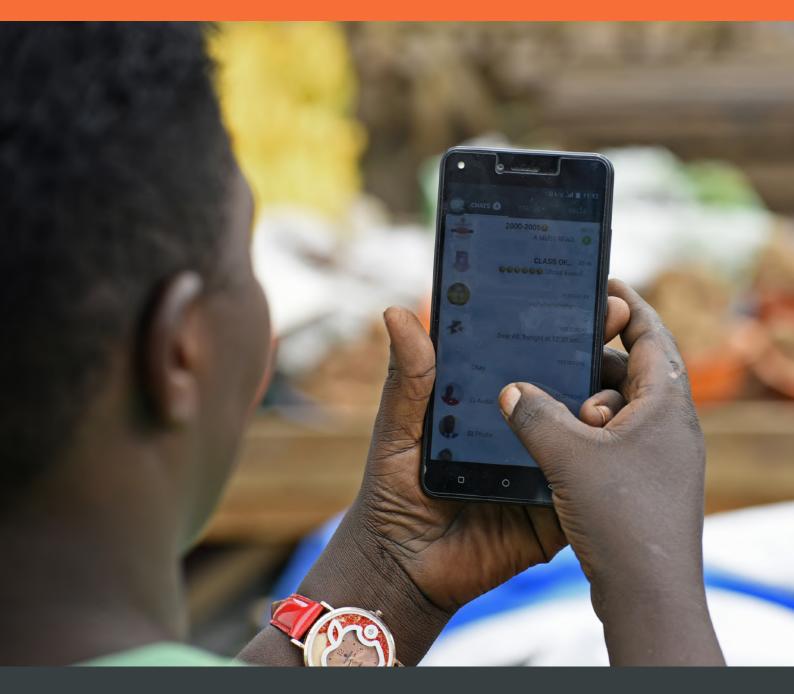# Tackling misinformation on WhatsApp in Kenya, Nigeria, Senegal & South Africa: Effective strategies in a time of Covid-19

December 2020

Published by Africa Check and the Africa Centre for Evidence.

Researchers

Nicola Theunissen
Africa Check
Johannesburg, South Africa

Linda Etale
Africa Centre for Evidence
Nairobi, Kenya

Andile Madonsela
Africa Centre for Evidence
Johannesburg, South Africa

Irvine Manyukisa
Africa Centre for Evidence
Johannesburg, South Africa

Ruth Stewart
Africa Centre for Evidence
Johannesburg, South Africa

Nkululeko Tshabalala
Africa Centre for Evidence
Johannesburg, South Africa

About Africa Check

Africa Check is the continent's leading independent, non-partisan, fact-checking organisation. Established in 2012, Africa Check has four offices across the continent: in Johannesburg (South Africa), Nairobi (Kenya), Lagos (Nigeria) and Dakar (Senegal). Its goal is to raise the quality of information available to society across the continent.
University of the Witwatersrand, South Africa

www.africacheck.org   *@AfricaCheck*

About the Africa Centre for Evidence

The Africa Centre for Evidence (ACE) is a research organisation based at the University of Johannesburg that was founded at the end of 2016. ACE's mission is to contribute to the reduction of poverty and inequality in Africa through the use of evidence. All of ACE's work is aligned with the Sustainable Development Goals and the National Development Plan of South Africa. The ACE team is multi-disciplinary and provides the secretariat to the Africa Evidence Network.
University of Johannesburg, South Africa

www.africacentreforevidence.org   *@ACE_UJ*

KONRAD ADENAUER STIFTUNG

Table of contents

Overview

Misinformation about Covid-19 threatens to exacerbate the impact of the pandemic, undermine efforts to tackle the disease, and permanently reduce our trust in the global health system.

Social media fuels the generation and spread of misinformation and, although there is widespread fact-checking of content on Twitter and Facebook, dark social-media platforms such as WhatsApp, where messages are encrypted, remain a challenge. Encryption implies that the messages and calls of the 2 billion WhatsApp users across the world are secured so that nobody other than the sender and receiver can read or listen to them, not even WhatsApp. Fact-checkers committed to tackling the Covid-19 infodemic face an uphill task in countering misinformation on WhatsApp.

As the most commonly used form of social media in Africa, WhatsApp is accessed daily by millions of Africans. Fact checkers across the continent urgently seek evidence to shape effective fact-checking strategies to help counter the impact of Covid-19 in Africa and further afield. This research aims to provide that evidence. It addresses the following questions:

1. What is the nature of misinformation on Covid-19 shared across WhatsApp in Africa and what level of risk does this misinformation present?

2. How are WhatsApp users engaging with health misinformation, and how are they responding in ways that reduce the risk it presents?

3. What do we know about how to mitigate against misinformation on social media?

To address these questions, we combined rapid evidence assessments with a survey of WhatsApp users, interviews with fact-checkers, and an analysis of misinformation messages reported to Africa Check since the start of the pandemic.

The rapid evidence assessment and the claims reported to Africa Check suggest that Covid-19 misinformation poses significant risk to society. Both methods highlight a large number and a wide range of risks and harm to physical health, and economic, social and political well-being. The evidence also suggests that the overarching psychological impact of the infodemic might be contributing to an impending Covid-19 mental-health crisis.

We find, through the analysis of WhatsApp claims, that health misinformation on WhatsApp as a private messenger app presents specific risks because of access, format and tone.

Our interviews with fact-checkers shine additional light on the scale of the issue as they highlight the wide range of sources of misinformation and the involvement of influential individuals, including religious leaders.

Our rapid evidence assessment of the ways in which social media users respond to misinformation identify eight user behaviour patterns that lead to the spread of misinformation.

The evidence suggests that although some users can and do assess the validity of information by checking for cues within the messages, trying to verify the source of the information and occasionally reporting false messages or posting a correction, most of the time social media users appear to delete messages, ignore them or just share them anyway. Some users worryingly acted on misinformation, as a study in Nigeria about the Ebola virus demonstrates.

When we explore the evidence base to understand what motivates users to respond to misinformation in particular ways, we find that the type of content, who had shared it with them, the emotions it triggered, their trust of social media (or their fear of it, *'cyberchondria'*), and their tendency towards conformity all shape their behaviour.

We discover a strong theme of responsibility to those within their social circle and a need to be of help. Trust in social media also plays a role.

Our survey of WhatsApp users in Africa suggests that people do share Covid-19 messages widely in their networks. They do also question their accuracy and report misinformation, by asking the sender about it. Less frequently, they will report the message to an official fact-checking organisation such as Africa Check.

When asked what motivated users to share messages with others, respondents referred to a desire to raise awareness about the pandemic and provide helpful information to those they care about. This sense of responsibility may be linked to the finding from our rapid review that social media users are strongly influenced by their own social circles and a desire to help. Users also told us that they acted on information that they felt would improve their health and those they care about.

Our survey findings further support the review finding that users' responses are shaped by whom they receive messages from. Of particular interest is that survey respondents rated their trust in messages from legitimate news sources to be higher than their trust in government sources (for example, the Ministry of Health), organisations they know, or people they respect.

Our rapid evidence assessment identified nine mitigating strategies to counter health misinformation on social media: credible information over misinformation, self-efficacy to detect misinformation, making misinformation illegal, infoveillance, technical solutions, debunking, social media companies tackling misinformation in their platforms, collective action against misinformation and social media campaigns.

The research was enhanced with contributions from our survey respondents and via our interviews with African fact-checkers. The three strategies common to all three parts of our research are those that relate to:

• Self-efficacy to detect misinformation – users want and need to be enabled to detect misinformation

• Verifying or debunking information via reliable organisations or groups that can assess the validity of information, and

• Public awareness campaigns about misinformation

The evidence base on the effectiveness of these mitigating strategies, if any, is thin, mostly conducted in the USA, without a focus on WhatsApp specifically and as such should be translated to African contexts with caution.

Taking our findings into account, we conclude with five evidence-based strategies for fact-checkers to consider in the fight against health misinformation:

- Use risk classification to enable more deliberate editorial and fundraising strategies

- Develop proactive key messages and positive reinforcement around information consumers need to be helpful in a time of crisis

- Leverage users' social circles to champion evidence-based health information about Covid-19

- Extend the fact-checking 'Circle of Trust' by building partnerships with trustworthy media, government bodies, civil society partners, religious leaders and big tech companies

- Promote self-efficacy and empower individuals to take control of misinformation through media literacy and social media campaigns

Background

If the 2020 Word of the Year is not 'COVID-19', it might well be 'Infodemic'. The waves of health misinformation that flooded social media and messaging apps at the onset of the coronavirus pandemic had UN Secretary General António Guterres warning the world against "harmful health advice", proliferating "snake-oil solutions" and "wild conspiracy theories" that sow hate and stigmatisation (NPR, 2020), the so-called infodemic (UN Department of Global Communications, 2020). The analogy of misinformation in epidemiological terms is not new at all. For years, people have been using the phrase "going viral" to refer to the rampant spread of digital information.

In fact, Eysenbach (2020) traces the term *infodemiology* back to 1996, defining it as "a knowledge translation gap between best evidence (what some experts know) and practice (what most people do or believe)". It is a problem with potentially fatal consequences that have plagued the medical profession for decades, particularly during epidemics.

On risks

Never in the history of digital communication has an infodemic received such wide-spread attention from medical, media and communications researchers and practitioners as in 2020, especially as the harm caused by the infodemic becomes undeniable. In Iran, 728 people died of alcohol poisoning between February and April 2020 (Al Jazeera, 2020); another 90 people have lost their eyesight or were suffering eye damage. In a letter to the editor of the *American Society of Tropical Medicine and Hygiene*, the authors refer to two men who respectively ingested 15 mℓ of surface disinfectant, and 100 mℓ of alcohol-based hand sanitiser based on social media advice to protect them from Covid-19 (Siddiqui *et al.*, 2020). The authors reiterate how the nonevidence-based medical advice about Covid-19 *"adds to the strain of the pandemic on medical and psychological healthcare resources"*. According to Khuroo (2020), the use of hydroxychloroquine as a desperate attempt for prophylaxis and treatment of Covid-19 has occurred despite a lack of evidence and the drug's potential cardiotoxicity.

To understand how to tackle misinformation in Africa, we need to understand the risks that it poses. By scoping out the problem, we can identify better solutions.

On user behaviour

While we know that some people will be directly impacted by misinformation, such as in the poisoning examples above, we also know that not everyone will believe or act on the information they receive. So how can we minimise the harm these messages cause for those who do believe them? To answer this question, we first need to understand how users respond to the messages they receive.

Evidence shows that, although some people respond positively to potential misinformation, by verifying it, posting a correction if it is false or reporting it, most social media users display passive behaviour, such as ignoring or deleting it, or negative behaviour such as

sharing it, or acting on it. When it comes to user behaviour, we need to recognise that those in Africa may not respond the same as those in the Global North or elsewhere, and of course that those across Africa may not respond the same as one another.

African WhatsApp users from our survey have confirmed that they frequently shared Covid-19 information on WhatsApp, either to individual contacts or to one or more WhatsApp groups. They also displayed passive behaviours like deleting messages or not doing anything. Lower percentiles demonstrated positive behaviour like asking the sender about the accuracy of the message or reporting it to a misinformation line. Respondents were motivated by a desire to raise awareness about the pandemic and provide helpful information to those they care about. Social media users are strongly influenced by their social circles, which seems to be specifically relevant to WhatsApp, where the moral obligation to share helpful information with family and friends is strong.

On mitigating strategies

Understanding the nature and danger of misinformation about Covid-19 and understanding user responses to misinformation helps us to understand the problem and points towards possible solutions. Fact-checking organisations and public health bodies around the world are urgently working to find and effectively implement mitigating strategies. The literature identifies nine mitigating strategies to counter online health misinformation. Our research finds that self-efficacy to detect misinformation is a key strategy, as it speaks to users' agency in responding to misinformation. Another strategy is to provide credible, accurate information to users (the antidote or inoculation approach, by fighting misinformation with high dosages of accurate information), while verifying or debunking information via reliable organisations or groups that can assess the validity of information, and public awareness campaigns about misinformation are also common strategies.

Our research questions

This report presents findings to the following three research questions:

1. What is the nature of misinformation on Covid-19 shared across WhatsApp in Africa and what level of risk does this misinformation present?

2. How are WhatsApp users engaging with health misinformation, and how are they responding in ways that reduce the risk it presents?

3. What do we know about how to mitigate against misinformation on social media?

Methods

This project sets out to use a range of methods to address these questions, combining: systematic approaches to rapid evidence assessments, a trend analysis of misinformation shared on WhatsApp, a user behaviour survey conducted in French and English, a series of online workshops engaging key stakeholders across the fact-checking community, mainly the Africa Facts Network, and key informant interviews with specific fact-checkers in the Network. These are described in more detail in the annexures and summarised in Table 1 below.

Table 1: An overview of our study methods

Question	Methods
1. What is the nature of misinformation on Covid-19 shared across WhatsApp in Africa and what level of risk does this misinformation present?	Risk framework development A rapid evidence assessment Interviews with fact-checkers working in Africa Trend analysis of misinformation reported to Africa Check for three months from April to June 2020
2. How are users responding to public health misinformation on WhatsApp?	A rapid evidence assessment A survey of WhatsApp users across Africa
3. How can we better respond to the risks that misinformation presents to help tackle the pandemic in Africa?	A rapid evidence assessment A survey of WhatsApp users across Africa Interviews with fact-checkers working in Africa

Findings

Section 1: Risks and harm of public health misinformation

This section tackles three questions relating to the risks and harm of misinformation on WhatsApp, focusing particularly on Covid-19 in Africa, while drawing lessons from other forms of social media, and from research outside of the continent too.

1.1.   How can we categorise misinformation risks and harm?

1.2.   What can we learn about the level of risk that misinformation presents from our rapid evidence assessment?

1.3.   What risks and harm of misinformation did African fact-checkers share with us in our interviews?

1.4.   What is the nature (and associated risks) of Covid-19 misinformation shared across WhatsApp in Africa and reported by users for fact-checking?

To address these questions, we have:

• Developed a framework for categorising risks and harm drawing heavily on the wider literature

• Conducted a rapid evidence assessment of the risks and harm of public health misinformation on social media

• Interviewed fact-checkers working in Africa

• Analysed the content of misinformation reported to Africa Check between April and June 2020

Details of our methods are reported in the Annexures.

1.1   How can we categorise misinformation risks and harm?

Preliminary scans of the academic literature found no standardised framework to help communication and media professionals, specifically fact-checkers, understand and classify the risks associated with health misinformation in a more systematic way. One of the biggest challenges for fact-checkers is the high volumes of misinformation matched against the low volumes of staff. As a caveat, fact-checkers publicly state that they cannot fact-check every single false claim and therefore prioritise claims based on factors such as virality (how often information has been shared) and the potential harm that could be caused by a claim if left unchecked. Yet, with the lack of a standardised risk classification framework for health

misinformation, how do fact-checkers determine which claims are the riskiest during a pandemic?

In a bid to develop more effective strategies to fight health misinformation during the Covid-19 pandemic and beyond, it is important for fact-checkers to gain a better understanding of the risks associated with health misinformation.

