

# Research Update

## Misinformation on Personal Messaging—Are WhatsApp’s Warnings Effective?

PUBLIC REPORT

Findings now confirmed in a nationally-representative survey of the UK public

Natalie-Anne Hall

Andrew Chadwick

Cristian Vaccari

Brendan T Lawson

Portia Akolgo

# Table of Contents

---

<b>This Research Update</b>	<b>3</b>
Recap: Messaging, Forwarding, and Misinformation	3
Infographic: What's Up with WhatsApp's Forwarded Tags?	4
Our Earlier Report	5
Today's New Evidence	5
<b>Summary of Key Findings</b>	<b>6</b>
<b>The Findings in Detail</b>	<b>7</b>
Categorising Responses	7
Overall Pattern of Responses	8
From Categorising Responses to Categorising People	9
What Categorising People Reveals	10
Exploring Links Between Social Factors and (Mis)perceiving the Tags' Function	11
Age	11
Educational Attainment	12
Levels of Personal Messaging Use	12
Trust in Information on Personal Messaging	13
Chat Settings and Group Size	14
<b>Conclusions</b>	<b>15</b>
Next Steps	16
1. Don't Rely on Description Alone	16
2. Introduce User Friction	16
3. Gain Media Exposure	17
4. Consider the Context	17
5. Think Beyond the Platforms	17
Infographic: Principles for Effective Misinformation Warnings	18
<b>Data and Research Method</b>	<b>19</b>
Sampling	19
Comparing the Demographic Characteristics of our Sample with the UK Population	19
<b>About the Everyday Misinformation Project</b>	<b>20</b>
<b>About the Authors</b>	<b>21</b>
<b>Disclosure and Integrity Statement</b>	<b>22</b>
<b>Notes</b>	<b>23</b>
<b>References</b>	<b>24</b>
<b>About the Online Civic Culture Centre (O3C)</b>	<b>25</b>

# This Research Update

---

This report provides new, population-level findings that confirm and expand the exploratory findings in our Online Civic Culture Centre June 2023 report, [Beyond Quick Fixes: How Users Make Sense of Misinformation Warnings on Personal Messaging](#).

In that earlier report, we reported findings from the [Everyday Misinformation Project](#). We asked: **How do personal messaging users understand WhatsApp’s “forwarded” and “forwarded many times” tags?** Insights from that qualitative, exploratory study with 102 members of the public cast serious doubt on whether these tags are effective as misinformation warnings.

The new evidence we present today comes from our nationally-representative survey of 2,000 members of the public, which we conducted in September 2023. This allows us to generalise about how those among the UK public who use personal messaging interpret the “forwarded” and “forwarded many times” misinformation warning tags.

## Recap: Messaging, Forwarding, and Misinformation

Personal messaging is extremely popular. WhatsApp is used by 79 percent of the UK adult population.<sup>1</sup> It weaves together social and personal interactions with discussion of news, politics, science, health, and many other publicly important topics. We therefore dub it *hybrid public-interpersonal* communication.<sup>2</sup> This hybrid

character means misinformation can and does make its way into people's everyday, interpersonal exchanges.

But misinformation is a difficult problem to tackle—for regulators and for the platforms themselves. WhatsApp’s end-to-end encryption means automated moderation, fact-checking, and content removal are not possible in the way they are on other forms of social media. WhatsApp is committed to end-to-end encryption, with user privacy a key selling point of the service. But this means harmful misinformation can spread unmonitored and unchecked. The burden is on people themselves to identify, challenge, and correct it.

WhatsApp’s message “forwarding” feature can be particularly conducive to the spread of misinformation. Due to encryption, forwarded messages come with no metadata about their origins. And the ability for people to forward messages to multiple users at once potentially enables the exponential diffusion of misleading content.


In 2018 and 2019, WhatsApp’s forwarding feature was implicated in a series of high-profile, misinformation-fuelled events. These included deadly mob violence in India and Mexico, harmful vaccine misinformation in Brazil, and election manipulation in Brazil and India.<sup>3</sup> These were complex events also rooted in broader social problems. Nonetheless, WhatsApp’s large presence in these countries’ communication markets saw it face pressure from media, the public, and governments to more decisively tackle misinformation on its service.<sup>4</sup>


The “forwarded” and “forwarded many times” tags are a light-touch variety of *misinformation warning*. Introduced in response to the high-profile, sometimes violent, events linked to forwarded misinformation on the platform, they are meant to prompt users to consider the source of forwarded information and take a moment to reflect on its accuracy. The infographic on the next page gives more details.

# What's up with WhatsApp's forwarded tags?

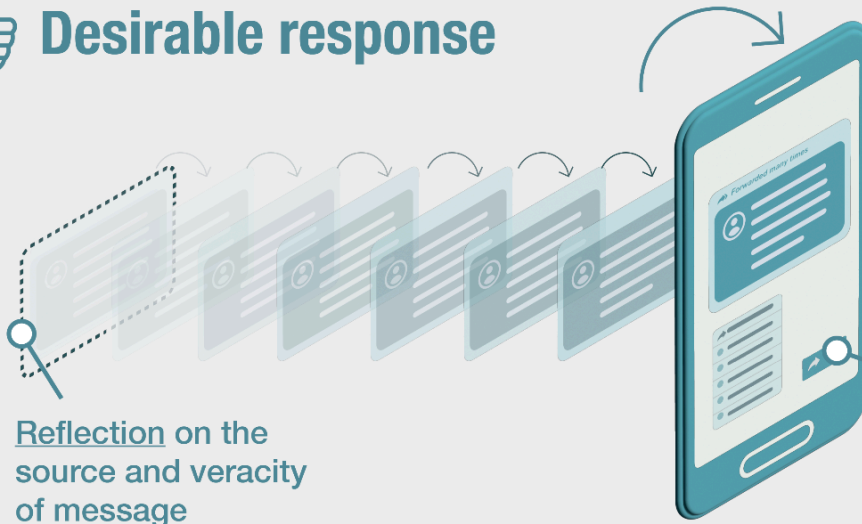
How a user should respond when they see the tag


**WHAT ARE THE TAGS?**

 *Forwarded*  
This means the message has been forwarded from another user.

 *Forwarded many times*  
This means the content has been forwarded 5 or more times.

## Desirable response




 Can reduce spread of misinformation

Due consideration before passing on further

## Undesirable response



 Can increase the spread of misinformation

No consideration before passing on further

## Our Earlier Report

Contrary to Meta’s intentions with the “forwarded” and “forwarded many times” tags, [our June 2023 report](#) found that the UK public has highly variable understandings of these tags and what they denote. Based on in-depth interview-based fieldwork with 102 personal messaging users in the UK over about a year and a half, we showed that this wide variety of interpretations throws into doubt these tags’ effectiveness as misinformation warnings.

