.

Hern, A. & Milmo, D. (2023, May 8). WhatsApp could disappear from UK over privacy concerns, ministers told. *The Guardian*. https://www.theguardian.com/technology/2023/may/08/whatsapp-could-disappear-uk-over-privacy-concerns-ministers-told.

Johns, A., & Cheong, N. (2021). The affective pressures of WhatsApp: From safe spaces to conspiratorial publics. *Continuum, 35*(5), 732–746. https://doi.org/10.1080/10304312.2021.1983256.

Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2021). Adapting Security Warnings to Counter Online Disinformation. 1163–1180. https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser.

Kazemi, A., Garimella, K., Shahi, G. K., Gaffney, D., & Hale, S. A. (2022). Tiplines to uncover misinformation on encrypted platforms: A case study of the 2019 Indian general election on WhatsApp. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-91.

Leonardi, P. M. (2013). Theoretical foundations for the study of sociomateriality. *Information and Organization, 23*(2), 59–76. https://doi.org/10.1016/j.infoandorg.2013.02.002.

Lu, X., Vijaykumar, S., Jin, Y., & Rogerson, D. (2022). Think before you Share: Beliefs and emotions that shaped COVID-19 (Mis)information vetting and sharing intentions among WhatsApp users in the United Kingdom. *Telematics and Informatics, 67*, 101750. https://doi.org/10.1016/j.tele.2021.101750.

Maddox, J., & Kanthawala, S. (2022). The Revolution Will Be Forwarded: Interrogating India's WhatsApp Imaginary. *Journal of Communication Inquiry*, *47*(3), 249-267. https://doi.org/10.1177/01968599221095177.

Majchrzak, A., Faraj, S., Kane, G. C., & Azad, B. (2013). The Contradictory Influence of Social Media Affordances on Online Communal Knowledge Sharing. *Journal of Computer-Mediated Communication, 19*(1), 38–55. https://doi.org/10.1111/jcc4.12030.

Martínez, M. (2018, November 12). Burned to death because of a rumour on WhatsApp. *BBC News*. https://www.bbc.com/news/world-latin-america-46145986.

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*(8), 1062–1084. https://doi.org/10.1080/21670811.2021.1927778.

Melo, P. de F., Vieira, C. C., Garimella, K., de Melo, P. O. S. V., & Benevenuto, F. (2019). Can WhatsApp Counter Misinformation by Limiting Message Forwarding? (arXiv:1909.08740). arXiv. http://arxiv.org/abs/1909.08740.

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. https://doi.org/10.1016/j.pragma.2013.07.012.

Molteni, M. (2018, March 9). When WhatsApp's Fake News Problem Threatens Public Health. *Wired*. https://www.wired.com/story/when-whatsapps-fake-news-problem-threatens-public-health/.

Mukherjee, R. (2020). Mobile witnessing on WhatsApp: Vigilante virality and the anatomy of mob lynching. *South Asian Popular Culture, 18*(1), 79–101. https://doi.org/10.1080/14746689.2020.1736810.

Nobre, G. P., Ferreira, C. H. G., & Almeida, J. M. (2022). A Hierarchical Network-Oriented Analysis of User Participation in Misinformation Spread on WhatsApp. *Information Processing & Management, 59*(1), 102757. https://doi.org/10.1016/j.ipm.2021.102757.

OFCOM. (2021). *Online Nation Annual Report 2021*. https://www.ofcom.org.uk/research-and-data/internet-and-on-demand-research/online-nation.

Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

Pasquetto, I. V., Jahani, E., Atreja, S., & Baum, M. (2022). Social Debunking of Misinformation on WhatsApp: The Case for Strong and In-group Ties. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW1), 1-35. https://doi.org/10.1145/3512964.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Management Science, 66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.3478

Rand, A., & Andrey, S. (2021, June 21). Private messages contribute to the spread of COVID-19 conspiracies. *The Conversation*. http://theconversation.com/private-messages-contribute-to-the-spread-of-covid-19-conspiracies-162725

Sahoo, S. (2022). Political Posters Reveal a Tension in WhatsApp Platform Design: An Analysis of Digital Images From India's 2019 Elections. *Television & New Media, 23*(8), 874–899. https://doi.org/10.1177/15274764211052997.