Towards an evidence-based framework for analysing the risks of health misinformation during Covid-19

A 2020 report on the consequences of misinformation, published by UK fact-checking organisation Full Fact, states that "a great deal of anecdotal evidence exists" about the provable, plausible or potential harm of misinformation, but that the field significantly lacks rigorous, high-quality research.

The Full Fact report groups the harm of misinformation in four main categories:

• Disengagement from democracy

• Interference in democracy

• Economic harm

• Risks to life

A preliminary scan of the literature confirms the negative health consequences of misinformation, which overlaps with the Full Fact report's "risks to life" category. From the literature, it appears as if harm to health should be viewed on a spectrum, ranging from impact on mental health at the one end, to physical health implications on the other.

It becomes evident that Covid-19 misinformation had notable implications for people's mental health. For example, one survey-based study with 516 participants in Iraqi Kurdistan found that social media had a significant impact on spreading fear and panic related to the Covid-19 outbreak and an influence on people's mental health and psychological well-being (Ahmad & Murad, 2020).

Research also reports physical harm caused by Covid-19 misinformation. In one innovative study, researchers analysed 16 729 calls to the Regional Centre for Poison Control and Prevention in Massachusetts and Rhode Island, and 25 231 tweets discussing the treatment of Covid-19 with house cleaners. They found a startling geospatial correlation between Covid-19 health misinformation and poisoning with household cleaners in the Greater Boston Area. The poison control call centre witnessed a spike in calls about house-cleaner poisoning that were preceded two to three days earlier by tweets about ingesting bleach to cure Covid-19.

Overall findings into the impact of believing conspiracy theories largely focused on negative health consequences, including poor health-seeking behaviour, or psychological harm such as inducing fear and paranoia.

However, some studies also assessed the socio-political implications of conspiracies and found that the belief in conspiracy theories led to a disregard of government guidelines and distrust of authorities (Freeman *et al.*, 2020). The same study suggests that people

who believed conspiracies were also less willing to take diagnostic or antibody tests or be vaccinated (flagging poor health-seeking behaviour). Conspiracy ideas were more commonly associated with paranoia and a distrust in institutions (Freeman *et al.*, 2020). Georgiou and colleagues (2020) found that beliefs in conspiracy theories correlated strongly with negative attitudes towards government responses.

In yet another study on conspiracy beliefs, a survey with 1540 university students from the University of Jordan, found that students who held the belief that the disease is part of a global conspiracy demonstrated lower knowledge and higher anxiety about Covid-19 (Sallam *et al.*, 2020).

Budhwani and Sun (2020) find evidence into the social harm of Covid-19 misinformation. In this study, the authors assessed whether there was an increase in the prevalence and frequency of the phrases "Chinese virus" and "China virus" on Twitter after US President Donald Trump's use of the term on 16 March 2020. The study found that prior to the presidential usage, 16 535 "Chinese virus" or "China virus" tweets were identified. In the period following his public utterances, 177 327 tweets were identified, illustrating a nearly ten-fold increase at the national level. All 50 US states witnessed an increase in the number of tweets exclusively mentioning "Chinese virus" or "China virus" instead of coronavirus disease or Covid-19. The authors concluded that both the increase in the number of tweets and its content referencing "Chinese virus" or "China virus" point to the fact that Covid-19 stigma was perpetuated on Twitter. The media has reported widely on the issue, highlighted in examples of prejudice and assault as a New York Times article demonstrates.

Other research touches on the economic harm of misinformation in Africa (Ahinkorah *et al.*, 2020). They reference how local scammers and internet bloggers have created web links to spread Covid-19 misinformation with 'provocative' headlines to lure users into visiting, generating advertising revenue for the owners, or promoting unverified treatment protocols or medication (Ahinkorah *et al.*, 2020). The Full Fact report reiterates that economic harm from misinformation has been a particular problem in Africa, where social media pages impersonating recruiters or global charities scammed job seekers into paying cash before applying for a job. Ahinkorah and colleagues (2020) also reference 'panic purchases' that led to interruptions in the supply chain and increased demand-supply loopholes as an economic consequence of misinformation during the Covid-19 pandemic.

If we reassess the risks associated with Covid-19 misinformation on social media based on the four categories identified by Full Fact, we can classify the harm or risks from Covid-19 misinformation at the *sphere of impact* at which the harm occurred, being health, economic, social and political.

Typically, some claims' greatest harm would be around people's physical health (such as, someone dying from drinking household cleaner). These claims would fall within the Health sphere of impact. In other claims, the greatest harm is the direct loss of income through scams. In this case, the harm occurs within the economic sphere of impact. If a claim instigates discrimination against a person or group based on perceivable social characteristics, which in return fosters social harm like systemic racism, polarisation, xenophobia, gender discrimination or civil unrest, it falls within the social sphere of impact. And lastly, in what is possibly the least understood or explored, at least in terms of the spread of misinformation during Covid-19, if claims sow mistrust of authorities and

government entities, which muddle how citizens engage in democratic and other decision-making processes, it falls within the political sphere.

There is a notable overlap within these spheres. For example, most classic conspiracy theories seem to demonstrate a mistrust of political power or leadership, yet, as the evidence showed, people who believed such theories were less likely to seek out science-based medical treatment for Covid-19. Equally, claims that create social stigma could stir such severe prejudice against a specific group that they lead to unwarranted attacks and injury of that group, which has physical and mental-health consequences.

Therefore, the classification of harm caused by misinformation during a health pandemic exists on an overlapping, intricate and highly complex continuum. Often, at the individual level, the psychological agony from believing health misinformation is an underlying consequence across the spectrum. For example, losing your savings to a scam has major psychological repercussions; the uncertainty about wearing a mask leads to anxiety and confusion; and reading racially polarising content creates a deep sense of unease.

The below model (Fig 1) demonstrates the spheres of impact at which harm occurs.

Figure 1: Spheres of impact

Table 2 goes further to include a list of evidence-informed types of harm that were documented in the academic literature during Covid-19:

Table 2: An evidence-informed risk framework of health misinformation

Sphere of negative impact	Implications of misinformation	Examples from the literature for each sphere of impact	
Physical harm	Lowered knowledge about health treatments; negative health-seeking behaviour, such unlikelihood to be tested or receive vaccinations; poisoning	Sallam *et al.* (2020); Allington *et al.* (2020); Freeman *et al.* (2020); Chary (2020)	Self-reported mental-health problems; spread of panic; increased paranoia; higher anxiety
Economic harm	Loss of revenue through scams and fake job adverts; panic buying that leads to supply chain interruptions and demand-supply loopholes	Ahinkorah *et al.* 2020	
Social harm	Stigma	Budhwani *et al.* (2020)	Ahmad & Murad. (2020); Sallam *et al.* (2020); Freeman *et al.* (2020)
Political harm	Reduced trust in government and other credible organisations; disregard for government guidelines; negative attitudes towards government's responses	Freeman *et al.* (2020); Georgiou *et al.* (2020)	

1.2 What can we learn about the level of risk that misinformation presents from our rapid evidence assessment?

Our rapid evidence assessment suggests that there is a wide range of risks caused by misinformation. These are presented below using the framework developed in Section 1.1 above. We draw on research from around the world, and on the risks of a range of types of social media (not only WhatsApp), about a number of public health topics (not only Covid-19). Where available, findings specific to misinformation about Covid-19 shared on WhatsApp in Africa are highlighted.

After systematic searching and screening to identify only those studies relevant to this review, we identified 22 studies. These are summarised in Table 3 below:

Table 3: An overview of the literature exploring risks and harm of misinformation on social media.

	WhatsApp	Facebook	TikTok	Instagram	Twitter	Not specified
Text	2	4	1	1	10	3
Video	1	2	1	1	1	1
Voice	0	0	0	0	0	0
Image	0	0	0	0	0	0
Not specified	0	1	0	1	1	7

It is worth noting that we found no studies that focused on voice or image messages.

Of the 22 identified studies, only two studies focused on WhatsApp specifically, one examined text messages alone (Allington *et al.*, 2020), and one examined both text and video messages (Kulkarni *et al.*, 2020). Both were focused on misinformation around Covid-19 specifically, but neither were based in Africa.

Eighteen of the 22 studies focused on specific countries or regions, and only 2 explicitly focused on Africa: one examined misinformation around Ebola outbreaks in West Africa (Roy *et al.*, 2020) and the other focused on Covid-19 misinformation across Africa (Ahinkorah *et al.*, 2020).

Our analysis of the risks and harm of misinformation on social media reported within these studies found these were often second-hand accounts or reported anecdotal cases of harm. Systematic analysis of the harm of misinformation about Covid-19 is largely lacking from the academic literature at this point in time.

Physical harm caused by misinformation

Our analysis of the three studies that reported physical harm due to misinformation on social media found no evidence specifically about WhatsApp, but all three did focus on harm caused by misinformation on Covid-19. In all cases, they reported anecdotal data on harm, with an awareness of risks (Moukarzel *et al.*, 2020), but no systematic analysis of the scale of potential harm due to misinformation.

The studies did report serious risks, including cases of:

- Poisoning as a result of swallowing or injecting bleach in the USA (discussed in Michael *et al.*, 2020).

- Hospital admissions and deaths due to methanol overdosing in Iran linked to reports on social media that it could prevent infection of Covid-19 (Soltaninejad *et al.*, 2020).

Economic harm caused by misinformation

In our analysis of the literature reporting economic harm due to misinformation on social media, we found no studies that focused on WhatsApp, and three that looked at social media in general. None focused on Africa specifically. All three did focus on harm caused by misinformation about Covid-19.

The studies describe serious risks of economic harm, including cases of:

- Food insecurity due to disruptions in food supply.

    - Fear spread on social media about food shortages and show closure led to stockpiling of food and chemicals in India (Sahoo *et al*., 2020). In addition to the financial burden of stockpiling, this behaviour is likely to have led to shortages and increased food insecurities among those who rely on small regular food purchases, particularly those with low socioeconomic status and other vulnerable populations.

- Indirect economic harm due to the destruction of physical resources, or diversion of resources away from their primary purpose. This includes

  - Cases of destruction of 5G towers across Europe caused by false claims that 5G caused Covid-19 (Naseem *et al.*, 2020).

  - Medical resources being diverted to cope with the results of misinformation, rather than the pandemic itself. The large number of cases of methanol poisoning in Iran led to increased demand for hospital beds, straining the health system (Soltaninejad *et al.*, 2020).

Social harm caused by misinformation

We found six studies that reported social harm due to misinformation on social media, although only one of the six included consideration of WhatsApp specifically (Allington *et al.*, 2020). Two focused on misinformation on Facebook (Allington *et al.*, 2020; Ahinkorah *et al.*, 2020), and three focused on Twitter (Allington *et al.*, 2020; Ahinkorah *et al.*, 2020; Budhwani & Sun, 2020). The others focused on social media in general. Only one discussed Africa specifically (Ahinkorah *et al.*, 2020). All six did focus on harm caused by misinformation about Covid-19.

The six studies did report serious risks of social harm, including:

- Reports of xenophobia and of stigma associated with Covid-19 (Ahinkorah *et al.*, 2020), and specifically concerns about xenophobia towards the Chinese because of the assertion that this is a 'Chinese virus', particularly in the USA (Budhwani & Sun, 2020).

- High levels of anxiety in China associated with social media messaging about Covid-19 (Gao *et al.*, 2020).

Studies that explore the potential of social pressures being associated with a greater tendency to believe, or perpetuate, misinformation found that:

- Stress levels about Covid-19 were not associated with a tendency to believe conspiracy theories. The authors of this study believe that this might be as a result of using a younger cohort for the study at a relatively short point in time since the inception of lockdown restrictions around the world (Georgiou *et al.*, 2020).

- Social media, particularly YouTube, was associated with negative social impact (Allington *et al.*, 2020).

Political harm caused by misinformation

While our rapid evidence assessment did not focus specifically on political misinformation, four of the studies about public health misinformation reported risks related to political harm. None of the four studies focused on WhatsApp specifically, and only one discussed Africa specifically (Ahinkorah *et al.*, 2020). Three of the four studies focused on harm caused by misinformation about Covid-19 outside of Africa, and the fourth focused on the 2014/15 Ebola epidemic in West Africa (Roy *et al.*, 2020).

The studies found:

- Negative associations with the Chinese/China because of the apparent source of the virus, linked to the USA President's description of this as the 'Chinese virus' (Budhwani & Sun, 2020).

- Anti-WHO sentiments arising from the Ebola epidemic in West Africa (Roy *et al.*, 2020).

It is worth noting that many governments are actively contributing to efforts to counter misinformation about Covid-19, as reported by Sahoo and colleagues (2020).

While not reporting political harm specifically, Duffy and Allington (2020) report findings of a UK-wide survey including associations between different political affiliations and beliefs about Covid-19. From these we might hypothesise that future elections may be influenced by misinformation about Covid-19 but will be highly dependent on the timing of the elections and the phase of the pandemic at the time.

Psychological harm caused by misinformation

Our framework for mapping risks of misinformation suggests that psychological harm is a potential primary impact (for example, depression), but also an indirect consequence of the physical, economic, social and political harm caused by misinformation.

Our rapid evidence assessment reports the risks of psychological harm as described in five different studies, all of which focus on Covid-19. None report psychological harm as a result of WhatsApp and none focus on Africa.

Ahmad and Murad (2020) report a strong association between panic and the use of social media, particularly among young people (18-35 years). Kawchuk and colleagues (2020) demonstrate how the promotion of therapies, especially by health professionals, that claim to boost immune systems – in their case spinal manipulation therapy – is widespread. Gao and colleagues (2020) report high levels of depression associated with Covid-19 information on social media in China. We also learn from Germany that those worried about pre-existing conditions were particularly likely to suffer psychological strain due to Covid-19-related media (Bendau *et al.*, 2020).

Lastly, there is an apparent association between those who tend towards conspiracy theories and believe misinformation on social media about Covid-19 and distrust government interventions designed to tackle the pandemic (Georgiou *et al.*, 2020). This suggests that misinformation on social media may feed into and perpetuate existing distrust of authority and disbelief about public health information.

1.3 What is the nature, and associated risks, of Covid-19 misinformation shared across WhatsApp in Africa and reported by users for fact-checking?

WhatsApp's end-to-end encryption which enables greater privacy, may have many advantages for individual users, but for fact-checkers, and anyone else tasked with the

battle against misinformation, it implies that misinformation continues to fester in the dark. Without access to this data, we know little about its nature and intent, nor its risks.

A summary of the nature and risks associated with claims that surfaced on WhatsApp for the period March to June 2020 in Kenya, Nigeria, Senegal and South Africa is presented below. Using a tip line where users flagged the messages to Africa Check, the content was then analysed according to the evidence-informed risk framework developed in Table 2.

Loosely categorised, 30% (n=67) of claims were associated with harm to physical health; 18% (n=39) of claims fell in the sphere of economic harm; 19% (n=42) of claims were categorised in the sphere of social harm; and 33% (n=74) of claims analysed were associated with political harm. There is a complex overlap in categorising the risks associated with misinformation. Each claim was categorised according to the sphere in which it was perceived to cause the most harm.

The description of the methodology and the data used to perform the analysis are available in the Annexures (A8).

Harm to physical health

- The majority of claims flagged to Africa Check that were categorised as harm to physical health were about false cures.