In summary, in our earlier report we found:

- The tags’ effectiveness must **rely on an association between forwards and misinformation** in the minds of users. But those who think they do not usually receive forwards containing misinformation are less familiar with this association.
- Some **associate forwards with viral and unwelcome jokes** because they routinely receive this forwarded content. They might therefore dismiss tagged content, but not critically engage with its veracity or origin.
- Others associate forwarded messages with more desirable characteristics, including being carriers of **valuable or useful information**. A minority of users even saw the tags as signalling **high-quality** information.
- Others are **unaware of or indifferent to the tags**.
- Seeing news media coverage of the reason the tags were introduced helped users make sense of these measures, pointing to the **valuable role of broader campaigns in promoting awareness**.

Based on these findings, we warned that the effectiveness of the “forwarded” and “forwarded many times” tags is currently limited. We therefore proposed five key principles for all messaging platforms to use when designing these and similar misinformation warnings in future. We reiterate these five principles later in

this current report. But first, what did our new survey find?

## Today’s New Evidence

In September 2023, the Everyday Misinformation Project surveyed a nationally-representative sample of 2,000 members of the public. Participants were drawn from Opinium Research’s national respondent panel.

To assess perceptions of the forwarded tags we designed a multiple response question:

**You may have seen or heard about a label that can appear on WhatsApp messages that says “Forwarded” or “Forwarded many times.” In your opinion, what do these labels indicate that the message potentially contains?**

As we show below, the results of this survey confirm and extend our June 2023 report’s findings. **Interpretations of the tags vary widely and few people in the UK understand the “forwarded” and “forwarded many times” tags as clear markers of potential misinformation.**

We also explored some demographic, attitudinal, and behavioural factors that may be related to misinterpretations of the tags.

The findings in today’s Research Update reinforce our claim that corporate design choices, which are often aimed at reducing user friction and avoiding negative associations between a platform and the spread of misinformation, can inhibit the effectiveness of misinformation warnings.

Today we renew our recommendation that Meta should reconsider the assumptions underpinning the design of its misinformation warnings on WhatsApp. The commercial goal of avoiding negative perceptions of a platform should not get in the way of measures to combat the spread of misinformation and protect the public.

# Summary of Key Findings

Our nationally-representative survey shows that members of the UK public have widely varying interpretations of the “forwarded” and “forwarded many times” tags:

- **Widespread ambiguity and lack of awareness:** About half of UK messaging users either have never seen the tags, do not know what they signify, or have uncertain perceptions about the quality of the content attached to them.
- **Few accurate interpretations:** Only a minority (fewer than 10%) of the messaging-using public interpret the tags in ways that align with Meta’s stated anti-misinformation aims when the company introduced the tags.
- **Frequent misinterpretations:** The most common interpretation of the tags is that they denote viral or entertainment content such as jokes and videos. The tags are often not seen as prompts to consider the veracity of forwarded messages.
- **Some dangerous interpretations:** A small but significant proportion of the messaging-using public (around 10%) completely misperceives the tags’ purpose. This group sees the tags as flagging accurate, trustworthy, useful, or relevant content.

When we explored links between these outcomes and selected demographic, behavioural, and attitudinal factors, we found that:

- Younger people, and people who place a great degree of trust in what they see on personal messaging, are **most likely to completely misperceive the tags’ purpose**.
- Older messaging users, and those with lower levels of formal educational attainment are the

**least likely to be familiar with the tags** and know how to interpret them.

- People who use personal messaging most frequently are **less likely to completely misperceive the tags’** purpose. However, rather than associating the tags with potentially untrustworthy content, **frequent messaging users tend to associate the tags with popular content, jokes, and multimedia**.
- Those who often participate in larger messaging groups, either of friends or of workmates, are **more likely to misperceive the tags’ purpose**.

We reiterate our recommendations in our earlier report. Meta should put the safety of the public first and reconsider the “forwarded” tags’ design and its strategy for informing people about how to interpret them.

## Five Principles for the Design of Effective Misinformation Warnings

1. **Don’t rely on description alone:** misinformation warnings should clearly indicate the potential for misinformation.
2. **Introduce user friction:** misinformation warnings may be overlooked unless they incorporate designs that force the user to stop and reflect.
3. **Gain media exposure:** platforms should engage in publicity campaigns about the intended purpose of misinformation warnings.
4. **Consider the context:** it is crucial to understand the different ways messaging platforms are used, shaped by social norms and people’s relationships with others.
5. **Think beyond platforms:** technological features need to be combined with socially-oriented anti-misinformation interventions, to empower people to work together to use personal messaging platforms in ways that help reduce misinformation.

# The Findings in Detail

In our survey we asked a representative sample of the UK public (2,000 people) what they thought the “forwarded” and “forwarded many times” tags mean.

Our aim was to capture the perceptions of the broad group of personal messaging users in the UK. So we first excluded from the analysis 192 people who stated that they never or only infrequently (less than once a month) use personal messaging.

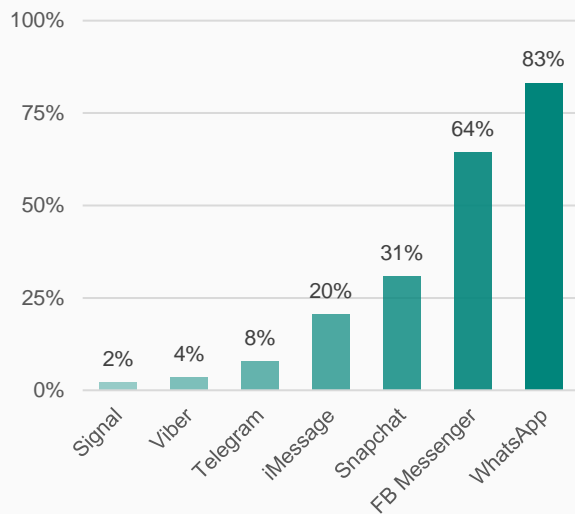
The survey question we asked was:

Typically, how often do you use a personal messaging service or app? Personal messaging services or apps include WhatsApp, Facebook Messenger, Snapchat, Telegram, iMessage (the default messaging app on iPhones and iPads), Messages (the default messaging app on Android phones and tablets), and similar services or apps.

This left 1,808 respondents—90% of the sample—that use personal messaging either one to three times a month, once a week, a few times a week, every day, or more than once per day.

Although not all personal messaging users are on WhatsApp (where the “forwarded” and “forwarded many times” tags are found) the vast majority are. According to the UK’s Ofcom, WhatsApp is the most popular social networking platform in the UK. 79% of UK adult internet users use it, which is 83% of those who use any messaging app. See Figure 1. The number of WhatsApp users continues to grow, meaning some of our respondents who do not use WhatsApp now are likely to do so in the future.