Sarkar, S. (2022). 'No forwards please': Indian mothers on school WhatsApp groups. In F. J. Green & J. M. Rogers (Eds.), *Parenting/Internet/Kids: Domesticating Technologies*. Demeter Press.

Saurwein, F., & Spencer-Smith, C. (2020). Combating Disinformation on Social Media: Multilevel Governance and Distributed Accountability in Europe. *Digital Journalism, 8*(6), 820–841. https://doi.org/10.1080/21670811.2020.1765401.

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2021). Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes (arXiv:2104.00779). arXiv. https://doi.org/10.48550/arXiv.2104.00779.

Sharma, S. (2018, July 13). WhatsApp's 'Forwarded' label to curb fake news has a loophole | Digit. Digit.In. https://www.digit.in/news/apps/whatsapps-first-step-towards-curbing-fake-news-is-forwarded-label-but-it-has-a-loophole-42196.html.

Sircar, S. (2018, July 25). India Tops the World... In Forwarding Messages! Umm, Why Though? *TheQuint*. https://www.thequint.com/voices/opinion/reasons-why-indians-forward-many-whatsapp-messages.

Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital Media, Youth, and Credibility* (pp. 72–100). MIT Press.

Swart, J., Peters, C., & Broersma, M. (2019). Sharing and Discussing News in Private Social Media Groups. *Digital Journalism, 7*(2), 187–205. https://doi.org/10.1080/21670811.2018.1465351.

Tandoc, E. C. Jr., Rosenthal, S., Yeo, J., Ong, Z., Yang, T., Malik, S., Ou, M., Zhou, Y., Zheng, J., Mohamed, H. A. B., Tan, J., Lau, Z. X., & Lim, J. Y. (2022). Moving Forward Against Misinformation or Stepping Back? WhatsApp's Forwarded Tag as an Electronically Relayed Information Cue. *International Journal of Communication, 16*. 1851-1868. https://ijoc.org/index.php/ijoc/article/view/18138.

Vasudeva, F., & Barkdull, N. (2020). WhatsApp in India? A case study of social media related lynchings. *Social Identities, 26*(5), 574–589. https://doi.org/10.1080/13504630.2020.1782730.

WhatsApp Blog. (2018, July 10). Labeling Forwarded Messages. *WhatsApp.Com*. https://blog.whatsapp.com/labeling-forwarded-messages.

WhatsApp Blog. (2019, January 21). More changes to forwarding. *WhatsApp.Com*. https://blog.whatsapp.com/more-changes-to-forwarding.

WhatsApp. (2020, February 12). Two Billion Users—Connecting the World Privately. *WhatsApp.Com*. https://blog.whatsapp.com/two-billion-users-connecting-the-world-privately/?lang=en

WhatsApp. (n.d.-a). About forwarding limits | WhatsApp Help Center. *WhatsApp Help Center*. Retrieved 1 December 2022, from https://faq.whatsapp.com/759286787919153/?locale=en_US

WhatsApp. (n.d.-b). How to prevent the spread of misinformation | WhatsApp Help Center. *WhatsApp Help Center*. Retrieved 1 December 2022, from https://faq.whatsapp.com/431498999157251

Xie, J., Yamashita, M., Cai, Z., & Xiong, A. (2022). A User Study on the Feasibility of Topic-aware Misinformation Warning on Social Media. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 66*(1), 621–625. https://doi.org/10.1177/1071181322661252

Appendix

A1. Data and Research Method

Ethical approval for this study was granted by Loughborough University's Ethics Review Subcommittee (2021-4516-3252; PI Chadwick).

We hired established opinion polling company Opinium Research to recruit participants. Opinium maintains its own panel of more than 40,000 members of the UK public who participate in surveys and market research. Opinium is a member of the British Polling Council, the Market Research Society, and the European Society for Opinion and Marketing Research (ESOMAR).

Sampling and Recruitment

We recruited people who used at least one of the following apps at least a few times a week: WhatsApp, Facebook Messenger, iMessages, Android Messages, Snapchat, Telegram, Signal. Figure A1 shows which apps our participants used and how frequently.