- A series of natural remedies and herbal 'cures' were flagged on the different WhatsApp lines. Although most of them may not lead to extreme harm to an individual's physical health, they could still prevent people from seeking evidence-based treatment for Covid-19.

- One should not underestimate how the promotion of these 'everyday cures' and falsities chip away at the evidence-base, with potentially devastating consequences. For example, the use of aspirin presented as a 'cure' to Covid-19 could lead to Salicylate/aspirin intoxication, which remains one of the most common accidental forms of poisoning, especially among children. A video that stated that Covid-19 is bacterial and can be treated with aspirin was shared by 50 different users, an indication of the virality of the claim, and its risk.

- Several messages on WhatsApp claimed that the medical profession wrongly attributed the cause of Covid-19 to being a virus. For example, one specific claim that "In Italy the cure for Coronavirus is finally found" relating to the fact that Covid-19 is a bacteria and not a virus, was detected with the same wording in Kenya and Nigeria (see the Annexure, A8).

- Similar text claims in South Africa state that Prevotella, the "bacterial genius" (sic) for Covid-19 has been found, while other messages attributed the cause of Covid-19 to "Intravascular Coagulation (Pulmonary Thrombosis)". It is interesting to note how the use of medical terms creates a sense of pseudo-accuracy that could mislead users into believing that it is credible information. Claims about masks often used the medical term "hypoxia", a condition in which the body or a region of the body is deprived of adequate oxygen supply at the tissue level.

- A claim that could be potentially harmful because it was presented in text format with no further context is that coronavirus cells could be killed by a 'head lice drug'. It shows how a misleading headline such as: "Researchers find common head lice drug kills coronavirus" (The New Daily, 2020) becomes even *more* misleading on WhatsApp in plain text format when the user cannot click to read the full article for more information. Fact-checkers such as Snopes (2020) found that, although preliminary results showed the potential for the effectiveness of Ivermectin (a drug used to treat animal parasites and head lice) to treat Covid-19, it was too early to draw conclusions. Additionally, some head lice treatments are neurotoxic. The Food and Drug Administration (FDA) in the US has warned that a drug named Lindane could lead to seizures or death, if taken incorrectly. Without the treatment name presented in the WhatsApp message, one can then easily understand how the circulation of a 'head lice drug as a cure for COVID-19' could have serious negative health implications.

- Another interesting observation is how the messages about health advice take on a very colloquial tone on WhatsApp. The words 'family' and 'friends' frequently appeared in the texts. Some of the claims literally made you think that you received a message from a distant, well-intentioned aunt. Because of the chatty, personal nature of many of the messages, it intuitively felt less threatening and more credible. This adds another dimension to the risks associated with health misinformation on WhatsApp. One of the findings in a 2020 study about mobile instant messaging in Nigeria and Pakistan was similar (Pasquetto *et al.*, 2020): *"Misinformation on MIMs is more personal, it sounds and looks like a suggestion from a close friend, rather than a top-down piece of information created by a specific group to influence another"*.

- According to Wardle and Derakhshan (2018), breaking news in a time of crisis is bound to unleash waves of misinformation, often with the intent to be *helpful* but failing to adequately inspect the veracity of the message, as highlighted during a terror attack on the Champs Elysees in 2017. This was also the case with Covid-19 misinformation on WhatsApp. In the messages, the help is offered by random individuals, some with clearly stated first names like "Victor" whose daughter is "a Pharmacologist (M.Sc) at Pfizer Pharmaceuticals in Italy" or "Leon" who tested positive for Covid-19, or a doctor or ICU nurse. It is important to note that the colloquial tone and seemingly helpful nature of misinformation on WhatsApp make it appear trustworthy, which could amplify the risks when compared with other social-media platforms.

- An interesting trend in Nigeria was how three classic conspiracies (the whistle-blower Chinese doctor, coronavirus being a bio-weapon, and the Madagascar fabrication) were used as a pretext to promote natural remedies like fever grass or the inhalation of hot water vapour (see A8 for more).

- Many of the messages also had the potential to discourage good behavioural practices like wearing a mask. In a series of slightly different messages, masks were presented as 'dangerous' to one's health as it led to reduced oxygen in the blood. Usage for "prolonged periods creates hypoxia".

- There were also several claims about masks and test kits being infected with Covid-19. In South Africa claims about masks being "doused with chemicals" as a criminal strategy to rob citizens also circulated.

Economic harm

▪ In a significant number of claims reported to Africa Check during the period, scammers have applied WhatsApp as a platform to capitalise on the economic vulnerability presented by the pandemic. This was a specific problem in Kenya and Nigeria. Examples of some of these claims are listed in the Annexure (A8).

▪ In many claims reported to the Africa Check tip line, convincing job opportunities at USAID or the WHO were advertised.

▪ It is also interesting to note that in two claims – one from Nigeria and one from Kenya – the wording of the claim remained exactly the same, only changing the currency and the link to enter personal details. It illustrates how many of these opportunistic claims travelled across countries.

▪ A secondary theme that presented itself in the recorded claims, was misinformation offering users a form of social relief. This was specifically the case with claims recorded in South Africa, where there were doubts about the SASSA grants and the distribution of food parcels.

▪ Some claims also encouraged panic buying with voice notes claiming that fuel or rice were running out. As the rapid evidence assessment shows, such claims can undermine supply chains and lead to a shortage of goods.

Political harm

▪ Conspiracy theories thrived on all WhatsApp lines. Scrutinising these theories with a closer lens showed that they all resembled a few overarching themes, none of which was unknown to fact-checkers, in that they also surfaced on other social-media platforms and were not WhatsApp-specific. They included conspiracies around 5G causing Covid-19; vaccination conspiracy theories, specifically the testing of vaccines on African citizens; conspiracies that the 'whistle blower' doctor who passed away from Covid-19 had a cure; that Covid-19 is a biochemical weapon from China; or variations of Madagascar quitting the WHO as member state because its Covid-19 treatment was denied by the international organisation.

▪ The Madagascar claim featured strongly on WhatsApp in most countries. The messages carried strong Pan-Africanist sentiments supported by views that could potentially be classified as anti-West. They were typically fuelled by claims, such as one in which the WHO is falsely quoted as saying that "a vaccine is very unlikely to come from Africa". Corresponding narratives unfolded in the public domain.

▪ The claim that the new coronavirus is a 'bio-chemical weapon' from China also surfaced frequently in all four countries, sowing mistrust of official narratives.

▪ Several "classic" Covid-19 theories around 5G, a new world order, Bill Gates depopulating the world, or general anti-vaccination theories emerged.

▪ Although most claims recorded were in text format (which often makes tracing their origin and nature harder because of the limited visual evidence that fact-checkers usually rely on to determine veracity), some claims were received in voice note format.

For example, one voice note claims that the Ghanaian President came across 'secret documents' about Covid-19, which prompted him to "release it to his people out of patriotism". It requires listeners to "please carefully listen to it and share widely to expose this evil". The voice in the voice note claims that he is reading extracts from an article on the Rockefeller Foundation website, which outlines a "detailed plan for the creation and spread of the Coronavirus and the lockdown of the world". One challenge identified during the analysis is that many of the external markers that fact-checkers usually rely on such as date, place, or identity of the speaker are obscured in voice note format.

- On the South African WhatsApp line, several conspiracy theories related to the SA government's ban on cigarettes and alcohol were documented. In one image, the headlines call for "Mrs Zuma to explain" showing boxes of cigarettes and alcohol in a warehouse owned by one of her family members. The claim states Dr. Dlamini-Zuma banned alcohol and drugs to make way for illicit dealing of her own stock estimated at R20 million. The WhatsApp messages all carry traces of the paranoia among SA citizens during South Africa's hard lockdown.

- The last category of claims that could lead to political harm, such as a distrust of authorities, were claims about information security. Claims from South African WhatsApp users about phone lines being monitored illustrate how misinformation bred growing paranoia during a time when citizens felt vulnerable to government authority.

Social harm

- Most of the broader conspiracy theories that were recorded were debunked by fact-checkers in one form or another. However, the real harm of WhatsApp lies in its more focused, regional-based misinformation that have not reached a level of popular debate. These are the clandestine, hate-sowing and polarising messages that flourish in closed WhatsApp groups between families and friends.

- Although the four Africa Check WhatsApp tip lines did not reveal high volumes of such claims, there were some examples recorded. It must be stated that the absence of such claims doesn't mean they were not circulating, it simply points to one of the limitations of using tip lines as a strategy for detecting harmful content on WhatsApp: the reliance on circumspect users to report such content.

- A few concerning examples were flagged to Africa Check in South Africa that fuelled the racist, right-wing narrative, linking it to newsworthy events related to the pandemic, such as the lockdown, relief packages, or police brutality. The Annexure (A8) gives examples.

- Other messages that fell into the category of social harm were conspiracies about a cruel strategy by the West "against Africans". For example, in one claim, Obama reportedly asks Africans not to accept a vaccine from the West starting with the phrase: "I'll be an accomplice if I don't denounce this evil act white people want to do to Africans". Several claims also circulated about the testing of vaccines by Western pharmaceutical companies on African children.

- When the anti-vaccination conspiracies were making harmful statements against another social group, they were classified in the sphere of social harm; yet it is evident that these

theories will also lead to a distrust in vaccination procedures, which will have an impact on people's physical health.

▪ Only 11 claims (5%) of the 222 analysed claims were received in voice note format. Of the 11 voice note claims, the majority (45%) were mapped at the sphere of social harm. This is an interesting observation, as voice notes are predominantly a WhatsApp-based format. We also know that voice-based messaging can be a very explosive and polarising format (think about how radio was utilised during the Rwandan genocide to spark the killings). Fact-checkers need to guard against the personal, yet convincing tone of voice notes in sowing social harm. In one of the voice notes flagged to us, a claim is made in Wolof that people have died after being vaccinated. It calls on the people of Mali to "fight the people who introduced the lethal vaccine in Mali and in Africa". In another voice note, the speaker says the entire staff of the Jumbo store in South Africa has Covid-19, which has been concealed by management who wants to "maximise profits at the cost of black lives". Both these claims could polarise.

▪ Consistent with the academic literature, many of the claims reported to Africa Check had the potential to sow Sinophobic sentiments, claiming that blankets and masks sent from China were full of the coronavirus.

Conclusively, we find that the nature of health misinformation on WhatsApp presents specific health, economic, social and political risks because of three factors: access, format and tone:

▪ Access

▪ Without fact-checkers being able to access claims due to end-to-end encryption, health misinformation on WhatsApp presents a high risk of causing harm as it may go undetected, continuing to survive and thrive in the underbelly of the WhatsApp information ecosystem.

▪ Format

▪ Text: health misinformation tipped to fact-checkers are often in plain text format, which provides little context, does not allow back-clicking and limits the evidence trail that fact-checkers usually rely on to debunk misinformation

▪ Voice: equally, voice notes limit many of the typical external markers that fact-checkers rely on to determine the veracity of a message, including date, place, visual cues, or speaker identity. Voice as a format can also be highly explosive and polarising. Because of the 'personal' nature of WhatsApp, it may appear to be more trustworthy

▪ Tone

▪ The friendly, helpful tone of many of the health misinformation messages on WhatsApp make them seem sincere and trustworthy, which amplifies the risk of people sharing, believing or acting on their content

▪ The use of medical terms in some health misinformation messages creates a sense of pseudo accuracy. It makes the misinformation seem as if it is credible advice from medical experts

1.4 What did we learn from African fact-checkers that helps explain risks and harm better?

In addition to reviewing the literature on the risks and harm of Covid-19 misinformation and assessing the risks evidence within the WhatsApp misinformation reported to Africa Check, we also interviewed nine individuals working with fact-checking organisations in Africa to understand their experiences. The details of our data collection and analysis are provided in the annexures. Table 4 below provides an overview of who we spoke to.

Our interviews with fact-checkers across Africa suggest they experience a wide range of risks and harm in their work. They shared with us some of the contributing factors to these risks, as the themes below illustrate.

Wide range of sources of misinformation

Fact-checkers told us about the lengths to which people in different capacities have gone in spreading misinformation. From the fact-checkers' experience, misinformation messages (where traceable) originate from a wide spectrum of sources, ranging from political and religious bodies to the common person. They highlighted how citizens were particularly susceptible to misinformation shared by people in authority, such as religious leaders. Further, it was interesting to see religious leaders adopting messages from political leaders to inform their own misinformation messages.

Conspiracy theories

Fact-checkers categorised many of the messages they reviewed as deceptive in their intention. As indicated above, they also stressed how many of these messages originated from religious leaders. The following excerpt from the Democratic Republic of the Congo illustrates this:

> *There were audios being sent by pastors who were denying the virus … Some said the vaccines were coming from Bill Gates to place chips on us to track us. So people should not take any interventions from the West – Editor, DRC.*

Such messages were judged by the interviewees to be likely to derail their fact-checking efforts, particularly those related to possible future vaccination drives. As governments around the world were facing challenges when providing medical gear for medical professionals, there were messages that would have frustrated these efforts. For instance, read below the experience from Ghana:

> *Donations from Jack Ma and China … videos were advising not to use the PPE that came from China, even though countries such as Ghana were grappling with finding this equipment. We researched and found out the owner of the videos, we went to a military warehouse where samples were taken and done – the tests showed that they were not contaminated. After producing the fact-check report we shared it with other journalists, social-media platforms to actively counter the misinformation – Editor, Ghana.*

Table 4: Overview of key informants during interviews

	Interviewees									
Fact-checker's role within organisation	Researcher	Editor	Reporter	Editor	Reporter	Editor	Fact-checker	Editor	Fact-checker	
Fact-checking organisation's main location	Nigeria	Kenya	Senegal	South Africa	Nigeria	DRC	Kenya	Ghana	France	
Fact-checking organisation's countries of operation	Guinea, Nigeria, Sierra Leone	Burundi, Cameroon, Central African Republic, Ethiopia, Kenya, Mali, Niger, Tanzania, Uganda	Côte d'Ivoire, Senegal, South Africa	Kenya, Nigeria, Senegal, South Africa	Nigeria	Central African Republic, Congo, Dem. Rep	Kenya, Malawi, Rwanda, South Sudan, Tanzania, Zambia	Ghana	Cameroon, Congo, Dem. Rep, Ethiopia, Ghana, Kenya, Mali, Morocco, Nigeria, Senegal, South Africa, South Sudan, Tanzania, Uganda, Zambia, Zimbabwe	
Experience of fact-checking organisation in the fact-checking field	Have some experience (3-8 years)	Have some experience (3-8 years)	Have some experience (3-8 years)	Established (8-15 years)	New to this work (0-3 years)	New to this work (0-3 years)	New to this work (0-3 years)	New to this work (0-3 years)	Have some experience (3-8 years)	

Fake cures / treatments

The interviewees explained that, although it might seem logical for people to lean on their health institutions to provide health guidelines to fight the pandemic, this was not necessarily the case everywhere. Messages in various formats, including audio and video messages, carried non-sanctioned seemingly health-oriented messages shared through WhatsApp and other platforms. These messages claimed to provide health remedies for those infected by Covid-19. For instance:

> *'People claiming chewing onions is a cure'. These claims were being shared in vernacular languages – Editor, Ghana.*

> *If you steam, you'll cure the virus – Misinformation reported by Editor, Ghana.*

Fact-checkers also told us of misleading messages that made xenophobic claims. For example:

> *Messages claim 'Infections are based on skin colour' – Fact-checker, France.*

In relation to this claim was a reference to COVID-19 as the Chinese/China virus, which led to harassment of Chinese nationals in various countries by members of the public.