Figure 1. Proportion of UK Adult Messaging Users Who Use the Most Popular Messaging Apps



Ofcom Adults’ Media Use and Attitudes 2023 data. Question: “Which if any of these apps or sites do you use to send messages, chat or make video or voice calls?” Multiple responses permitted. Percentages calculated as proportion of those who selected at least one of 18 messaging app options. N=5,336.

We presented 19 options, listed in Table 2. These were based directly on the findings of our long-term qualitative fieldwork, which we reported in our [June 2023 report](#). Participants could choose multiple options, unless they answered that they had never heard of these labels or that they did not know.

## Categorising Responses

We then categorised the various options these 1,808 respondents selected. For this, we developed a simple *traffic-light-plus* system: Red, Amber, Green, and Grey. See Table 1.

Table 1. How Responses Relate to Aims to Tackle Misinformation

Red	Completely contradicts aims
Amber	Mostly irrelevant to aims
Green	Aligned with aims
Grey	Uncertainty or no awareness

Now see Table 2, which shows how this system worked and a snapshot of the overall responses.

**Table 2. Overall Responses**

Response option	No. who selected	% of sample
Content that is currently popular on WhatsApp	424	23.45
Content that is currently popular on social media	396	21.90
Jokes or satirical content	360	19.91
Content that is currently a big topic in the news	287	15.87
Pictures or GIFs	252	13.94
Untrustworthy content	193	10.67
Content that is likely to be irrelevant to me	187	10.34
Low quality content	183	10.12
Useful information	190	10.51
Video or audio	188	10.40
Content that is false or misleading	167	9.24
Content that is likely to be relevant to me	164	9.07
Links to other websites	159	8.79
Important information	144	7.96
Trustworthy content	120	6.64
Reliable information	108	5.97
High quality content	97	5.37
I have never seen or heard of these labels	299	16.54
Don't know	198	10.95

*Question: "You may have seen or heard about a label that can appear on WhatsApp messages that says 'Forwarded' or 'Forwarded many times.' In your opinion, what do these labels indicate that the message potentially contains? Tick all that apply." N=4,116 selections by 1,808 respondents who use personal messaging "One to three times a month," "Once a week," "A few times a week," "Every day," "More than once per day." Respondents could select only one of the two possible Grey responses. Selecting one Grey response meant a respondent could not select any other response. Response options were not presented in colours in the questionnaire. We randomised the presentation order of the list of options for each respondent, apart from "I have never seen or heard of these labels" and "Don't know," which were always fixed to the bottom of the list.*

Perceptions of the tags we labelled Red completely contradict Meta's stated aims to use the tags to reduce misinformation on WhatsApp, and are therefore the most worrying.

Perceptions we labelled Green clearly align with Meta's aim of reducing misinformation with the tags.

The perceptions we labelled Amber are mostly irrelevant to anti-misinformation aims because they have only weak links with a message's veracity. They include the perception that the tags mean a message is "popular" on WhatsApp or on broader social media, or indicate that the message contains a joke, video, or image. Although these perceptions are not categorically incorrect, they are troubling, first, because they fail to prompt reflection on a message's accuracy, and second, because people often make positive judgments about the reliability of "popular" content. Due to what communication researchers call the "bandwagon heuristic," people often associate popularity with credibility online.<sup>5</sup>

In the Grey category are the responses "I don't know" and "I have never seen or heard of these labels." These perceptions are substantively important here because they suggest the labels are unlikely to achieve their goals among these users.

## Overall Pattern of Responses

Our 1,808 messaging-using members of the UK public made an average of 2.28 selections each, and there were 4,116 selections in total.

As Table 2 on the left reveals, no single perception dominated, confirming our June 2023 report's finding that people have an extremely wide variety of interpretations of the functions of the "forwarded" and "forwarded many times" tags.



Troublingly, perceptions we labelled Green—ones that indicate that a person “gets” the point of the tags—in each case tended to be individually selected by only about a tenth of messaging users.

The tags do not perform well as misinformation warnings.

The most widely-held perceptions are those in the Amber category—ones mostly irrelevant to anti-misinformation aims. The most common here are that the tags highlight content currently popular on WhatsApp or social media more broadly. These are closely followed by “jokes or satirical content.” The high numbers among the UK public who selected these misdirected perceptions reveal that the tags do not perform well as misinformation warnings.

The responses in the Red category are especially worrying. These include perceptions that the tags show that a message contains useful, relevant, important, or even trustworthy information.

Although these Red responses were individually selected by smaller numbers of respondents than selected the Green responses, the numbers for Red came worryingly close to the numbers for Green.

A substantial minority of people misinterpret the tags in ways that potentially leave themselves vulnerable to misinformation.

These findings show that surprisingly few messaging users in the UK understand the tags straightforwardly as markers of potentially untrustworthy content.

And a substantial minority misinterpret the tags in ways that potentially leave themselves vulnerable to misinformation.

Lastly, a large group fell into the Grey category—lack of knowledge or awareness. We do not have

data on precisely how many of those who chose these responses are messaging users who use WhatsApp. But the number in the Grey category is surprisingly large if we bear in mind that Ofcom data show that 83% of messaging users in the UK use WhatsApp.

That close to a third (27.5%) of UK messaging users were either entirely unaware of the tags or did not know how to interpret them at all reveals there are currently big limits to the tags’ effectiveness.

## From Categorising Responses to Categorising People

Our survey question allowed for multiple selections: those who did not choose one of the Grey options could tick any number of options that they felt applied.

The advantage of this approach is that it allows people to quickly express complex views, while allowing us to identify patterns in the overall distribution of selections.

However, to better understand how *individuals* perceive the tags requires a different approach. We need a system for categorising each person based on the combination of different options they selected.

So, for this study we devised the following method.

Let’s use the Red options as an example. We placed a person in the Red category if they met either of two criteria:

- 1) They selected only Red options from the list, or
- 2) The number of Red options they selected was significantly greater—at least two more—than the number of selections they made from either the Amber or Green options.

We applied the same principle to assign people to the Amber and Green categories. See Table 3.

For the Grey category, our survey did not allow respondents to select other options in addition to a Grey option (“Don’t know” or “I have never seen or heard of these labels”), which made categorising these people straightforward.

Lastly, we placed all those who did not meet the thresholds for being assigned a colour—Red, Amber, Green, or Grey—into a separate “Mixed” category.