To reduce self-selection bias and ensure sincere responses at the screening stage, potential participants were not informed of the precise topic of the in-depth interviews before they provisionally agreed to take part. After screening, each participant was then provided with full Loughborough University research ethics committee approved information describing the study. Each participant completed an informed consent form before their first interview.

Demographics

A screening questionnaire ensured that the demographic composition of our participants roughly reflects the diversity of the UK population across gender, age, ethnicity, educational attainment, and basic digital literacy. Figures A2 and A3 display the distributions across these variables. Exact matching is impossible and of little meaningful value in a qualitative study with a sample of 102 participants, but our recruitment method meant we avoided over-recruiting from a narrow range of social groups. This is particularly important when using online panels. One of the Everyday Misinformation Project's aims is to explore the role of community and neighbourhood in shaping sharing on personal messaging. With this in mind, we recruited participants who resided in three distinct regions: London, the East Midlands, and the North East.

Basic Digital Literacy

To ensure that participants' basic digital literacy roughly matched the distribution across the UK adult population we used a single screening question we took from OFCOM's Adult Media Literacy Tracker survey:

"When you use a search engine to find information, you enter a query in the search box and the search engine will then show some links to websites in the results pages. Which one of these is closest to your opinion about the level of accuracy or bias of the information detailed in the websites that appear in the results pages?"

1. I think that if they have been listed by the search engine, these websites will have accurate and unbiased information
2. I think that some of the websites will be accurate or unbiased and some won't be
3. I don't really think about whether or not they have accurate or unbiased information, I just use the sites I like the look of.
4. Don't know.

The distribution on this question in OFCOM's 2020 survey was 21% for option 1, 60% for option 2, and 19% for option 3[50]. The distribution among our participants deviated only slightly from the national distribution. See Figure A2. Respondents who chose "Don't know" to this question were screened out and were not interviewed.

Calibration

We employed an iterative sampling strategy, with six recruitment rounds on a rolling schedule. This allowed us to adjust for discrepancies between potential participants selected from Opinium's panel and the individuals who progressed to interview stage. For example, the first group of interviewees scored relatively highly on educational attainment and London residence. So, in the second round, we balanced the sample by recruiting on North East and East Midlands residence and lower educational attainment. We continued to calibrate recruitment in this way throughout the five-month fieldwork for this first phase of work (April-November 2021), adjusting as necessary to ensure balance on demographics, personal messaging use, and basic digital literacy across the sample of actual interviewees.

[50] OFCOM (2020). Adults' media literacy. https://www.ofcom.org.uk/__data/assets/pdf_file/0028/196372/adults-media-use-and-attitudes-2020-data-tables.pdf

Figure A1. Participants' Use of Personal Messaging (n=102)