The influence of public figures

As mentioned earlier, fact-checkers explained that people are more inclined to believe and further propagate information they come across from figures of authority. One interviewee described a case from Nigeria where a member of the clergy shared what seems to be false hope:

> *One member of the clergy was telling his congregation in Nigeria that a cure for the virus has been discovered in the US. Also, that it is merely like a fever. The cure was Hydroxychloroquine. This was as a result of the claims that were made by frontline doctors based in the US and their press briefing. The harm is that people who have underlying conditions such as Arthritis and Lupus who need Hydroxychloroquine for regular use – Reporter, Nigeria.*

As explained by the fact-checker, the effect of such messages on a section of the society (those vulnerable), if not debunked, would be grave.

"Fake news"

Inaccurate news also flooded the messaging platforms, including WhatsApp. For instance, there were videos carrying information perceived as credible. This is probably due to the novel nature of the virus and the uncertainty of a cure. One interviewee explained:

> *Claims made by certain clergy claiming they have drawn links between 5G and COVID-19. That 5G causes the virus. Not in Nigeria, but elsewhere there have been people pulling down 5G installations because they claim it causing the virus. This results into wastage of resources – Reporter, Nigeria.*

Threats and violence during the COVID-19 pandemic

Finally, interviewees explained how misinformation about Covid-19 was causing people to be violent towards others. The messages propagated seemed to appeal to the element of fear and distrust in recipients. Consider the quote below telling people to distrust those who would come to them in the name of providing solutions to the pandemic:

> *Beat people who are coming to talk to you about the virus because they are coming to buy corpses of people who have died to get financial support from financial institutions such as the World Bank – Misinformation reported by interviewee from the DRC.*

Section 2: How are users responding to public health misinformation on WhatsApp?

This section tackles questions relating to how users respond to misinformation on WhatsApp, focusing particularly on Covid-19 in Africa, while drawing lessons from other forms of social media, and from research outside of the continent too.

To address these questions, we have:

- Conducted a rapid evidence assessment of the literature on public-health misinformation on social media

- Conducted a survey of WhatsApp users in Africa, targeting particularly those in Kenya, Nigeria, Senegal and South Africa

Details of our methods are reported in the annexures.

2.1 What did we learn about user responses to public-health misinformation on social media from our rapid evidence assessment?

Our findings draw on research from around the world, and on user responses to a range of social media types (not only WhatsApp), about a number of public-health topics (not only Covid-19). Where available, findings specific to misinformation about Covid-19 shared on WhatsApp in Africa are highlighted.

After systematic searching and screening to identify only those studies relevant to this review, we identified 18 studies. These are summarised in Table 5 below:

Table 5: An overview of the 18 studies included in this rapid evidence assessment

	WhatsApp	Facebook	TikTok	Instagram	Twitter	Not specified
Text	4	9	1	1	5	5
Video	1	1	1	1	0	3
Voice	0	0	0	0	0	1
Image	0	1	0	0	0	2
Not specified	1	2	0	1	2	1

Of these 18 studies, five studies included a focus on WhatsApp, although only one of these looked at information about Covid-19 (Kulkarni *et al.*, 2020). Of the 18 studies, two focused on information about Ebola in Nigeria (Ahmed *et al.*, 2019; Adebimpe *et al.*, 2015), while

Wasserman and colleagues (2019) looked at misinformation sharing across five African countries (Kenya, Namibia, Nigeria, South Africa and Zimbabwe), and Pasquetto and colleagues (2020) looked at misinformation in Nigeria, India and Pakistan.

How users respond

Within the rapid evidence assessment, 11 studies describe user responses in terms of the following behaviours (see Table 6):

Table 6: Themes of user responses within the literature

User responses to information	Size of the literature (number of studies)
Check for cues	2
Try to verify the information	3
Reported the message	1
Delete if suspected misinformation	0
Post a correction	1
Ignore	1
Share anyway	6
Acted on the message	1

Whilst the literature is still limited on exactly who and why users take the decisions to respond in these ways, we can learn from the available evidence. Undergraduate students, for example, check for cues about the validity of information, including looking for the verification check on Twitter messages (Wasserman *et al*., 2019). The same students report that they don't use fact-checking services but agree that checking the source of information would be useful. We also know that students are more likely to try and verify information before sharing when they are identifiable. Talwar and colleagues (2019) found that users who are more likely to share personal information online with others are more likely to check the validity of information before sharing it, while those who have higher trust in information on social media are more likely to share it without checking it. Only one study found that recipients of information (just over 12% of them) report misinformation (Tandoc *et al*., 2020). However, this study, which combined a survey and in-depth interviews, was conducted in Singapore, which is technologically and economically advanced relative to most African countries. It is perhaps not surprising that this was also the only study that found that people would post a correction to the misinformation on their own social media and notify the sender of the misinformation. Despite the 'positive' responses to misinformation among these Singaporean respondents, the majority (73%) still said they would ignore the messages altogether (Tandoc *et al*., 2020).

Evidence from Indonesia explored the link between users' scope for recognising misinformation and their tendency to share it without verifying it first (Khan & Idris, 2019). These were not linked, suggesting that, even when users can understand that something is misinformation, they share it anyway. Five other studies similarly found that users share information anyway; some do not distinguish whether it is verified or true (Zollo *et al*., 2015), particularly those who are 'epistemologically naïve' (Chua & Banerjee, 2017),

while others suspect that it is not true and still share it (Kulkarni *et al*., 2020; Adebimpe *et al*., 2015). Young people surveyed in Nigeria about the Ebola virus had relatively high understanding of the virus, and yet many still reported acting on misinformation (Adebimpe *et al*., 2015).

Of particular concern is a finding by researchers in the UK and Europe that those social media users with more followers are more likely than those with few followers to share information that they know to be false, to do so with confidence, and to attach 'evidence' to their messages by quoting an external source (Arkaitz *et al*., 2015).

What motivates users' responses?

We examined the motivations for the different responses of users and found that the type of content, who had shared it with them, the emotions it triggered, their trust of social media (or their fear of it, 'cyberchondria'), and their tendency towards conformity, all shaped their behaviour (see Table 7).

Table 7: Themes of user motivations within the literature

Motivation behind their responses	Size of the literature (number of studies)
Type of content (Funny, Helpful)	2
Who shared the content (Interpersonal network, Government source or officials, Legitimate news source)	8
Emotions (Good news or bad news, Anger or fear)	4
Trust info on social media to be true	3
Cyberchondria	1
Conformity	4

From these studies we learn the following:

▪ Users share information if they think it is helpful to others, out of a sense of civic duty, and they do so even if they are not sure if the information is accurate, if they judge its potential benefit to outweigh any potential harm (Wasserman *et al*., 2019). They are also more likely to share on WhatsApp information which they think is funny or weird (Wasserman *et al*., 2019), but do not respond at all to misinformation if they are not particularly interested in the topic (Tandoc *et al*., 2020).

▪ We explored whether user responses to misinformation differed according to whether or not the information came from a trusted network, an official organisation, or a legitimate news source. We did not find any evidence about legitimate news sources. Perhaps not surprisingly, the evidence suggests that users are more likely to share information received from within a trusted personal network (Cronkhite *et al*., 2020; Pasquetto *et al*., 2020; Tandoc *et al*., 2020; Wasserman *et al*., 2019). Users are also more likely to re-share corrections received from a family member, close friend or like-minded individual (Pasquetto *et al*., 2020). They are also more likely to act on misinformation received from a family member (Ahmed *et al*., 2019). Users tend to aggregate around similar content (Bessi *et al*., 2015), creating echo chambers in which certain views are reinforced and others rejected.

- If information was received from perceived official sources, it was more likely to prompt a user response. Ahmed and colleagues (2019) found respondents were more likely to act on advice if medical professionals were also trying the treatment, while users' own trust in political institutions shaped their trust (or distrust) of information from governments (Cronkhite *et al.*, 2020). Lastly, Huang and Carley (2020) found that news and government sources were less likely to share misinformation than personal accounts.

- Emotions play a role in how people respond to misinformation, including if they are indifferent to news in general (Tandoc *et al.,* 2020; Tiago *et al.,* 2019) and if the information itself arouses specific emotions: Wasserman and colleagues (2019) found that users are more likely to share information which makes them feel emotional, and which makes them feel patriotic. Chua and Banerjee (2017) focused on what they described as epistemic belief, which they define as 'perceptions about the characteristics of knowledge and the process of knowing'. They found that those who are 'epistemologically naïve' are more likely to spread misinformation, while those who are 'epistemologically robust' stifle misinformation by stopping its spread.

- Perhaps not surprisingly, trust in social media played a key role in how users responded to information they received, even if they were uncertain how true it was (Ahmed *et al.*, 2020; Samuli *et al.*, 2020; Talwar *et al*, 2019). Fear of social media, referred to in some of the literature as *cyberchondria*, is explored by Samuli and colleagues (2020). They find that women are likely to suffer from it.

There is some evidence that a desire to conform to the 'norm' shapes users' responses to misinformation:

One study found that those who use social media more often are more likely to share content (Tiago *et al.*, 2019). Some evidence suggests that respondents are less likely to support a fake news story if they read comments by others that are critical of it (Colliander, 2019). This may be due to pressures of social comparison and social enhancement online (Talwar *et al.*, 2019). Wasserman and colleagues (2019) even suggest that sharing fake news online is a form of social currency.

What characteristics of users are associated with different responses?

The evidence base on whether men or women are more likely to share or act upon misinformation is inconclusive. While there is some indication that women are less likely to respond at all to misinformation (Samuli *et al.*, 2020; Tandoc *et al.*, 2020), this is contradicted in other studies (Adebimpe *et al.*, 2015; Ahmed *et al.*, 2020), while others found no difference by gender (Khan & Idris, 2019).

The evidence on age is similarly mixed. Some studies suggest young people are more likely to act on misinformation (Ahmed *et al.*, 2020). There is some indication that older people are less likely to share unverified information (Samuli *et al.*, 2020), even though young people blame older people for sharing misinformation (Wasserman *et al.*, 2020).

The evidence base on the role of education level in user responses is both small and inconclusive (Cronkhite *et al.*, 2020).

2.2 What did we learn about responses to Covid-19 information from our survey of WhatsApp users?

To help us understand WhatsApp user responses to Covid-19 misinformation, we designed an electronic self-administered survey and disseminated the survey to the Africa Check and ACE networks.

The survey targeted adults living in Africa, particularly residents in South Africa, Senegal, Nigeria, and Kenya, which was the geographic focus of this study. About 4% of respondents were from other African countries, mainly Benin, Côte d'Ivoire, Cameroon, and Zimbabwe. The electronic survey was snowballed primarily using the Africa Evidence Network and Africa Check mailing lists, and respective social media channels. As a result of this technique, and the fact that the research team was largely South African based, the sample showed a bias towards South African respondents.

A total of 286 WhatsApp users responded to our survey about misinformation on WhatsApp; 53.5% were women and 45.5% were men (1% of respondents withheld this information). While the majority of respondents were based in our target countries of Kenya, Nigeria, Senegal and South Africa, respondents were based across 17 African countries (see Table 8).

Table 8: The geographical spread and gender of respondents to our survey

		Male	Female	Withheld	Total	Percentage
		130	153	3	286	100%
Benin		2	0	0	2	0.70%
Burkina Faso		1	0	0	1	0.35%
Burundi		2	0	0	2	0.70%
Cameroon		4	1	0	5	1.75%
Côte d'Ivoire		3	2	0	5	1.75%
DRC		2	0	0	2	0.70%
Gambia		0	1	0	1	0.35%
Guinea	Gender by country	1	0	0	1	0.35%
Kenya		14	23	0	37	12.94%
Mali		1	0	0	1	0.35%
Nepal		1	0	0	1	0.35%
Nigeria		34	19	1	54	18.88%
Rwanda		2	0	0	2	0.70%
Senegal		25	5	0	30	10.49%
South Africa		36	102	1	139	48.60%
Tanzania		0	0	1	1	0.35%
Zimbabwe		2	0	0	2	0.70%

Respondents spanned the age range from 18 to over 61, although the majority were between 21 and 60 years (see Table 9). Our largest response base was in the 31-40 age group.

Table 9: The age profile of respondents to our survey

		18-20	21-30	31-40	41-50	51-60	61 and over	Total
		4	57	103	63	35	24	286
Benin		0	1	1	0	0	0	2
Burkina Faso		0	0	1	0	0	0	1
Burundi		0	0	1	1	0	0	2
Cameroon		0	0	2	3	0	0	5
Côte d'Ivoire		0	1	1	1	2	0	5
DRC	Age distribution by country	0	0	1	1	0	0	2
Gambia		0	0	0	1	0	0	1
Guinea		0	0	0	0	0	1	1
Kenya		1	6	22	5	3	0	37
Mali		0	1	0	0	0	0	1
Nepal		0	0	1	0	0	0	1
Nigeria		3	15	16	15	5	0	54
Rwanda		0	0	0	0	2	0	2
Senegal		0	15	9	4	0	2	30
South Africa		0	18	46	31	23	21	139
Tanzania		0	0	1	0	0	0	1
Zimbabwe		0	0	1	1	0	0	2

We asked respondents to tell us about their education level, specifically which of the following categories they fell into (Table 10). Despite our attempts to reach a wide range of people, the majority of our respondents had further education after high school (see Table 11). When we look at the responses that users provided, we need to bear in mind that they are not by any means representative of all WhatsApp users in Africa.

Table 10: Education levels

1	I did not finish primary school
2	I finished primary school
3	I started high school but I did not finish
4	I finished high school
5	I attended further education after high school
6	I prefer not to say

Table 11: Respondents by education level (key provided in Table 10)

		1	2	3	4	5	6	Total
		0	0	8	22	246	10	286
Benin	Education	0	0	0	0	2	0	2
Burkina Faso		0	0	0	0	1	0	1
Burundi		0	0	0	0	2	0	2
Cameroon		0	0	0	0	5	0	5
Côte d'Ivoire		0	0	0	0	4	1	5
DRC		0	0	0	0	2	0	2
Gambia		0	0	0	0	1	0	1
Guinea		0	0	0	0	1	0	1
Kenya		0	0	0	0	35	2	37
Mali		0	0	0	0	1	0	1
Nepal		0	0	0	0	1	0	1
Nigeria		0	0	1	5	47	1	54
Rwanda		0	0	0	0	2	0	2
Senegal		0	0	0	3	23	4	30
South Africa		0	0	7	14	116	2	139
Tanzania		0	0	0	0	1	0	1
Zimbabwe		0	0	0	0	2	0	2

We were keen to understand how WhatsApp users behaved towards the messages that they received on Covid-19. We did not explicitly ask them about misinformation, as we were aware that respondents did not necessarily know whether the information they received was accurate; we wanted to hear from them what they found helpful (or not) and what they did about it. They provided explanations for why they forwarded messages on to others and why they might change their behaviour in response to a message.