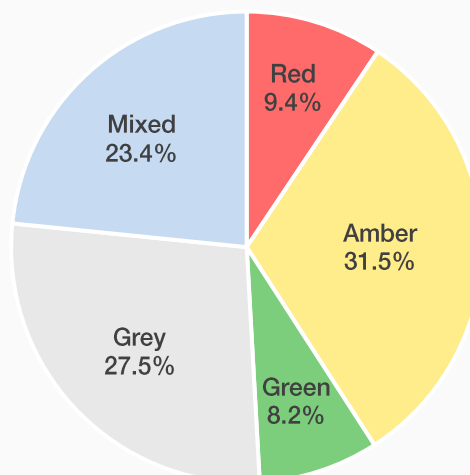
This is a simple and effective way of summarising the distribution of selections for each individual. It also represents a cautious approach. Individuals are not placed in one of the main categories of Red, Amber, or Green without crossing a demanding threshold.

**Table 3. Method for Categorising Individuals Based on Perceptions of the Tags’ Function in Reducing Misinformation**

<b>Red</b> Perceptions contrary to tags’ function	<ul style="list-style-type: none"> <li>• Respondents who selected only Red options.</li> <li>• Respondents who selected at least two more Red options than either Amber or Green options.</li> </ul>
<b>Amber</b> Perceptions mostly irrelevant to tags’ function	<ul style="list-style-type: none"> <li>• Respondents who selected only Amber options.</li> <li>• Respondents who selected at least two more Amber options than either Red or Green options.</li> </ul>
<b>Green</b> Perceptions aligned with tags’ function	<ul style="list-style-type: none"> <li>• Respondents who selected only Green options.</li> <li>• Respondents who selected at least two more Green options than either Red or Amber options.</li> </ul>
<b>Grey</b> Uncertain, unaware	<ul style="list-style-type: none"> <li>• Selected a Grey category response</li> </ul>
<b>Mixed</b>	<ul style="list-style-type: none"> <li>• No clear dominating category</li> </ul>

Figure 2 shows how many people there were in each category.

**Figure 2. Individuals by Category**



*Question: “You may have seen or heard about a label that can appear on WhatsApp messages that says ‘Forwarded’ or ‘Forwarded many times.’ In your opinion, what do these labels indicate that the message potentially contains? Tick all that apply.”*

*N=1,808 respondents who use personal messaging “One to three times a month,” “Once a week,” “A few times a week,” “Every day,” “More than once per day.” For the method we used to categorise individuals see Table 3.*

## What Categorising People Reveals

We can now see that about half of people in the UK who use messaging are unclear about the function of the “forwarded” and “forwarded many times” tags. They either fall into the Grey category or hold a mixture of inconsistent, jumbled perceptions (the Mixed category).

About half of people in the UK who use messaging are unclear about the meaning of the “forwarded” tags. Only 8.2% consistently perceive the tags in ways that align with the tags’ intended role as misinformation warnings.

This finding about lack of clarity is reinforced if we consider that the largest group (31.5%) were

Amber—those who hold perceptions of the tags that are mostly irrelevant to the tags’ role as misinformation warnings.

Only 8.2% are Green—they perceive the tags in ways that align with their intended role as misinformation warnings.

And almost 1 in 10 UK messaging users (9.4%) are in the Red category. This group holds fundamentally inaccurate perceptions that are diametrically opposed to what the forwarded tags are designed to elicit.

## Exploring Links Between Social Factors and (Mis)perceiving the Tags’ Function

We included some other questions in our survey. This enabled us to explore links between some social factors and people’s perceptions of the “forwarded” and “forwarded many times” tags.<sup>6</sup>

Let’s start with age.

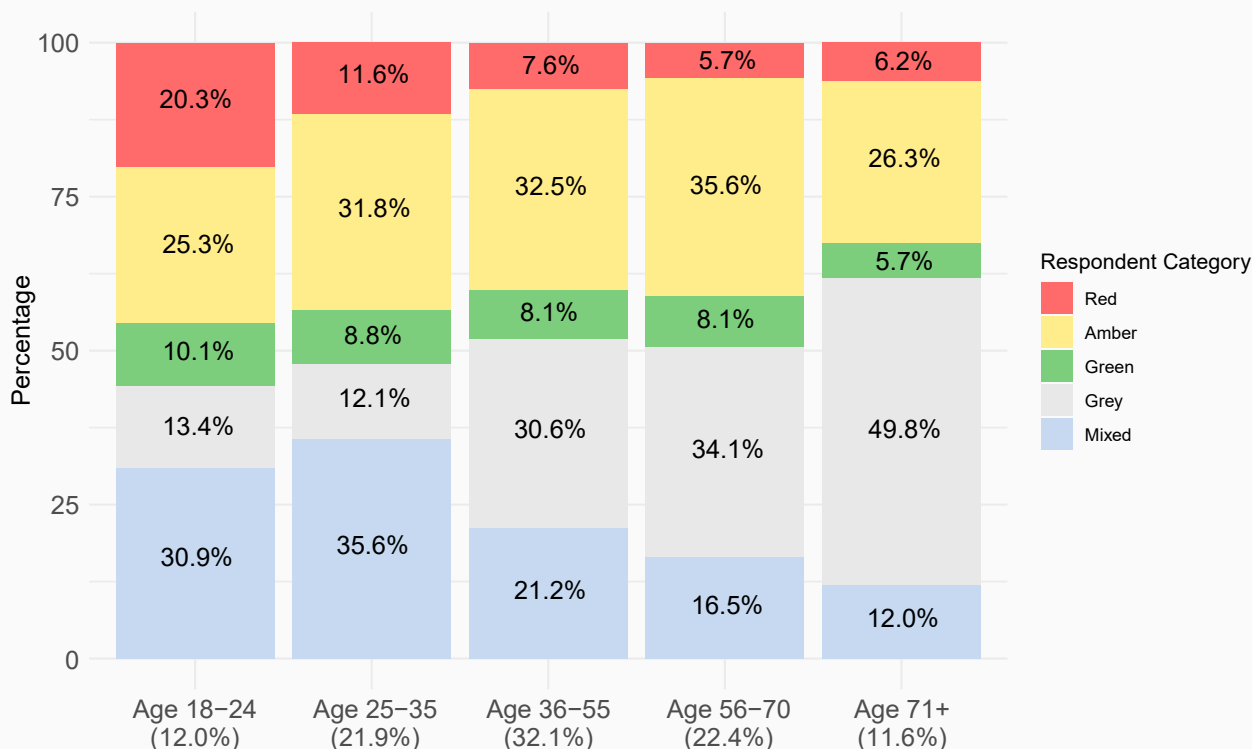
### Age

The relationship between age and interpretations of the tags is illustrated in Figure 3. Older messaging users are more likely to be unaware of the tags or not know how to interpret them at all (Grey).

However, the younger age groups, particularly those under 35, contain the largest proportions of people who completely misperceive the tags’ function (Red). This is most evident with the youngest group we surveyed (18–24 years).

That younger people are more confident in stating what they think the tags mean is perhaps unsurprising. But there is a trade-off here: this confidence can be misplaced. This finding belies the often-stated view that young people are “digital natives” and best placed to navigate online misinformation.