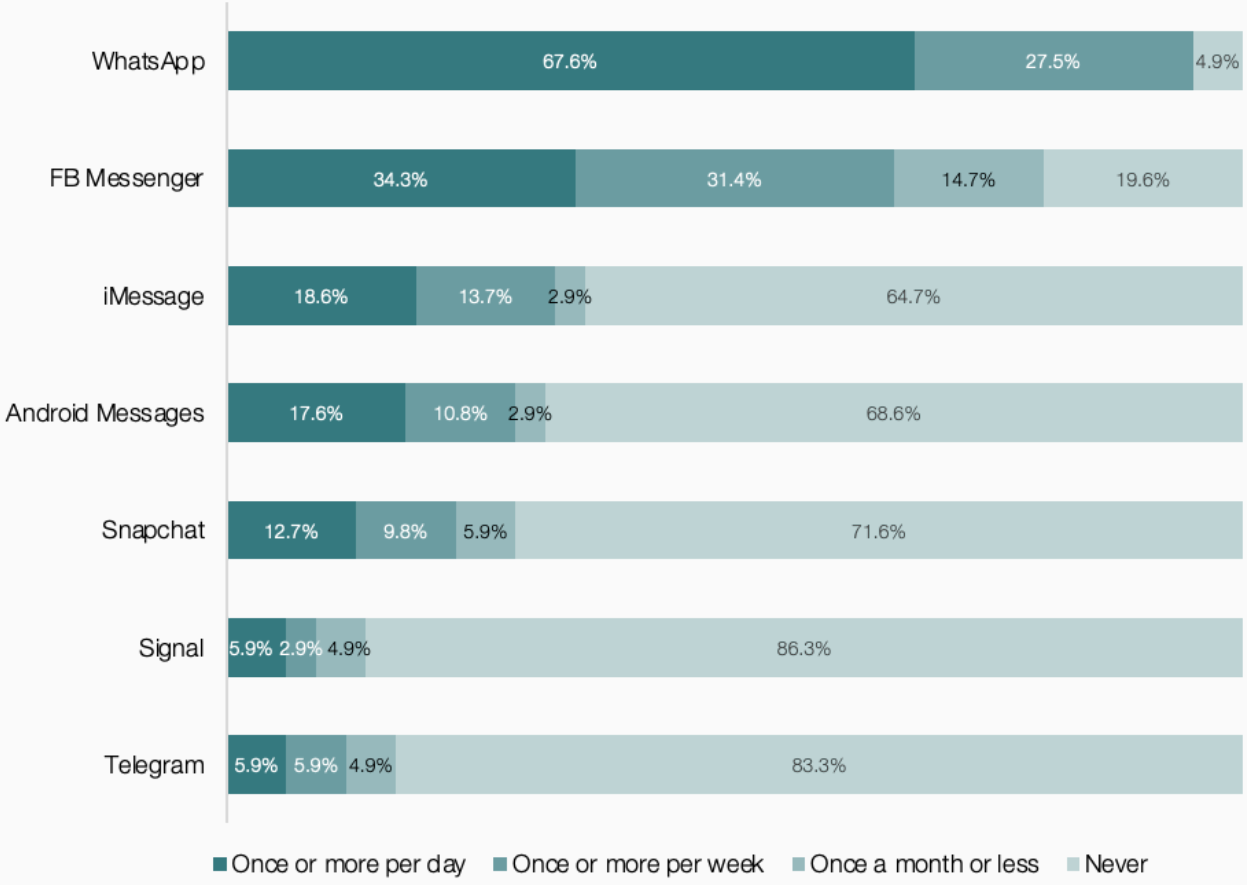


Figure A2. Participant Characteristics (n=102)