How users responded to misinformation

When asked to explain how they reacted to Covid-19 information, WhatsApp users gave a range of answers (selected from a list of provided options; they were encouraged to tick all that apply). Table 12 shows the frequency of each response and the percentage of total responses. All six respondents who ticked 'other', went on to describe actions linked to verifying the accuracy of the messages.

Table 12: Responses from users to Covid-19 messages on WhatsApp

Response	Frequency of response	Percentage of response
I have forwarded a message to individual contacts	101	17.24%
I have forwarded a message to one or more WhatsApp groups	93	15.87%
I have asked the sender of the message about its accuracy	82	13.99%
I have deleted a message because I thought it was false	94	16.04%
I have reported a message (for example, via Africa Checks' misinformation line)	29	4.95%
I have acted on the information changing my behaviour	84	14.33%
I have done nothing	97	16.55%
Other	6	1.02%

Why users forwarded messages to others

When WhatsApp users were asked to explain why they forwarded messages about Covid-19 to others (either individuals or groups), they provided a range of explanations. These included a desire to raise awareness about Covid-19, their feelings (confidence) in the message itself, a desire to impact others positively, a sense that the messages provided valuable alternative sources of information, a direct desire to counter misinformation messages, and sometimes just because they found messages funny. Each of these reasons was identified from survey respondents' answers and are illustrated below using direct quotes from the survey (see Table 13).

Table 13: Why users forward Covid-19 messages on WhatsApp

Theme	Illustrative Quotes
They wanted to raise awareness about Covid-19	"My husband pastors a church. We created a distribution network for general plus COVID info. If something was excellent e.g. a video on how to sanitise or wash hands, or the value of masks, we distributed it". "I conveyed the message to warn against the danger of this pandemic, because many people in my country did not believe in it". "Because I believed in them and I wanted to protect them too".
They had confidence in the integrity of the message content	"I forwarded reliable and authentic information and data to other groups for good governance and accountability". "Occasionally, I share news or reports that I deem to be credible".
They wanted to positively impact others and achieve behaviour change	"To guide them against getting affected". "Thought the information was educative and can help in behavioural change during the pandemic".
They saw it as an alternative source of information for others	"Relevant and correct Info that is not being said by the mainstream media". "Some friends also do not follow news platforms, so I share relevant information with them sometimes".
They considered it to be a way of dealing with misinformation	"It'll also help curb misinformation that's rampant". "Once verified, it was relevant to forward - e.g. the risks of using masks with vents". "I have forwarded some Africa check messages to neighbours and employees because they are useful guides for detecting lies. And because my neighbours have sent me false messages".
They found it humorous	"Usually, I forward COVID-19 jokes".

What shapes users' responses to messages

We asked WhatsApp users which factors influence their response to Covid-19 messages. They selected from a menu of options and were invited to tick all that were relevant. Regard for the sender and individual perception on whether the information is correct or incorrect were the most common influences listed. Table 14 provides a summary of users' responses. The 65 people who indicated that their response 'depends on other factors', described these other factors in more detail. They talked about the credibility of the message, and about their scope to verify its contents. They also talked about experiential and behavioural factors, such as their own training or experience, or how the message made them feel. A few people also referred to whether or not they felt the message would help others.

Table 14: What shapes users' responses to messages

Responses	Frequency of response	Percentages
My response depends on which person sends it to me	129	23.45%
My response depends on whether it comes from an organisation that I have heard of	131	23.82%
My response depends on whether it gives good news or bad	39	7.09%
My response depends on whether I think the message is true or not	143	26.00%
My response depends on which language the message is in	16	2.91%
My response depends on what format it takes (for example, text or image)	27	4.91%
My response depends on other factors	65	11.82%
Total responses	550	100.00%

Why users might change their behaviour in response to messages

When asked to explain why they acted on information and changing their behaviour, respondents gave a number of reasons, as illustrated in Table 15. These included the argument that the information helped them to improve their personal hygiene and implement physical protection measures. They felt that it helped them to implement behaviour change, and to take preventative supplements. They also valued the sense of greater personal awareness.

Table 15: Why users change their behaviour in response to messages

Theme	Illustrative Quotes
The information helped to improve hygiene	"Mandatory use of hand sanitizer, handwashing" "I started washing my hands frequently"
The information helped to implement physical protection measures	"When I got a WhatsApp message on how to put on the mask and saw it was a better way of putting on the mask" "Took masks more seriously based on a video"
The information helped with behaviour change	"I continue to physical distance and not go to places I would normally go" "I understood that young people could be infected and I started by respecting the protective measures"
The information advised on preventative supplements	"I started taking supplements to improve my immune system" "I took some vitamins / minerals to help my immune system like the message said it would help"
The information increased personal awareness / created awareness more generally	"I reassessed my risk profile and took extra precautions" "I have become more responsible about managing my mental and physical health to optimise my body's disease fighting chances"

What shapes whether users believe a message is true?

Survey respondents were given several statements from which to pick multiple options, on which sources they have the most trust in. Responses were varied (see Table 16). Thirty-five percent of users told us they had the most trust in legitimate news sources.

Table 16: What gave social media users confidence in a message

Response	Frequency of response	Percentage of response
I trust a message when it comes from someone whom I respect	87	15.18%
I trust a message when the message comes from a legitimate news source	201	35.08%
I trust a message when it comes from an organisation that I know	125	21.82%
I trust a message when it comes from a government source (for example, the Ministry of Health)	160	27.92%
Total responses	573	100.00%

How users choose to share messages they believe to be true

When users were asked to explain whether their responses to information depended on whether or not they think the information is true, they gave a range of answers. They explained how their responses depended on their ability to verify the information, on their personal judgement, and on the source of the information. They also explained how it depended on their own previous experience with and knowledge of false information on social media, and on the strength of their own feeling of responsibility. These explanations are summarised in Table 17 and illustrated with direct quotes from respondents.

Table 17: How users choose to share messages they believe to be true

Theme	Illustrative Quotes
It depends on my ability to verify the information	"If it's true, share, if it's not I discard" "I research more about it. E.g. WHO says there is no cure yet… but someone says there is then I go with WHO"
It depends on my personal judgement	"If I believe it's true, I will share it" "Some messages are sent out to create panic and I don't respond / act on those"
It depends on the source of the message	"There are so many false information being spread I usually don't respond to anything until it was noted by the president" "Yes, here, even if the message is coming from W.H.O for instance, I still need to do a fact-check to ensure it is true and verifiable"
It depends on respondents' previous experience with and knowledge of false information on social media	"I am very wary of fake news, conspiracy theories and the like. My first response is always a sceptical one"
It is shaped by the respondent's feeling of responsibility	"Spreading misinformation is wrong and a problem" "There was too much negativity going around. If one doubts the authenticity, it is easier just to delete the message rather than encourage more fake news"

What shapes users' response to Covid-19 messages

When users were asked to explain what shaped their response to Covid-19 messages on WhatsApp, they gave a variety of explanations (see Table 18). Some told us it depended on the reliability of the message, or on the possibility for independent verification, while others referred to the type and source of the message. Some respondents told us that their response to a message was driven by their desire to be helpful towards others. Some respondents simply said their response to Covid-19 messages depended on their mood.

Table 18: What shapes users' responses

Theme	Illustrative Quotes
It depends on the reliability of the message	"Scientific references from a reputable organisation"
	"I believed in the Corona virus at first, but I don't believe in it anymore. I think it's a setup against the Africans"
	"Medical or it comes from a reputable organisation."
It depends on the possibility of independent verification	"Reliability, plausibility based on my scientific knowledge, cross-checked by several established organisations"
	"If I am able to identify that the source of the message (organisation / other) is reliable / credible"
It depends on the type and source of the message	"I try not to forward videos. I don't trust them. And also they are often big files that block up my storage on my phone so I usually just delete videos"
	"Religion: If it's from a religion other than mine and its directed at members of that religion, I usually just ignore to avoid unsavoury confrontations with adherents of that religion"
	"If a message has no name or the person has no qualification I will be suspect and will first check it out. If the message is scary, inflammatory or sensational I will first check it out"
If it helps others to be informed	"If I think it will be helpful to others"
	"I will only send something on if I believe it is true and it will add value to other recipients."
It's a personal decision	"My response depends on my mood. I ignore most messages. Information overload and often dubious information from sensationalist sources. I only scan some messages very briefly and only properly read messages if I have a question I want an answer to"
	"I am a doctor with many years of experience in managing epidemics"

Why users reported messages they determined were false

When users decided that a message was false, some indicated that they reported the message. We asked them to tell us more about what they did, and they gave the following responses (see Table 19).

Table 19: Why users reported misinformation to fact-checkers

Theme	Illustrative Quotes
Because it is a way of dealing with alarming messages with possible negative impact on the masses	"When the message is taking people astray" "I reported it if the message made me suspicious or denigrated a sector of the population in sweeping terms. An example is the message about police confiscating food given by church members to a white community in the Vaal Triangle early on in lockdown. It had photos of thin pathetic children" "I reported messages that I felt could cause harm, influencing people to believe that the virus isn't real and not to follow the protocols"
To stop spread of fake / unverified information / feeling of responsibility	"Because, is just to avoid spreading fake news, else when the true news come peoples might not take it seriously..." "In order to put an end to the action of disinformation"
Prior knowledge of fact-checking platforms that I could report to for fact-checking	"Facebook makes reporting fake news easy" "Forwarded to What's Crap on WhatsApp and or Africa Check"
Because of uncertainty about the validity of the message	"If the message sounds true and sincere, but I cannot find verification online"

Which organisations users trusted

When users indicated that they trusted messages from particular organisations, we asked them to explain more. Their answers spanned government ministries, international organisations, NGOs, research institutes and even specific journals and academics, media houses, civil-society organisations, and finally private-sector organisations including private-health associates.

In summary: User responses to information about Covid-19 on WhatsApp

It was evident that users of WhatsApp continued to be bombarded with a great deal of information regarding the Covid-19 pandemic. Some information they accessed by searching for various sources, but also those shared by colleagues, relatives, and loved ones, among others. At the back of these developments was a population most likely gripped by fear and uncertainty about what the future holds in terms of managing the effects of the pandemic. There were various factors that determined how WhatsApp users interacted with the information being passed to them. This spanned from sharing the same messages, to changing their behaviour based on the information, or doing nothing about messages, while others sought clarification about dubious messages.

Various factors determined the way WhatsApp users dealt with information and, in some instances, it depended on personal judgement. Some managed to verify the information through various platforms, including the WHO, while others solely relied on personal judgement. We also realised that there were various factors that influenced people's responses, including seeking to help others, the reliability of the information, and the nature of the message. Some made the decision to report information that they thought was alarming to fact-checking organisations. One of the reasons given for reporting alarming messages was to curtail any further spread of false information. Prior knowledge of fact-checking platforms also played a role.

We learnt that in as much as WhatsApp's security features present challenges to fact-checking processes, a good number of the respondents knew about verification processes. Some use the most accessible tools, such as Google, to try and verify information before deciding what to do with the information they receive. This might indicate that awareness drives about verification tools and fact-checking organisations must continue and be accelerated, so that people are better equipped to deal with misinformation directly on the platform.

Section 3: How can we better respond to the risks that misinformation presents to help tackle the pandemic in Africa?

This section tackles questions relating to how fact-checking organisations and others can address misinformation on WhatsApp, focusing particularly on Covid-19 in Africa, while drawing lessons from other forms of social media, and from research outside of the continent too.

To address these questions, we have:

▪ Conducted a rapid evidence assessment of the evidence base to both identify strategies and understand how effective they are

▪ Spoken to WhatsApp users across the continent to gather views on mitigating misinformation

▪ Conducted an interview study of fact-checking organisations in Africa, specifically in Kenya, Nigeria, Senegal and South Africa

Details of our methods are reported in the annexures.

3.1 What we found about mitigating strategies from our rapid evidence assessment?

We identified 43 studies describing mitigating strategies for tackling misinformation. These mostly focused on Twitter, but also included WhatsApp, Facebook, TikTok, Instagram and general social media (see Table 20). They also focused on various message formats, the most common being text (see Table 21). The majority of the studies either examined Covid-19 specifically or public-health misinformation more generally (see Table 22).

Table 20: Social-media platforms considered in relation to mitigating strategies

Social-media platform	N
WhatsApp	9
Facebook	14
TikTok	2
Instagram	5
Twitter	19
General 'social media'	10
Not specified	2

Table 21: Message formats considered in relation to mitigating strategies

Message formats	N
Text	36
Video	8
Voice	4
Image	8
Not specified	9

Table 22: Message content considered in relation to mitigating strategies

Topic of misinformation	N
Covid-19	16
Political	2
Other public-health information	13
Other not listed above	15

Of these 43 studies, only four studies focused on Africa. One of the four covered Pakistan and India, as well as Nigeria (Pasquetto *et al.*, 2020).

If we were to focus only on those studies exploring mitigating strategies for Covid-19 misinformation on WhatsApp in Africa, our evidence would be limited to two studies:

▪ Africa Check & Volume (2020). Case Study: "What's Crap on WhatsApp?"

▪ Bowles, Larreguy and Liu (2020): Countering Misinformation via WhatsApp: Evidence from the Covid-19 Pandemic in Zimbabwe

We therefore cast our net more widely to include all studies on strategies to counter public-health misinformation (including, but also broader than just Covid-19) on social media anywhere in the world. We then examined the mitigating strategies that these studies focused on. These can be grouped into nine approaches (see Table 23). Each one is described in more detail below.

Approach to mitigating misinformation	Number of studies describing use of these strategies
Credible information over misinformation	15
Self-efficacy to detect misinformation	9
Make misinformation illegal	4
Infoveillance	5
Technical	8
Debunking	20
Social media companies tackling misinformation on their plat-forms	8
Collective action against misinformation	6
Social media campaigns	4

Promoting credible information to counter misinformation

A number of studies promote the idea of providing credible accurate information with a wide range of public-health information about awareness, healthy behaviour, and improved outcomes, among others (Al-Dmour *et al*., 2020), as a means for countering misinformation. It is suggested that a wide range of organisations could be mobilised to achieve this, including local community organisations as well as mass media (Ahinkorah *et al*., 2020). Health agencies in particular are urged to communicate over social media (Alvarez *et al*., 2020). The argument is that, by providing accurate information via trusted sources, you can both combat misinformation and promote positive behaviour change (Bowles *et al*., 2020; Pulido *et al*., 2020).

Supporting self-efficacy to detect misinformation

Supporting self-efficacy to detect misinformation is a strategy which aims to teach people how to identify and recognise false information on social media (Naeem *et al*., 2020). Some studies outline factors to consider before liking, sharing or ignoring public-health information (Alvarez *et al*., 2020; Armitage *et al*., 2020), while others have put an emphasis on the importance of ehealth literacy to support self-efficacy to detect misinformation (Alvarez *et al*., 2020; Eysenbach, 2020). It is believed that ehealth literacy plays a key role in improving the extent to which individual citizens are resistant to health misinformation. However, some researchers argue that this literacy should be focused on specific socio-demographic groups (Cronkhite *et al*., 2020; Khan & Idris, 2019; Morley *et al*., 2020).