Figure 3. Age and (Mis)perceiving the Tags’ Function



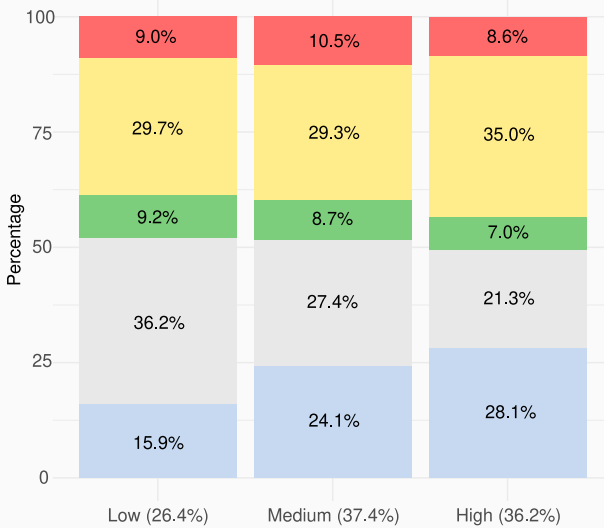
Percentages labelled on the horizontal axis are the proportion of the 1,808 messaging-using survey participants.

## Educational Attainment

When people’s interpretations of the “forwarded” and “forwarded many times” tags are considered in light of their formal educational attainment (Figure 4) we see a fairly clear pattern for people in the Grey category: the group with the lowest educational attainment contains the largest proportion of people who appear to be unaware of the tags or don’t know how to interpret them.

However, overall there is no clear evidence of a link between educational attainment and holding *inaccurate* perceptions of the tags’ purpose. And in the group with the highest level of education, there are still plenty who hold inconsistent or misdirected perceptions—see the Blue and Amber bars on the right-hand side of Figure 4.

Figure 4. Educational Attainment and (Mis)perceiving the Tags’ Function



*Low:* GCSE, Standard Grades or equivalent; or no formal qualifications.

*Medium:* A Level, Highers or equivalent; Certificate of Higher Education or equivalent; or Diploma of Higher Education or equivalent.

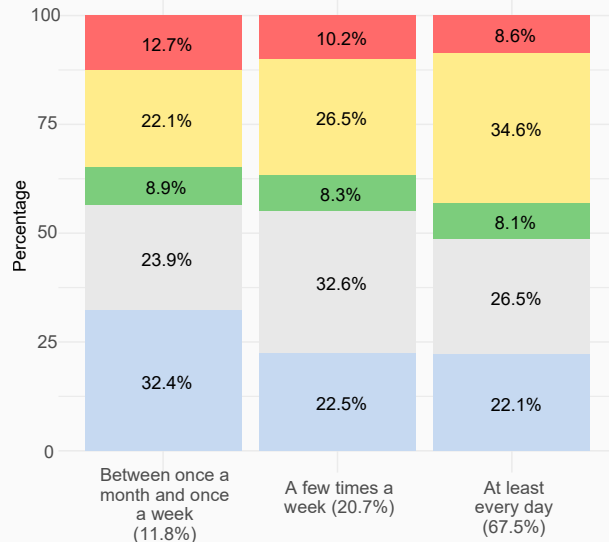
*High:* Undergraduate degree or above.

Percentages labelled on the horizontal axis are the proportion of the 1,808 messaging-using survey participants.

## Levels of Personal Messaging Use

Frequent users—the group who use personal messaging at least once a day or more often—are slightly less likely than infrequent users to misperceive the “forwarded” tags’ functions.

Figure 5. Frequency of Personal Messaging Use, and (Mis)perceiving the Tags’ Function



Question: “Typically, how often do you use a personal messaging service or app? Personal messaging services or apps include WhatsApp, Facebook Messenger, Snapchat, Telegram, iMessage (the default messaging app on iPhones and iPads), Messages (the default messaging app on Android phones and tablets), and similar services or apps.” Options: “Never” (excluded from analysis), “Less often than once a month” (excluded from analysis), “One to three times a month,” “Once a week,” “A few times a week,” “Every day,” “More than once a day”. Percentages labelled on the horizontal axis are the proportion of the 1,808 messaging-using survey participants.

At the same time, however, there is no strong evidence of a link between frequency of messaging use and perceiving the tags’ functions correctly, i.e. as denoting untrustworthy content (Green).

In fact, the group of most frequent users of personal messaging contains the highest proportion of those who fall into the Amber category—perceptions of the forwarded tags that are mostly irrelevant to anti-misinformation aims. Recall that the most frequently-selected

responses in this category are evaluations of the online “popularity” of the content, followed by what type of content the message contains (e.g., jokes, pictures, videos).

It may be the case that those who use personal messaging most often are more likely to be exposed to “viral” visual content and more often see the “forwarded” tags applied to it. But this, on its own, does not appear to make such individuals more attentive to the tags’ function as misinformation warnings.

## Trust in Information on Personal Messaging

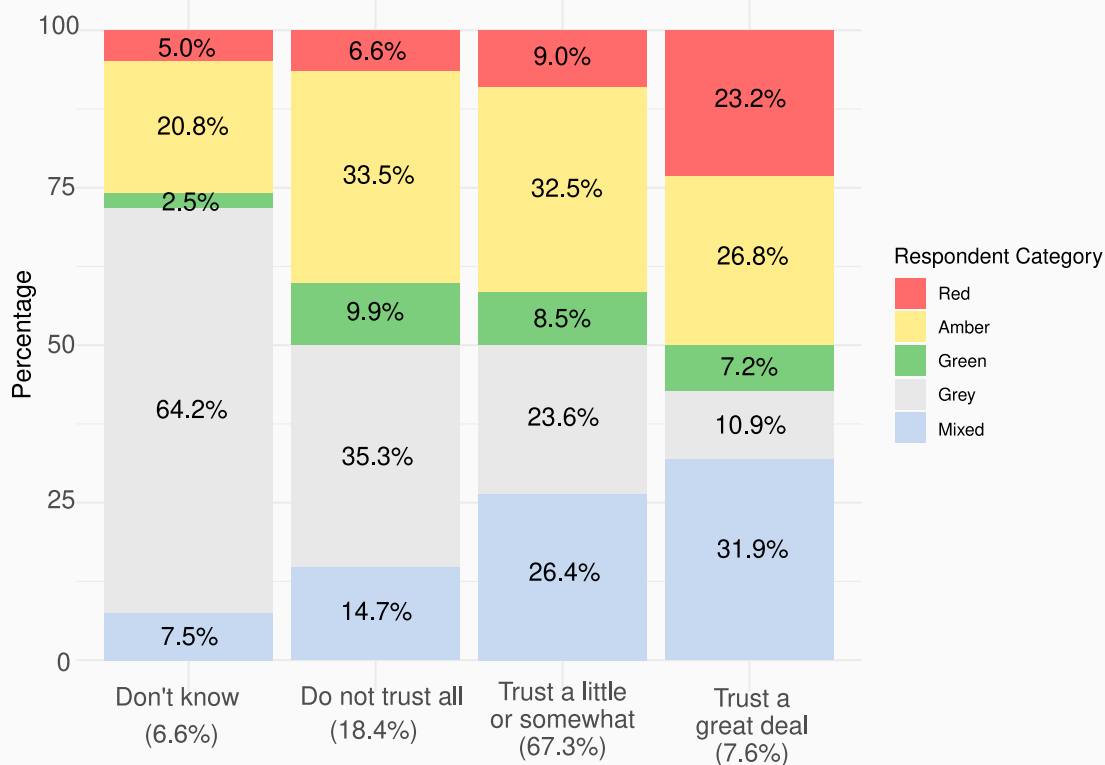
We also explored links between people’s interpretations of the tags and their general levels