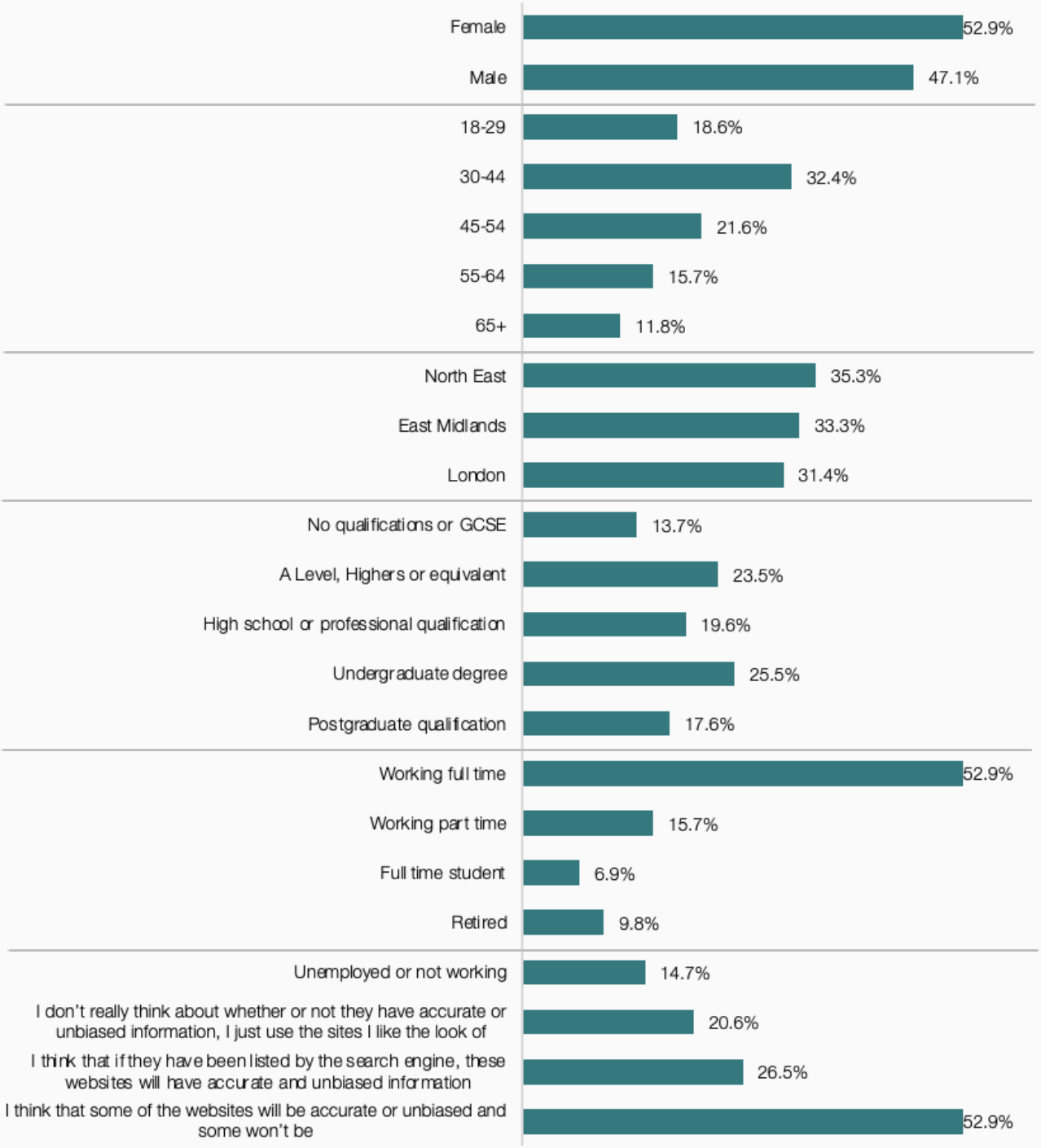
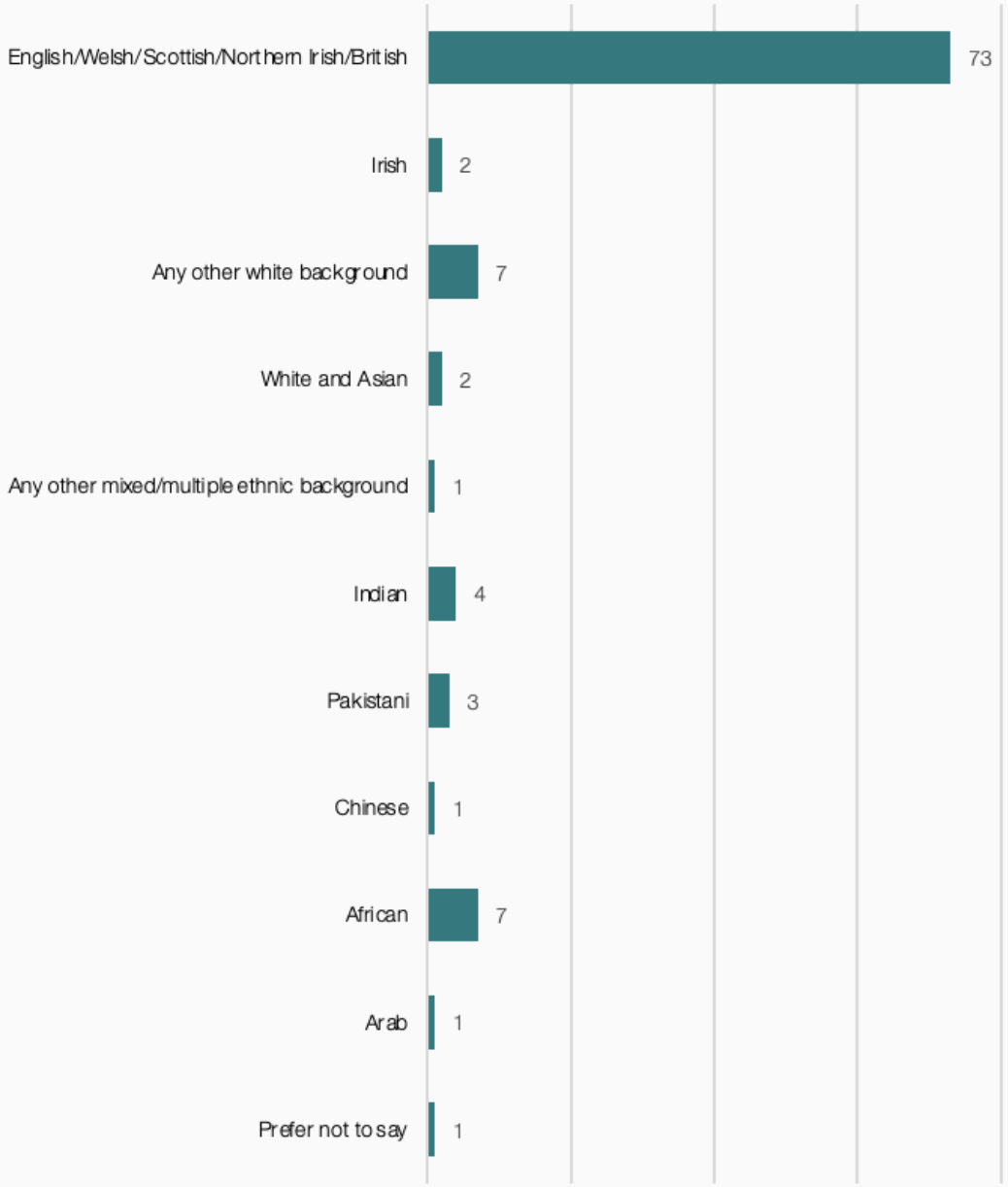