Making misinformation illegal

In response to the rising Covid-19 infodemic, some countries criminalised misinformation. Which means that those who are found responsible for the production and propagation of Covid-19 misinformation online are charged and made to face the law; this is intended to serve as a warning to other potential culprits (Ahinkorah *et al*., 2020). In Peru, the Ministry of Justice and Human Rights released a statement that people who share fake news or misinform others to obtain a benefit or to perturb the public tranquillity can be sanctioned

with a prison sentence (Alvarez *et al*., 2020). Other countries, including the UAE, have had stronger legal restrictions on misinformation since before the pandemic hit. Kabha and colleagues (2019) demonstrated how the UAE's legal framework on misinformation is much stronger than those in India and the UK, enabling them to tackle misinformation shared on WhatsApp much more effectively. This includes penalties for social media users who share fake news.

Infoveillance

According to Eysenbach (2020), infoveillance is the "continuous monitoring and analysis of data and information exchange patterns on the internet". Infoveillance allows one to detect outbreaks of misinformation, rumours, or falsehoods; and to counter them in a timely fashion with facts or other interventions. This can be done through data-mining algorithms (Ahinkorah *et al*., 2020). Infoveillance requires a proactive and agile public-health online presence (Alvarez *et al*., 2020).

Technical approaches

A technical approach to curbing misinformation means employing innovative technologies to detect and debunk misinformation. Aldwairi and Alwahedi (2018) suggest a project that could be used to detect and filter out sites containing fake news. Anderson (2018) also identified technological tools that could help those who aim to debunk misinformation, tools such as the "fake tweet generator" and the reverse image search tool. A number of studies used computer algorithms to identify users who are more likely to be susceptible to misinformation (Baeth & Aktas, 2019; Krishna Kuma & Geethakumari, 2014).

Debunking

Most of the studies identified in the review explored different ways of debunking misinformation. They argued that debunking should take place on the same platform as the misinformation (Africa Check & Volume, 2020; Ahmed *et al*., 2020). This is because people will not go to a website to read the counteracting report, but they will watch a video or a voice memo sent via WhatsApp or posted on a social-media platform (Ahmed *et al*., 2020). Some studies argued that presenting accurate information to counter a rumour will decrease the likelihood of sharing behaviour (Ozturk *et al*., 2015; Bode & Vraga, 2018; Van der Meer & Jin, 2019), while others argued that a correction is not enough and that there is a need to consider the worldview of the receiver, their cognitive ability and media literacy (Cronkhite *et al*., 2020; De Keersmaecker & Roets, 2017; Krishna Kuma & Geethakumari, 2014). It was recommended that counter-information should be timely, simple, brief and tailored for the target audience (Zollo, 2019; Ozturk *et al*., 2015).

Social media companies tackling misinformation in their platforms

To curb the spread of misinformation, social-media companies have been encouraged to act on misinformation on their platforms. The studies we identified explored ways in which this could be done. These included social-media companies partnering with health institutions to build a system that detects misinformation and scammers, to minimise the effects of health

misinformation (Ahinkorah *et al.*, 2020; Xu *et al.*, 2020; Iosifidis & Nicoli, 2019). Social media companies can also act quickly in deleting accounts which are set up to spread conspiracy theories (Ahmed *et al.*, 2020; Alvarez *et al.*, 2020; Arkaitz *et al.*; 2015). Finally, social media companies should design their technologies in a way that allows users to self-correct (Ozturk *et al.*, 2015). As described above, the legal framework in the UAE has enabled social media companies to be held to account on the spreading of fake news, since before the pandemic (Kabha *et al.*, 2019).

Collective action against misinformation

A number of studies explored the idea that interventions from multiple stakeholders are essential in curbing misinformation. From the studies, stakeholders include local communities, social-media platforms, health organisations, civil society, public authorities and figures, tech companies, mass media, physicians and medical associations (Ahinkorah *et al.*, 2020; Kouzy *et al.*, 2020; Naeem *et al.*, 2020; Sahoo *et al.*, 2020; Ahmed *et al.*, 2020). All these stakeholders should make efforts to circulate accurate information and correct misinformation. Collective action also means action from social media users who play an important role in curbing the spread of misinformation. Users have the responsibility of correcting other users when they share misinformation and checking the validity of the information before sharing (Pasquetto *et al.*, 2020; Vraga & Bode, 2017a).

Social-media campaigns

Some studies explored the use of social-media platforms to raise awareness through social media campaigns. These campaigns were used to inform the public with the aim of changing behaviour (Al-Dmour *et al.*, 2020; Kulkarni *et al.*, 2020). In Peru, the Ministry of Justice and Human Rights encouraged people to share only official information using the hashtags "Don't Spread #FakeNews" (Alvarez *et al.*, 2020). The World Health Organisation (WHO) has also been able to closely track and identify the most prevalent rumours relating to the coronavirus through a campaign titled #KnowTheFacts (Sahoo *et al.*, 2020). This has enabled social-media platforms to be part of the solution, a space to communicate health advice in real time in order to curb misinformation (Swire-Thompson & Lazer, 2019).

We further went on to explore the people involved with these strategies and found huge variety.

▪ First, we found a range of people reported as responsible for misinformation on social media, including verified accounts (for example, on Twitter), which are particularly problematic because people trust what they publish and when they coordinate with one another to propagate misinformation (Abbasi & Liu, 2013; Zollo, 2019). The evidence also points to accounts which are dedicated to spreading conspiracy theories (Abbasi & Liu, 2013). The third group which propagates misinformation is social media and news outlets which deliberately publish misinformation to grab attention and increase readership (Aldwairi & Alwahedi, 2018).

▪ Second, we identified a range of people focused on implementing strategies to tackle misinformation on social media. These include civil society focused on social media (Bowles *et al.*, 2020) and other aspects of local communities, the media itself (Alonso & Alemañy-Castilla., 2020), including social media companies (Vraga & Bode, 2017a;

Iosifidis & Nicoli, 2019), fact-checkers (Africa Check & Volume, 2020), and lastly, social media users themselves (Sojung & John, 2020).

- Third, these strategies to mitigate against misinformation were targeted at particular groups of people: WhatsApp users (Bowles *et al*., 2020; Reis *et al*., 2020), broader social media users (Pennycook *et al*., 2020; Baeth & Aktas, 2019; Sang *et al*., 2020), traditional media, and both the public and those perceived to be at high risk from or susceptible to misinformation on social media (Cronkhite *et al*., 2020). The latter included strategies focused on older people (Swire-Thompson & Ecker, 2017), young people (Naeem *et al*., 2020), and those with low cognitive abilities (De Keersmaecker & Roets., 2017).

3.2 What do we know about the effectiveness of strategies to tackle misinformation on social media?

While the literature includes a wide range of strategies for tackling misinformation, there is a relative dearth of information on whether they actually work. Of the 43 studies in our rapid evidence assessment that described strategies for mitigating misinformation, only ten also tested their effectiveness in relation to five outcomes areas (see Table 24).

Table 24: Mitigating strategies described in the research

Outcome area	Size of the evidence base (number of studies)
Attitude adjustment	2
Behavioural changes	1
Truth discernment	6
Responsiveness to correction	5
Psychological outcomes	4

Table 25 below illustrates which studies were measuring which outcomes in relation to which mitigating strategies. It is striking which interventions have not been evaluated at all (Making misinformation illegal; Infoveillance; Technical; Social media campaigns), and the evidence base on the others is very small.

Table 25: Outcomes and interventions

Interventions	Outcomes				
	Attitude adjustment	**Behavioural changes**	**Truth discernment**	**Responsiveness to correction**	**Psychological outcomes**
Credible information over misinformation	0	0	1	1	0
Self-efficacy to detect misinformation	1	0	1	0	0
Make misinformation illegal	0	0	0	0	0
Infoveillance	0	0	0	0	0
Technical	0	0	0	0	0
Debunking	1	1	5	4	4
Social media companies tackling misinformation in their platforms	0	0	2	2	0
Collective action against misinformation	0	0	1	1	0
Social media campaigns	0	0	0	0	0

When we examine the evidence that is available in relation to each of the outcomes reported, we find the following:

Adjusting users' attitudes and beliefs

We found two studies that examined how to adjust users' attitudes. One study found that it may be more effective to affirm facts than to counteract misinformation as a mitigation strategy (Swire-Thompson *et al.*, 2017). The second study, conducted in the USA, found that those with lower (vs higher) levels of cognitive ability are less able to adjust their attitudes to incorrect information after it has been corrected (De Keersmaecker & Roets, 2017).

Behaviour change

The single study we found that examined behaviour change (above merely changing attitudes), was conducted in Israel. It found that credible messaging can shape the behaviours and actions of health students and professionals in seeking further information to correct misperceptions in public-health discourses (Gesser-Edelsburg *et al.*, 2018).

Truth discernment

We found six studies which examined social media users' ability to discern the truth within messages. Most of this evidence was from the USA; no studies were conducted in Africa. A study conducted in the US found that, when people are forced to consider the accuracy (or inaccuracy) of information, they are more likely to engage actively in truth discernment (Pennycook *et al.*, 2020). Indeed, this paper also found that there is a greater likelihood that people will not share information if they suspect that it is a misrepresentation of the truth (Pennycook *et al.*, 2020). Using Twitter as a platform, another study found that strategies that are designed to counter the peddling of rumours on social media were effective in reducing the tendency of people to share misinformation (Ozturk & Sakamoto, 2015). Focused on a US university, a study found that the use of humour-related strategies in dispelling misinformation encouraged robust engagement on correcting inaccurate information (Sojung & Cook, 2020). Another study, also conducted in the US, found that the provision of corrective information was a viable strategy for debunking misinformation by improving awareness and discounting false beliefs about crises (Van Der Meer & Jin, 2019). In order to ensure the effectiveness of messaging in public health, a further study from the US concluded that clear messaging was a pivotal part of refuting and correcting false information (Vraga & Bode, 2017a). Another key finding from a similar study in the same country was that agreement and disagreement with false statements can play a role in reducing misperceptions about outbreaks (Vraga & Bode, 2017a; Vraga & Bode, 2017b).

Responsiveness to correction

We identified five studies, all from the USA, which zoomed in on how social media users responded to correction. These suggest that people respond positively to social media and algorithm correction that is aimed at tackling belief in misinformation (Vraga & Bode, 2017a). They also suggest that there is a higher degree of credibility that is put on non-humour (vs humour) correction to aid the demystification of misinformation (Sojung & Cook, 2020). Van der Meer and Jin (2019) determined that sensitivity to untruths is enhanced by perceptions and underlying attitudes towards crises, while Vraga and Bode (2017b) found that people could refute false information when corrected or encouraged to engage more with viable sources of facts.

Psychological outcomes

Three studies from the USA, and one from Israel, all examined psychological outcomes of strategies to mitigate misinformation. In the USA, researchers found that people with higher cognitive ability were more willing to accept corrections to misrepresentations of truth. Those with lower cognitive ability were more sluggish in their reactions but adjust well to correction (De Keersmaecker & Roets, 2017). Another study in the USA established that corrective information, as an intervention, can improve attitudinal perception and emotional stability (Vraga & Bode, 2017a). For a US sample of 700 adults, there was greater self-efficacy for those who received targeted communication from an authority on health matters than those who received general information (Van der Meer & Jin, 2019). Lastly, the Israel study found that the correction of misinformation for those of medium self-efficacy could encourage the search for more robust information and fact-checking.

3.3 What did we learn about mitigating strategies from our survey of WhatsApp users?

While our user survey focused primarily on how users respond to Covid-19 misinformation on WhatsApp, respondents also described strategies and solutions to misinformation.

When asked to provide a general feedback on Covid 19 misinformation on WhatsApp, 20% of survey respondents provided what could be classified as strategies and solutions to misinformation in general and the Covid-19 pandemic itself. These include the use of accessible language, access to credible information, scope to verify information, and responsible social-media engagement. Some of these mitigating strategies have been extensively reported on by various authors in the rapid evidence assessment and include infoveillance, debunking, caution, social-media campaigns and collective effort and action against misinformation, among other potential strategies and solutions. Table 26 illustrates these approaches using quotes from survey respondents.

Table 26: Approaches for mitigation proposed by survey respondents

Approaches proposed	Illustrative quotes
Accessible Language	"Accurate translation of English version of COVID-19 awareness and prevention messages to local languages and pretesting before circulation" "The information should be in a simple language and graphics"
Access to credible information	"Simple messages on how to deal with some of the symptoms of COVID 19 is effective and easily believable" "It would be great to have proper info in terms of messages from the ministry of health"
WhatsApp information verification	"The owners of WhatsApp can create a feature that verifies whether a message about COVID19 is true or false, just as we have on twitter"
Verification marks on WhatsApp and responsible social-media engagement	"It would be useful, in the case of messages originating from verifiable organisations, for the forwarded message to contain a marker (such as the Verified Account on Twitter) that persists with the message" "There is need to verify information received before forwarding it to someone or groups"
Preventative measures including public awareness / education	"The public education and enlightenment should be continuous" "I have noticed that many people share the first thing they receive so if the government puts out graphics immediately after the president's speech, they share that. So, a good way to counter misinformation is to get correct info out early and in ways people can share"

3.4 What we found about mitigating strategies currently used in Africa (especially Kenya, Nigeria, Senegal and South Africa) from our interviews with fact-checking organisations

Our second exploration of mitigating strategies involved interviewing eight individuals working with fact-checking organisations in Africa, to understand their experiences. The details of our data collection and analysis are provided in the annexures, and an overview of interviewees is provided earlier in this report, in Table 4.

From the interviews, we have identified six distinct mitigation strategies used by fact-checkers. These are listed below and then described in turn.

1. Setting up or joining WhatsApp groups with feedback mechanisms

2. Using other social-media platforms, for example, Twitter & Facebook

3. Accepting individual submissions

4. Use of organisational media platforms, such as newsletters and Facebook

5. Collaboration with tech companies and relevant institutions to track misinformation on WhatsApp

6. Collaboration with media organisations and journalists to track misinformation on WhatsApp

Fact-checkers shared interesting approaches that they use to track misinformation on WhatsApp. Overall, there were systems that had been put in place and some of these were derived from previous fact-checking projects. The most common strategy involved **setting up or joining WhatsApp groups with feedback mechanisms**. The fact-checkers were already members of numerous WhatsApp groups, some of which were created for the sole purpose of tracking misinformation about the Covid-19 pandemic. Some of the fact-checkers utilised private WhatsApp groups to track misinformation despite the fact that such groups carried considerable 'noise', information that was not necessarily Covid-19 related. Below are the experiences of a fact-checker from the DRC:

> *We have joined over 30-40 WhatsApp groups and they are about many topics, but they are also used to spread information on COVID-19 – Editor, DRC.*

Equally interesting was **using other social-media platforms**, such as Twitter and Facebook, to strengthen the efforts of monitoring on WhatsApp groups. The experience of fact-checkers had been that, due to the restrictive nature of WhatsApp, other social-media platforms had to be utilised to track misinformation:

> *We rely on our Facebook and Twitter accounts to track misinformation – Fact-checker, France.*

Members of these platforms have shared misinformation that originated from WhatsApp with fact-checking organisations.

Fact-checking organisations also **accepted individual submissions from the public** after they actively sought health misinformation. For instance:

> *We are actively asking people to send us questionable claims and we also have the newsletter which contains the most popular debunks of the content that we have published – Editor, Kenya.*

Use of other organisational media platforms was another approach used and periodic newsletters featured prominently. These newsletters are used to request readers to share information that originated from WhatsApp that required verification, as seen in the excerpt below:

> *Its really hard to track misinformation on WhatsApp because of the end-to-end encryption, we therefore solicit [claims] through the newsletter asking people to share what they receive through WhatsApp – Editor, Kenya.*

There are specific channels provided in these newsletters for submitting this information to the fact-checking organisations.