of trust in the information and news they see on personal messaging. See Figure 6.

A clear pattern here is that the group with the greatest trust in content on these platforms contains the largest percentage of people who believe the “forwarded” tags denote accurate and trustworthy information and therefore directly misperceive the tags’ function.

It is worth noting here that across the messaging-using UK public, we found a significant majority—62%—said they had a high level of confidence in their own ability to judge the accuracy of information on personal messaging.

Figure 6. Trust in Information and News on Personal Messaging, and (Mis)perceiving the Tags’ Function



Question: “How much do you trust the information and news you see on personal messaging services or apps?” Options: “A great deal,” “Somewhat,” “A little,” “Not at all,” “Don’t know.” Percentages labelled on the horizontal axis are the proportion of the 1,808 messaging-using survey participants.

## Chat Settings and Group Size

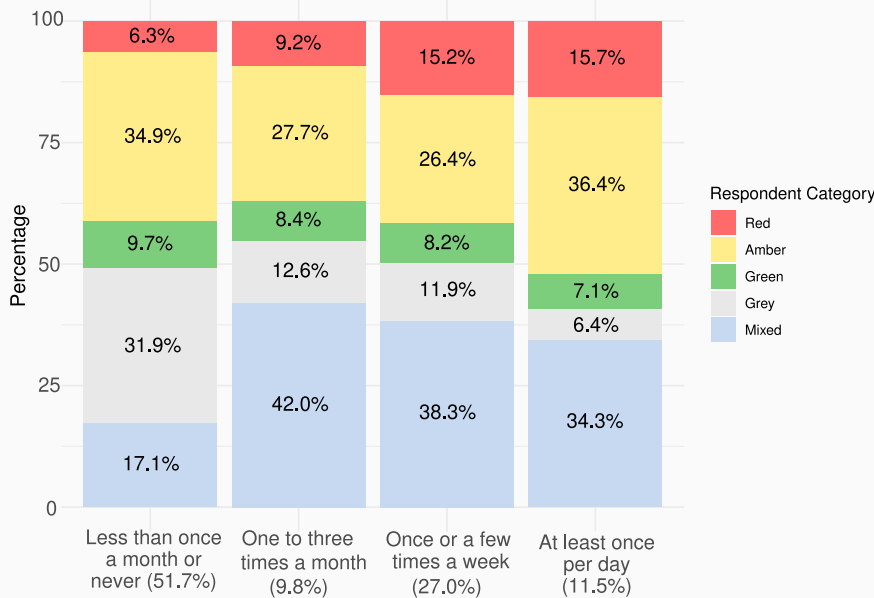
Finally, we considered whether participating in different types of chats and group settings on personal messaging has links with how the “forwarded” tags are perceived. We asked about chats and groups involving family, friends, workmates, neighbours, and hobbies/interests. For each setting we asked each survey participant how often they engaged in one-to-one chats, small groups (less than 5 people), and large groups (more than 10 people). See Figures 7 and 8.

Two settings—workplace and friend groups—appear to have links with an increased likelihood of misperceiving the tags’ purpose. But the key here appears to be the role of group size. Those who often participate in large groups, either of friends or of workmates, are more likely to misperceive the tags’ function.

There is no simple explanation for this finding. It might be that people who often participate in large groups are less vigilant because it is easier to delegate that responsibility to others. This is a key finding from a different strand of the Everyday Misinformation Project.<sup>7</sup>

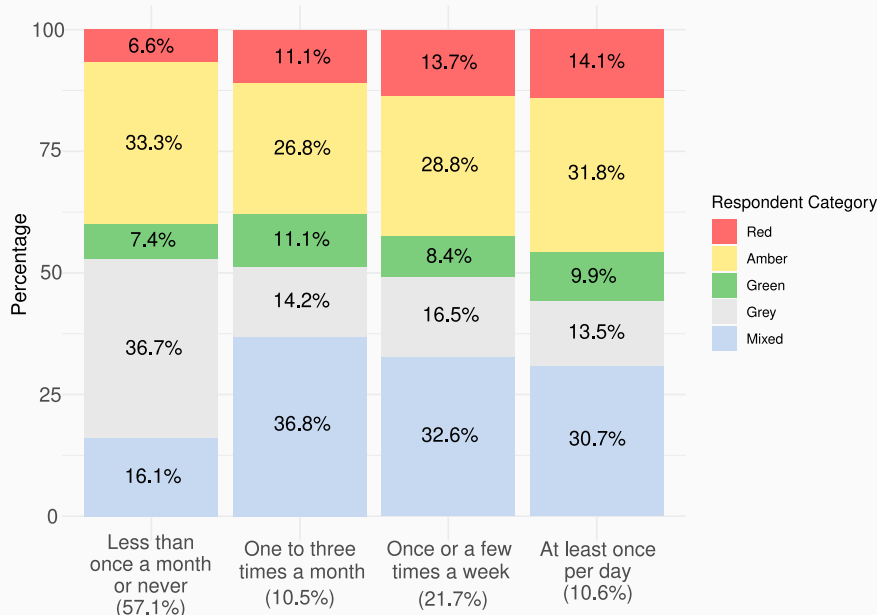
However, this finding might also be explained by how people behave in large messaging groups made up of friends or workmates. People who often spend time in large groups are more likely to see multiple “forwarded” messages. If the group is made up of people whom they trust (friends) or with whom they often exchange important information (workmates), they may more readily associate the messages tagged as “forwarded” with useful, reliable, and important information. Further research is needed to unpack this, but the key point here is that these are still misperceptions that go against what Meta said it wanted to achieve with the “forwarded” and “forwarded many times” tags.