Figure A3. Participant Characteristics: Ethnicity (n=102)



Procedure

Due to pandemic social distancing, all interviews were held on Zoom, video-recorded and fully transcribed. Interviews were semi-structured and guided by an indicative list of themes we finalised following pilot interviews we conducted before fieldwork formally commenced. Interviews averaged about an hour and five minutes in length.

We then asked participants to securely contribute examples of 'information' they had sent or received on personal messaging, via a dedicated smartphone app. To avoid ex-ante definitions of misinformation, we asked participants to categorise content as either 'accurate and helpful' or 'false, inaccurate or misleading'.

Finally, we invited participants to a second interview session. We used these sessions to discuss changes over time and ask participants to elaborate on their interpretations of the content they uploaded. These interviews lasted about one hour each, and 80 of the original 102 participants took part.

A2. Disclosure and Integrity Statement

The research received funding from the Leverhulme Trust (RPG-2020-019; PI Chadwick).

Andrew Chadwick is currently an advisory board member (unpaid) of Clean Up The Internet. Cristian Vaccari is currently a co-rapporteur (paid) for the Council of Europe's Committee of Experts on the Integrity of Online Information and an advisory board member (unpaid) of Clean Up The Internet. Any opinions in this report are those of its authors and not those of funders, affiliates, or other advisory board members.

This report is a summary of ongoing academic research and has been written for a broad readership. It has not undergone formal academic peer-review prior to publication. To stay up to date with peer-reviewed academic publications from this project as they publish, please visit the website at https://everyday-mis.info.

About the Online Civic Culture Centre (O3C)

Established in February 2018 with initial funding award from Loughborough University's Adventure Research Programme, the Online Civic Culture Centre (O3C) seeks to understand the role of social media in shaping our civic culture. Led by Professor Andrew Chadwick, it features academic staff and postdoctoral and doctoral researchers drawn from the disciplines of communication, social psychology, sociology, and information science. O3C enables teams of researchers to work together on issues of misinformation, disinformation, and hate speech and intolerance online. It develops evidence-based knowledge to inform policy and practice to mitigate the democratically dysfunctional aspects of social media. For more information, visit the O3C website and follow O3C on Twitter.

Online Civic Culture Centre (O3C)
Department of Communication & Media
School of Social Sciences & Humanities
Loughborough University
Loughborough
LE11 3TT
United Kingdom

@O3CLboro

lboro.ac.uk/research/online-civic-culture-centre

O3C 4 / June 2023