Another strategy that stood out was **collaborating with tech companies and relevant institutions** to track misinformation on WhatsApp.

> *We benefited from the International Fact-Checking Network (IFCN) … which helped us access a chatbot which helped people have access to our platforms and share information – Editor, Ghana.*

The IFCN chatbot has been built to address the challenge of misinformation especially related to Covid-19. With this technology, individuals and fact-checkers in various countries are connected to the largest database of untruths about Covid-19.

There was yet another fact-checking organisation sharing their experience with engaging WhatsApp in using their API to track misinformation, as demonstrated in this excerpt:

> *With the pilot of the WhatsApp API the process of tracking claims is more automated, and it automatically logs a submission. It also would tell us how many people have submitted that claim, and how much it is being shared elsewhere online – Editor, South Africa.*

Finally, fact-checking organisations **collaborated with media organisations and journalists to track misinformation on WhatsApp** and some of their experiences were as follows:

> *The mechanisms are working with the local radio managers so that the managers contact us where there is information that they cannot clarify. They would like the organisation to fact-check and provide info that can counter the claim – Fact-checker, Kenya.*

> *Because we are experienced journalists, we feature in several groups, where we receive information that is being posted on WhatsApp – Editor, Ghana.*

Specific strategies that fact-checking organisations are using to fact-check content on WhatsApp

The study realised that fact-checking organisations were continuously adopting ways to ensure that their fact-checking projects keep up with prevailing situations. Depending on the context (country), the strategies have included using learnings from experiences with outbreaks such as Ebola, for example, in the DRC.

The study noted the following strategies that fact-checkers have in place to verify information on WhatsApp (see Table 27). We have also included quotes from those we interviewed on these strategies:

Table 27: Mitigating strategies described by African fact-checkers

Strategy used	Illustrative quotes
Developing capacities of fact-checkers, for example, training, providing guidelines for fact-checking	"We have done some trainings with journalists. Four journalists in every radio station were trained, given tools, video verification, how to verify websites, given a manual on how to do fact-checks. After this training, most of the trainees were interested in doing fact-checking" "We helped create 13 fact-check desks across five regions of the country. They are located in major media organisations in the North, empowering them how to fact-check, also collaborating on producing fact-checks. They report in local languages. We expect them to produce fact-checks in their local languages. We trained them on video verification platforms such as Invid, reverse image, Tin eye on the phone, Google images etc. to be able to look at what has been shared on WhatsApp"
Using editors as fact-checking gatekeepers and deploying journalists for on-the-ground verification	"We have a newsroom on WhatsApp, we also have our offices. We share first the claims in the newsroom and there are editors such as me who try to understand the context, then we give instructions to journalists on where to get verification. Sometimes we have to send journalists on the ground"
Developing / using technology for fact-checking	"We use Google reverse image, Invid for videos to see the metadata or videos, to see if the videos have been altered etc."
Partnerships / collaboration with institutions for fact-checking purposes	"We are currently having a conversation with people at WhatsApp about this service (API) that they are offering to fact-checking organisations. The API is an interface where you can build tools, and have a database of what we have debunked"
Ascertaining the potential harm of a claim	"We search to see if a piece of information has been shared on other platforms such as Facebook and Twitter to see how widely information has been shared. If found in other platforms it is considered worth pursuing to debunk/verify"
Partnerships with local journalists / media for translation and fact- checking	"We have reporters who are able to speak and understand local languages to get the information that we want. We also have stringers, as a wire, because of the bureau; we have photographers and stringers on the ground that we go to for translation. We look at the messages that our reporters receive which is also challenging depending on the countries of origin"

Discussion and learning

This project is best described as rapid research. We have collected and analysed a large amount of data. In less than four months from contract to report, we have conducted three rapid evidence assessments, a survey of social-media users, a series of interviews with fact-checkers and an analysis of claims reported to Africa Check. As with most research, we have more data than we have been able to analyse, and the analysis we have conducted is not as complex as we might have wished. Nevertheless, at a time of urgent need, we have collated a vast body of knowledge on public-health misinformation on social media.

Perhaps not surprisingly, we found a large number and a wide range of risks and harm due to misinformation on social media (see Table 28). Both the rapid evidence assessment and the claims reported to Africa Check suggest that this is a serious issue; people's lives and their livelihoods are at risk. While Covid-19 has largely been treated as a health crisis, the evidence suggests that there is a mental-health crisis following close behind, and that misinformation is playing an active role in this. While the psychological harm caused by misinformation clearly needs to be considered and addressed, it is not surprising that anxiety and depression have increased as a result of the pandemic.

Table 28: An overview of the risks and harm identified in this project

Risks and harm	Identified from the rapid evidence assessment	Identified within the WhatsApp messages reported to Africa Check
Physical harm	Hospitalisations and even deaths from attempted preventative measures, and presumed cures	30% of claims were associated with harm to physical health
Economic harm	Food insecurity and wasted resources	18% of reported claims fell in the sphere of economic harm
Social harm	Xenophobia and stigma	19% of claims describe social harm
Political harm	Distrust of the state	33% of claims were associated with political harm
Psychological harm	Anxiety and depression	All the claims were potentially damaging to psychological health

The claims reported to Africa Check on WhatsApp revealed how the private messenger app presented platform-specific risks in terms of access, format and tone.

Our interviews with fact-checkers shone additional light on the scale of the issue, as they highlighted the wide range of sources of misinformation and the involvement of influential individuals, including religious leaders. They also helped us understand the depth of the conspiracy theories being shared, and the number of preventative therapies and fake cures linked to Covid-19. They highlighted the influence of 'fake news' sites, and the potential political unrest as messages referred directly to anticipated violence and political instability.

Our rapid evidence assessment of the ways in which social media users respond to misinformation suggested that users can and do assess the validity of information. They check for cues within the messages, try to verify the source of the information and its content, and occasionally report false messages or post a correction. Our review does suggest, however, that not everyone responds in these ways by any means, and those who do check the accuracy of messages are more likely to be better educated, students, or older. Some populations may also be more predisposed to recognise and act on misinformation. For example, in Singapore the population is known to have relatively high education levels and be familiar with technology. Not all these lessons can be generalised to the African context. Most of the time social-media users appear to delete messages, ignore them, or even just share them anyway. Some change their behaviour in response to messages, which is clearly potentially risky, depending on whether they have understood the validity of the information.

When we explored the evidence base to understand what motivates users to respond to misinformation in particular ways, we did find a strong theme of responsibility to those within their social circle. For example, people are more likely to share information that they judge to be helpful with those they love and to alert their social circle about misinformation. However, they also feel social pressure from those they know and, if a message has been shared and supported by people they know, they are less likely to question its validity.

Also of relevance to those trying to counter misinformation is the level of trust that social-media users put into certain sources of information. Official organisations are more trusted, such as departments of health, but so is anyone claiming to be a medical professional, or even a religious leader. This presents issues, if these individuals are intentionally or inadvertently spreading misinformation.

Our survey of WhatsApp users in Africa suggests that people do share Covid-19 messages widely in their networks. They do also question their accuracy and report misinformation, either by asking the sender about it, or by reporting the message to an official organisation such as Africa Check. This last finding may be shaped by the nature of respondents to our survey, which had a particularly high education level, and who had self-selected to respond to us, suggesting an awareness of Covid-19 misinformation.

When asked what motivated them to share messages with others, respondents talked about a desire to raise awareness about the pandemic and provide helpful information to those they care about. This sense of responsibility may be linked to the finding from our rapid review that social-media users are strongly influenced by their own social circles and a desire to help. Users also told us that they acted on information that they felt would improve their health and those they care about.

Our survey findings also supported the review finding that users' responses to messages are strongly shaped by who they receive messages from, and that they have greater trust in messages from legitimate news sources and those they consider to be in positions of greater knowledge or greater authority, including health professionals.

In considering how to counter public-health misinformation on social media, we identified potential mitigating strategies from a rapid evidence assessment of what has been implemented to tackle public-health misinformation on social media. This research was enhanced with contributions from our survey respondents and via our interviews with African fact-checkers. In all, we identified nine broad approaches to mitigating misinformation (see Table 29). The three approaches common to all three parts of our research are those that relate to:

- Self-efficacy to detect misinformation – users want and need to be enabled to detect misinformation

- Verifying or debunking information via reliable organisations or groups that can assess the validity of information and

- Public awareness campaigns about misinformation

Table 29: An overview of the misinformation-mitigating strategies that we identified

Mitigating strategies identified from the rapid review	Mitigating strategies proposed by survey respondents	Mitigating strategies described in our interviews with African Fact-checkers
Credible information over misinformation	Access to credible information	
Self-efficacy to detect misinformation	Responsible social-media engagement Accessible language	Fact-checkers accepting individual submissions
Make misinformation illegal		
Infoveillance		Use of organisational tracking tools
Technical		
Debunking	WhatsApp and information verification	Setting up / joining WhatsApp groups with feedback mechanisms
Social-media companies tackling misinformation in their platforms		Collaboration with tech companies & relevant institutions to track misinformation on WhatsApp
Collective action against misinformation		Collaboration with media organisations and journalists to track misinformation on WhatsApp
Social-media campaigns	Preventative measures including public awareness/ education	Using other social-media platforms, for example, Twitter & Facebook

Although criminalising misinformation was not raised as a strategy by WhatsApp users or fact-checkers, it deserves a mention in the context of African leadership and the implication

that such approaches could have on freedom of expression and human rights. Under the guise of protecting the population against the virus, the Covid-19 pandemic led to some African countries introducing new laws, digital information surveillance, and heavy-handed military control reminiscent of previous authoritative regimes (BBC, 2020). Any strategy proposing the criminalisation of misinformation needs consideration of the socio-political risks and possible human rights red flag related to the country tabling such solutions.

Further to the various mitigating strategies identified, the evidence base on which, if any, of these strategies work is extremely thin, mostly conducted in the USA, without a focus on WhatsApp specifically and as such should be translated to African contexts with caution.

What we can learn from this evidence is:

- It may be more beneficial to affirm facts with credible messaging than retrospectively attempting to debunk misinformation. This has the potential to adjust social-media users' attitudes and beliefs about the pandemic and change their behaviour.

- Users can be supported to discern the accuracy of information through prompts to encourage them to question the validity of information, and through clear messaging about misinformation. It is not clear whether humour can help people to critique misinformation, or whether non-humorous messages are taken more seriously.

- Correction of misinformation does appear to support users in adjusting their beliefs, although some social-media users are unlikely to respond to any debunking efforts. Users' confidence in their own abilities to sort fact from fiction appears to play a role in whether they will change their belief in false information. This suggests that there may be a role for campaigns that promote self-efficacy in relation to Covid-19 information on social media.

Five evidence-based strategies for fact-checkers fighting future infodemics

The Covid-19 infodemic is not only a communications emergency, as declared by the UN when it launched the #PledgeToPause campaign in October 2020. Considering the risks, it is also a crucial public-health issue. Like most public-health issues, prevention is central to mitigation. The global public-health drive to reduce car accidents by campaigning for safety-belt use is a good example. It required a multi-sectoral approach in which car manufacturers improved safety-belt technology, traffic departments legislated its usage, governments launched road-safety campaigns and individual behaviour eventually shifted towards more positive outcomes. It will take time and concerted efforts by many stakeholders to fight future infodemics. Approaching the problem from a public-health paradigm can help with mitigation.

From this research, we identify five evidence-based strategies for fact-checkers to consider during future infodemics.

Using risk classification to enable more deliberate editorial and fundraising strategies

In the now chronic fight against misinformation, there is practical value in attempting to classify the harm of misinformation in the sphere in which it can do the most damage. This study identified four major risks, with the fifth one (harm to psychological health) being overlapping: harm to physical health, economic harm, social harm and political harm.

Risk classification could help fact-checkers to determine where they are making a difference and guide them to be more deliberate in their editorial decisions and discussions with donors. For example, they can tailor a theory of change to focus on reducing the economic harm caused by health misinformation, only targeting claims that are scam-related, and have funding discussions with donors interested in increasing financial literacy. If they decide to address social issues, they can purposefully select claims that could lead to stigma, start conversations with social-justice non-profits, and partner with like-minded organisations for greater impact.

Many fact-checkers may choose editorial balance over niching specific topics and may therefore tackle the underlying risks of misinformation holistically. In this case, the various spheres of harm provide a useful classification system for fact-checkers to demonstrate that they covered the news agenda proportionally and addressed issues of concern in a fair and balanced manner.

Further to mapping risk more broadly, the research also highlighted WhatsApp's specific risks as a medium because of access, messaging format and tone. Some fact-checking initiatives, such as Africa Check and Volume's What's Crap on WhatsApp, have already deployed innovative strategies to tackle these risks, by working around the access hurdle, and debunking misinformation in voice note format directly on the same platform as the post containing the misinformation.

When designing strategies to fight health misinformation on WhatsApp, fact-checkers must consider the platform-specific risks.

Developing proactive key messages and positive reinforcement around information consumers' social need to be helpful in a time of crisis

A common thread that runs through this report centres around the theme of helpfulness. The nature of the health misinformation analysed on WhatsApp often took on a personal, helpful tone that deceivingly made it seem sincere and earnest. Respondents in our survey said they shared Covid-19 information because they were motivated by a desire to raise awareness about the pandemic and provide helpful information to those they care about. The literature confirms that users share information if they think it is helpful to others, out of a sense of civic duty, especially in a time of crisis.

This may have been the case with the onset of Covid-19 when humanity was confronted with existential ideas about the meaning of life and the human condition. Could the social desire to contribute positively during a global health crisis have enabled people to unwittingly share or even deliberately create 'helpful' misinformation?

Although our research did not aim to find the answer, the finding that information consumers have a social desire to be helpful in a time of crisis has implications for fact-checkers. For example, media literacy campaigns can promote the concept of helpfulness in key messages; debunks posted on social media can include phrases such as "help us bring an end to misinformation by sharing this message"; even fundraising strategies could appeal to audiences' desire to be helpful by asking for individual donations.

The human need to be helpful may have contributed to misinformation spreading unabatedly on WhatsApp, but the same need could be used by fact-checkers to appeal to bringing an end to the infodemic.

Leveraging users' social circles to champion evidence-based health information about Covid-19

Linked to the idea of helpfulness is the concept that WhatsApp users are strongly influenced by their own social circles, specifically friends and family. From the literature, we learn that users are more likely to share information received from within a trusted personal network. From a positive perspective, they are more likely to reshare corrections received from a family member, close friend or like-minded individual. Yet, more negatively, they are also more likely to act on misinformation if they received it from a family member or friend.

When people have access to the antidote or inoculation against misinformation (ie, the accurate, fact-based information or the debunk), they can widely disseminate this information to their social circles.

Africa Check is experimenting with the use of "Fact Ambassadors" who will distribute fact-checking content (including fact-checking reports, factsheets and guides) and media literacy content, using their social-media accounts among their network of peers. The lessons learnt from these innovative approaches will be important in crafting future mitigation strategies around the idea of social influence.