Figure 7. Frequency of Participating in a Large Workplace Group, and (Mis)perceiving the Tags’ Function



Question: “We would now like you to think about the different kinds of chats and groups you have on personal messaging services or apps. How often do you see messages from...” “Someone from your workplace, in a one-to-one chat between just the two of you?”, “Someone from your workplace, in a small work related group of up to 5 people?”, “Someone from your workplace, in a large work related group of more than 10 people?” Options for each: “Never,” “Less often than once a month,” “One to three times a month,” “Once a week,” “A few times a week,” “Every day,” “More than once per day.” Participants who indicated that they were not in work were not asked this question, so N=1,218. Percentages labelled on the horizontal axis are the proportion of the 1,218 messaging-using, in-work survey participants.

Figure 8. Frequency of Participating in a Large Group of Friends, and (Mis)perceiving the Tags' Function



Question: “We would now like you to think about the different kinds of chats and groups you have on personal messaging services or apps. How often do you see messages from...”, “A friend, in a one-to-one chat just between just the two of you?” “A friend, in a small friends’ group of up to 5 people?”, “A friend, in a large friends’ group of more than 10 people?” Options for each: “Never,” “Less often than once a month,” “One to three times a month,” “Once a week,” “A few times a week,” “Every day,” “More than once per day.” Percentages labelled on the horizontal axis are the proportion of the 1,808 messaging-using survey participants.

## Conclusions

The findings outlined above cast serious doubt on the effectiveness of WhatsApp’s strategy against misinformation.

WhatsApp’s “forwarded” and “forwarded many times” tags are a light-touch intervention for tackling misinformation. The tags are intended to prompt critical reflection on a message’s origin and whether its content is accurate and trustworthy. But the tags do not explicitly say this. So their effectiveness relies on people knowing that they should be associated with potential misinformation.

Our nationally-representative survey shows that members of the UK public who use personal messaging have widely varying interpretations of WhatsApp’s “forwarded” and “forwarded many times” tags. This broadly confirms the findings

from our June 2023 report based on in-depth interviews.

Although these tags are meant to prompt the recipient to stop and reflect on where a message originated and whether it is trustworthy, very few people understand the tags in the intended way. It is much more common for users to see the tags in ways that are not relevant to spotting misinformation, that is, as markers of viral or entertainment content.

There is also a small but significant group within the UK messaging-using population who see the tags as flagging accurate, trustworthy, useful, or relevant content. These misperceptions can potentially render people more vulnerable to misinformation.

Our additional analysis shows that those most likely to hold such direct misperceptions are the youngest messaging users, those who place greatest trust in content they see on messaging,

and those who often participate in larger messaging groups of friends or workmates.

Meta needs to work harder to increase awareness of the tags' intended purpose, particularly among younger people, who are confident they understand the tags yet more likely to get it wrong, and older people, who use the service in large numbers but are more likely to be unaware of the tags.

There is also work to do to increase awareness of the potential for deceptive content among messaging users, because those who place the greatest trust in the content they see on messaging are also more likely to misperceive the tags' purpose in dangerous ways.

Those who participate in larger messaging groups should also be the focus of Meta's efforts because they are more likely to misperceive the tags.

## Next Steps

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 for tackling the spread of misinformation.

In our previous report, we put forward five principles that personal messaging platforms should consider when designing such measures. Although we devised these based on our findings about WhatsApp's "forwarded" and "forwarded many times" tags, they can be applied to the design of warnings on any messaging platforms. The principles can be implemented without

compromising end-to-end encryption, which we anticipate and hope will continue as a feature of personal messaging.

## 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. That means they may not prompt 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 unintended and contradictory interpretations. Explicit warnings 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 warnings' effectiveness, particularly when the warning is plain and inconspicuous. In our previous report, we found some people come across such warnings but do not recall seeing them. Introducing friction in the user experience can help draw attention to misinformation risks. For example, 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 message’s trustworthiness 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 companies 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. Active efforts to raise awareness are particularly needed where warnings are vague and open to different interpretations, or in contexts where awareness of the role of personal messaging in misinformation may be low. 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 the variety of ways in which personal messaging platforms are used across social contexts is crucial to designing relevant and useful misinformation warnings. Differences in the way personal messaging is used by different groups mean different degrees of exposure to forwarded misinformation. And, for some groups, message characteristics other than being forwarded might actually be more salient markers of potential misinformation. Platforms need to consider the ways misinformation spreads in specific contexts, and whether a broader variety of anti-misinformation measures is needed.

## 5. Think Beyond the Platforms

Finally, it is also important that personal messaging platforms recognise the limitations of features like warnings for tackling misinformation. Misinformation is a complex social problem that cannot be wholly addressed through the introduction of new technical features alone. Relationships and social norms are crucial here, because personal messaging platforms are what we call “hybrid public-interpersonal communication environments.”<sup>8</sup> 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 in order to successfully reduce the spread of misinformation on personal messaging. These should focus on building social capacities. This means 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.

These principles are meant to serve as a foundation for further research. We encourage researchers to build on our findings to continue to investigate how misinformation warnings can be made most effective in practice.

# Principles for effective misinformation warnings

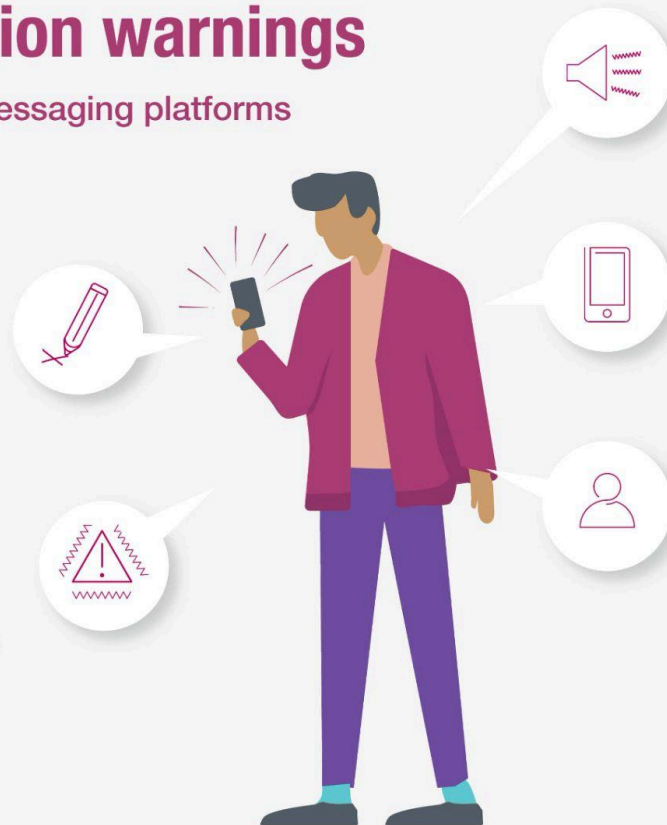
Advice for personal messaging platforms

## 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.

# Data and Research Method

Ethical approval for this study was granted by Loughborough University’s Ethics Review Subcommittee (2023-16044-15938; PI Chadwick).

We designed a survey and hired established opinion polling company Opinium Research to administer it to a nationally representative sample of the UK public. 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

People were eligible for the survey if they resided in the UK and had access to the internet (via any device) to complete it. Quotas matched to the latest UK Office of National Statistics data ensured the final sample was representative of the UK population in terms of age, gender, region, educational attainment, and ethnicity. The final sample size was 2,000.

## Comparing the Demographic Characteristics of our Sample with the UK Adult Population

	UK Office for National Statistics	Our Sample (N=2,000)	Difference
<i>Age &amp; Gender<sup>1</sup></i>			
Male 18-34	14%	14%	
Male 35-54	16%	15%	-1%
Male 55+	19%	18%	1%
Female 18-34	14%	14%	
Female 35-54	17%	17%	
Female 55+	21%	21%	
<i>Region</i>			
East Midlands	7%	8%	+1%
East of England	7%	8%	+1%
London	13%	13%	
Northern Ireland	4%	4%	
North East	4%	4%	
North West	11%	11%	
Scotland	10%	9%	-1%
South East	14%	13%	-1%
South West	9%	8%	-1%
Wales	5%	5%	
West Midlands	9%	9%	
Yorkshire & Humberside	8%	8%	
<i>Educational Attainment</i>			
Low	28%	28%	
Mid	37%	37%	
High	35%	35%	
<i>Ethnicity<sup>2</sup></i>			
White	82%	83%	+1%
Multiple ethnicity	3%	3%	
Asian	9%	7%	-2%
Black	4%	4%	
Not White/Multiple/Asian/Black	2%	1%	-1%

Notes: <sup>1</sup> ONS does not currently collect data for non-binary gender identity therefore matching a quota for this group is not possible. In our sample, 15 participants selected non-binary, 28 participants preferred not to state their gender, and 31 participants opted out of the question about gender. <sup>2</sup> Figures for Ethnicity exclude 37 participants who opted out of the question about ethnicity. Totals for Age & Gender and Ethnicity in our sample do not equal 100% due to rounding.

# About the Everyday Misinformation Project

Based in the Online Civic Culture Centre (O3C) at Loughborough University, the Everyday Misinformation Project is a three-year study funded by the Leverhulme Trust. Our aim is to develop a better, more socially-contextual understanding of why people share and correct misinformation online. We have 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 and around the world, 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 and 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 these 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 is a Postdoctoral Research Associate; Portia Akolgo is a Research Assistant. Dr Brendan Lawson was a Postdoctoral Research Associate 2021-22 and is now a Lecturer in Communication and Media at Loughborough University.

The fieldwork has three strands:

- Longitudinal in-depth qualitative interviews with 102 members of the public based in three regions of the UK.
- 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, designed based on findings from the first two strands of fieldwork.

This is the third public-facing report from the project. It presents findings based on the third strand of the fieldwork. Visit <https://everyday-mis.info> for more information.

# 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](mailto:n.hall@lboro.ac.uk)

[www.natalieannehall.com](http://www.natalieannehall.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](mailto:a.chadwicki@lboro.ac.uk)

[www.andrewchadwick.com](http://www.andrewchadwick.com)



## Cristian Vaccari

Chair of Future Governance, Public Policy and Technology, School of Social and Political Science, University of Edinburgh

[cvaccari@ed.ac.uk](mailto:cvaccari@ed.ac.uk)

[www.cristianvaccari.com](http://www.cristianvaccari.com)



## Brendan T Lawson

Lecturer in Communication and Media, Department of Communication and Media, and Researcher, Online Civic Culture Centre, Loughborough University.

[b.b.lawson@lboro.ac.uk](mailto:b.b.lawson@lboro.ac.uk)



## Portia Akolgo

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

[p.m.akolgo@lboro.ac.uk](mailto:p.m.akolgo@lboro.ac.uk)

# Disclosure and Integrity Statement

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

Andrew Chadwick and Cristian Vaccari are currently advisory board members (unpaid) of Clean Up The Internet. Any opinions in this report are solely those of its authors.

# Notes

1. Ofcom (2023).
2. Chadwick, Vaccari & Hall (2023).
3. Avelar (2019); Kazemi et al. (2022); Martínez (2018); Molteni (2018); Sahoo (2022); Vasudeva & Barkdull (2020).
4. Vasudeva & Barkdull (2020).
5. Sundar (2008).
6. For ease of interpretation, throughout this report we only discuss simple links between two variables. A caveat applies to this type of analysis. Take, for example, the relationship between age and perceptions of the tags. Older people may be more likely to be unaware of the “forwarded” tags but this does not automatically mean that being older is the most important variable associated with being unaware of the tags. Statistical analyses that help disentangle the relative importance of multiple different variables, known as multivariate analyses, are more difficult to interpret for those unfamiliar with statistics, which is why we avoid them in this report. Throughout, we are careful to present justifiable summaries of the different variables that matter in each case.
7. Chadwick, Hall & Vaccari (2023).
8. Chadwick, Vaccari & Hall (2022); Chadwick, Vaccari & Hall (2023); Chadwick, Hall & Vaccari (2023); Hall, Chadwick & Vaccari (2023).

# 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>.
- 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*, 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>.
- Hall, N-A., Chadwick, A. and Vaccari, C. (2023). Online Misinformation and Everyday Ontological Narratives of Social Distinction. *Media, Culture & Society*, OnlineFirst, <https://doi.org/10.1177/01634437231211678>.
- Hall, N-A., Lawson, B.L., Vaccari, C. and Chadwick, A. (2023). Beyond Quick Fixes: How users make sense of misinformation warnings on personal messaging. Online Civic Culture Centre. [https://www.lboro.ac.uk/media/media/research/o3c/pdf/O3C\\_4\\_Beyond%20Quick\\_Fixes\\_Misinformation\\_Warnings\\_Personal\\_Messaging.pdf](https://www.lboro.ac.uk/media/media/research/o3c/pdf/O3C_4_Beyond%20Quick_Fixes_Misinformation_Warnings_Personal_Messaging.pdf)
- 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>.
- 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>.
- 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/>.
- Ofcom (2023). Adults' Media Use and Attitudes 2023. <https://www.ofcom.org.uk/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes>
- 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>.
- 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.
- 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>.



# 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) analyses 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, intolerance, and trust online. In addition to publishing high-quality interdisciplinary social science research, O3C develops evidence-based knowledge to inform policies and practices that mitigate the democratically dysfunctional aspects of social media. For more information, visit the [O3C website](#).

**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](http://lboro.ac.uk/research/online-civic-culture-centre)**

---