Extending the fact-checking 'Circle of Trust' by building partnerships with trustworthy media, government bodies, civil-society partners, religious leaders, and big tech companies

The research emphasises the important role of trust in certain information sources. For example, users place more trust in official organisations, such as departments of health, but also in anyone claiming to be a medical professional, or even a religious leader. Our survey findings confirm that users have greater trust in messages from legitimate news sources and those they consider to be in positions of greater knowledge or greater authority, including health professionals.

Although many fact-checkers are already following media and other partnership models to achieve greater impact and reach, the notion to involve government departments of health, civil-society organisations, and vetted religious leaders could be further explored in Africa. The involvement of reputable religious leaders, specifically, may be a potentially under-explored strategy.

In our evidence assessment of mitigating strategies, we also highlight the role that social-media companies could play in discouraging misinformation on their platforms.

During the pandemic, several examples emerged of WhatsApp seeking out partnership models to fight health misinformation. For example, at the start of the outbreak, WhatsApp partnered with the WHO to introduce a health-information chatbot called Health Alert, developed by a South African NGO, Praekelt.Org (WHO Newsroom, 2020). In four days, the chatbot garnered 10 million subscribers (Protocol, 2020).

WhatsApp also launched the *WhatsApp Coronavirus Information Hub* in partnership with UN partners, the WHO, UNICEF, and UNDP. The Hub provides simple, actionable guidance for health workers, educators, community leaders, non-profits, local governments, and local businesses that rely on WhatsApp to communicate.

Of specific importance is the WhatsApp Coronavirus Information Hub's partnership with the IFCN.

As mentioned in one of the interviews with fact-checkers as a strategy to fight health misinformation, the IFCN launched a chatbot on WhatsApp to connect millions of users with the work of 99 fact-checking organisations from 70 countries worldwide. As the largest global collaboration of fact-checkers yet, the IFCN's #CoronaVirusFacts Alliance produced a database of over 9000 Covid-19 fact-checks, which have been translated from English into Spanish, Portuguese and Hindi. The WhatsApp chatbot feature allows citizens to check whether content about Covid-19 has already been rated as false by professional fact-checkers. In November 2020, the global initiative received international recognition when it announced that it received specialised support from the Paris Peace Forum (Poynter, 2020). It is indeed a laudable initiative, indicative of what could be achieved if fact-checkers collaborate with big tech companies at a global scale.

It is important that initiatives like the #CoronaVirusFacts Alliance also find expression through regional channels such as the Africa Facts Network.

Close collaboration between WhatsApp and fact-checking organisations is critical so that fact-checkers can provide inputs into new or proposed technological changes, especially when these changes call for lobbying and awareness-raising of their potential negative impact. A recent example is WhatsApp's announcement in a blog that it is introducing a 'disappearing messages' feature. When this feature is turned on, new messages sent to a chat will automatically disappear after seven days, a feature, the messenger app claims, will improve privacy. However, this feature could also eliminate evidence trails for fact-checkers on an encrypted platform where access to harming content already presents a challenge.

To effectively fight health misinformation on WhatsApp, African fact-checkers need to seek out regional relationships with the messaging platform.

In summary, the strategy for African fact-checkers to extend their 'circles of trust' is relevant because of the significant role trust plays in information-consumption patterns.

Trust-based partnerships can provide credibility and extended reach, especially for newer African fact-checking organisations that are still building their reputations and public awareness; it could be an avenue to reach new and wider audiences through reputable religious leaders; and it could enable regional collaboration with big tech companies.

Promoting self-efficacy through media literacy and other social-media campaigns

Self-efficacy as first defined by social psychologist Albert Bandura refers to people's ability to exert control over their motivation, behaviour, and social environment. In a misinformation context, the literature refers to it as a strategy teaching people how to identify and recognise misinformation on social media.

Self-efficacy to detect misinformation was identified as an overarching strategy in all three of our methods: the literature, user behaviour survey and the interviews with fact-checkers.

Using again the public-health analogy to promote the use of safety belts, the final onus rests on the individual car user to change their behaviour.

Equally, it is within the individual information-consumer's control how they choose to respond and act on misinformation. A message such as, 'COVID-19 Misinformation – it's within *your* control' could counter the mental-health implication of the infodemic in that it communicates users' agency in response to health misinformation on digital channels. If campaigns focus on the fact that the spread of health misinformation falls within their personal control, they may feel less anxious and overwhelmed.

Self-efficacy affirms that individual empowerment may be a crucial strategy against health infodemics and one that fact-checkers need to integrate into editorial messaging and social-media campaigns.

Acknowledgements

References (general)

Eysenbach, G. 2002. Infodemiology: The Epidemiology of (Mis)information. *The American Journal of Medicine.*

Hill, R. (2020). Fighting the causes and consequences of bad information, The Full Fact report.

Khuroo, M.S., Sofi, A.A., & Khuroo, M. (2020). Chloroquine and Hydroxychloroquine in Coronavirus Disease 2019 (COVID-19). Facts, Fiction & the Hype. A Critical Appraisal. *International Journal of Antimicrobial Agents*, 106101

Pasquetto, I.V., Jahani, E., Baranovsky, A. & Baum, M.A. (2020). Understanding Misinformation on Mobile Instant Messengers (MIMs) in Developing Countries. *Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School*

Sallam, M., Dababseh, D., Yaseen, A., Al-Haidar, A., Ababneh, N.A., Bakri, F.G., & Mahafzah, A. (2020). Conspiracy beliefs are associated with lower knowledge and higher anxiety levels regarding COVID-19 among students at the University of Jordan. International Journal of Environmental Research and Public Health, 17(14), 4915.

Siddiqui, M.Y.A., Mushtaq, K., Mohamed, M.F., Al Soub, H., Mohamedali, M.G.H., & Yousaf, Z. (2020). "Social Media Misinformation"—An Epidemic within the COVID-19 Pandemic. The American Journal of Tropical Medicine and Hygiene, 103(2), 920.

Wardle, C., & Derakhshan, H. (2018). UNESCO. Journalism, Fake News & Disinformation. Module 2: Thinking about Information Disorder

News & online sources referenced throughout this study:

- UN urges people to #PledgetoPause before sharing information online, UN News, 2020

- UN Chief Targets 'Dangerous Epidemic of Misinformation' on Coronavirus, NPR, 2020

- UN tackles 'infodemic' of misinformation and cybercrime in COVID-19 crisis, UN Department of Global Communications, 2020

- Iran: Over 700 dead after drinking alcohol to cure coronavirus, Al Jazeera, 2020

- Spit On, Yelled At, Attacked: Chinese-Americans Fear for Their Safety, New York Times, 2020

- Researchers find common head lice drug kills coronavirus, The New Daily, 2020

- Can Ivermectin Cure Coronavirus?, Snopes, 2020

- Food & Drug Administration database, Lindane Lotion

- Salicylate Intoxication in an Infant: A Case Report, 2017

- Coronavirus in Africa: Whipping, shooting and snooping, BBC, 2020

- Inside WhatsApp, Instagram and TikTok, a race to build COVID-19 tools, Protocol, 2020

- New WhatsApp chatbot unleashes power of worldwide fact-checking organizations to fight COVID-19 misinformation on the platform, Poynter, 2020

- WHO Health Alert brings COVID-19 facts to billions via WhatsApp, WHO Newsroom, 2020

- WhatsApp Coronavirus Information Hub, Local 2020 – localising the SDGs, 2020

- The CoronaVirusFacts Alliance will get specialized support from the Paris Peace Forum, Poynter, 2020

- Introducing disappearing messages on WhatsApp, WhatsApp blog, 2020

Annexures

A1: Overview of rapid evidence assessment methods

A2: Included studies in each of the three rapid evidence assessments

A3: Survey methods

A4: Survey tool

A5: Interview methods

A6: Interview guide

A7: Ethics board clearance

A8: Methodology & examples of WhatsApp claims for analysis

A1. Overview of rapid evidence assessment methods

Our aim in conducting these rapid reviews is to identify all relevant research to help address our questions. We therefore designed search strategies that were broad and far reaching, and then screened our results against pre-set criteria to decide whether or not to include the studies. While initially we had hoped to drill down only on studies about misinformation on WhatsApp in Africa, we quickly realised that this scope would be too narrow to glean relevant lessons. We therefore focused on public-health misinformation on social media from anywhere in the world.

Note that even as we finalise our report, we have identified two additional potentially relevant studies. These are both listed here for transparency and to enable anyone building on this work to engage further with them:

- Xaudiera, S. & Cardenal A.S. (2020). Ibuprofen Narratives in Five European Countries During the Covid-19 Pandemic. *The Harvard Kennedy School (HKS) Misinformation Review, Volume 1, Special Issue on Covid-19 and Misinformation*

- Barua, Z., Barua, S., Aktar, S., Kabir, N., & Li, M. (2020). Effects of misinformation on Covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science, 8*, 100119. https://doi.org/10.1016/j.pdisas.2020.100119

1. SEARCHING

We had a search strategy to identify relevant studies to be included in this review.

We firstly conducted citation searches of key systematic reviews. We thereafter conducted academic and grey literature searches. We also searched two mega hubs.

Citation search of systematic reviews (Reviews searched)
- A survey on fake news and rumour detection techniques
- A Systematic Review on Fake News Themes Reported in Literature
- Ebola virus disease and social media: A systematic review
- Facade of media and social media during Covid-19: A review
- Fact-checking as risk communication: the multi layered risk of misinformation in times of Covid-19
- Information Overload and Infodemic in the Covid-19 Pandemic.
- What is the impact of misinformation on public health in Africa and around the world? What is the effectiveness of the different approaches to mitigating these impacts?
- Health misinformation in Africa, Latin America and the UK: Impacts and possible solutions
- Public Health and Online Misinformation: Challenges and Recommendations
- Who is most likely to believe and to share misinformation?
- Systematic Literature Review on the Spread of Health-related Misinformation on Social Media

Academic searches

Database searched	Terms used in the search
EbscoHost	• Health misinformation on WhatsApp • Misinformation on Social media • Health misinformation on social media • Strategies to correct misinformation on social media • Countering misinformation on social media • Misinformation and WhatsApp and Covid-19 • How WhatsApp users react to misinformation
UCT Database	• Misinformation on WhatsApp • Health misinformation on WhatsApp • Strategies to counter misinformation on WhatsApp • Why do people share health misinformation? • Effective strategies to counter misinformation
Pubmed	Social media[Title/Abstract] OR Whatsapp[Title/Abstract] OR Instagram[Title/Abstract] OR tiktok[Title/Abstract] OR Twitter[Title/Abstract] OR Facebook)[Title/Abstract] AND "COVID*"[Title/Abstract] OR "coronavirus"[Title/Abstract] OR "corona*"[Title/Abstract] OR "pandemic"[Title/Abstract] OR "epidemic s"[Title/Abstract] OR "epidemical"[Title/Abstract] OR "epidemically"[Title/Abstract] OR "epidemicity"[Title/Abstract] OR "epidemics"[Title/Abstract] OR "epidemics"[Title/Abstract] OR "epidemic"[Title/Abstract] OR "epidemiology"[Title/Abstract] OR "epidemiology"[Title/Abstract]

Grey literature searches

Where the search was conducted	Terms used in the search
Google scholar	• Misinformation on WhatsApp • Misinformation on Social media • Strategies to correct misinformation on social media • Covid-19 misinformation on social media • Health misinformation on social media
Google	• Covid-19 misinformation on WhatsApp • Health misinformation on WhatsApp • Misinformation on Social media • Health misinformation on social media • Strategies to correct misinformation on social media • Countering misinformation on social media • How WhatsApp users react to misinformation
AFP Fact Check	• Misinformation on WhatsApp • Health misinformation
Congo Check	• Misinformation on WhatsApp • Health misinformation
Dubawa	• Misinformation on WhatsApp • Health misinformation
Zimfact	• Misinformation on WhatsApp • Health misinformation
Africa Check	• Misinformation on WhatsApp • Health misinformation
DW Akademie	• Misinformation on WhatsApp • Health misinformation
KCOMNET	• Misinformation on WhatsApp • Health misinformation
WADR	• Misinformation on WhatsApp • Health misinformation
Blackdot Research	• Misinformation on WhatsApp • Health misinformation • WhatsApp
FactCan	• Health misinformation on social media • WhatsApp • Misinformation and WhatsApp and Covid-19 • Health misinformation on social media
Anneberg School of Communication	• Misinformation and WhatsApp and Covid-19. • WhatsApp • Health misinformation on social media

Mega hub searches

Master Hub:
Covid-19 Evidence Network to support Decision-making (COVID_END)

Hub	Terms searched/filers applied
Covid-19+ by McMaster PLUS	Searched under high quality studies & studies currently under review: information, news, media Searched under studies that don't meet critical criteria: information, news, media
Evidence Aid	Searched: misinformation, media, news
L*VE by Epistemonikos	Searched: misinformation, news, media
LitCOVID from PubMed	Searched: misinformation AND review misinformation
TRIP database	Searched: misinformation AND COVID disinformation AND COVID news media AND COVID social media AND COVID
U.S. Veterans' Affairs (VA) Evidence Synthesis Program	Searched: misinformation, social media, media, information
AHRQ EPC Program	Searched: misinformation AND COVID AND review
Campbell Collaboration	Searched: misinformation, social media
Cochrane - Special collections of Cochrane systematic reviews	Searched: misinformation Also screened through their COVID reviews and special collections
JBI	Screened their evidence-based resources for health professionals and evidence-based resources for health organisations
DistillerSR	Used their database of COVID-only studies and searched for misinformation
Literature Review	Searched: misinformation, social media

Master Hub:	
Living Hub of Covid-19 Knowledge Hubs	
Hub	**Terms searched/filers applied**
Hub	Terms searched/filers applied
Coronavirus Research Repository	Misinformation AND COVID AND review = 19 Misinformation AND COVID
Google Scholar	Searched for misinformation impact (predetermined search term). Used the filters 'specific to COVID' 'peer reviewed', 'paper' specific to COVID, peer reviewed, paper
Covid-19 Research Explorer	Asked the question: What is the impact of COVID misinformation?

2. INCLUSION CRITERIA

The final inclusion criteria are comprehensive across each review and summarised in this section of the appendix.

2.1. Inclusion criteria for Review 1

For studies to be included in this review, they had to focus specifically on Covid-19 and on any form of misinformation. We follow a broad definition of misinformation which relates to any messages that conflict with the best-available evidence about Covid-19, and that would likely not be corrected if they were challenged. We included primary research studies with any empirical basis. We also did not restrict studies by geography and included studies from any country and focusing on any subpopulation. Studies had to also focus on the implications of misinformation at individual or community level. We did not develop a priori criteria for the type of harm resultant from misinformation but kept in mind broad categories of disengagement from democracy, interference in democracy, economic harm and risks to life.

2.2. Inclusion criteria for Review 2

The second review considered studies that focused on user behaviours that are related to Covid-19 and user-focused platforms such as WhatsApp, Facebook, Twitter and other social media. It did however excluded YouTube unless the content was shared through any of the included media. Importantly, we considered public-health studies but excluded any content related to elections and political discourse unless they were in relation to public health. We also considered a multitude of various formats that the media could be disseminated as: video, images, voice and text were included.

2.3. Inclusion criteria for Review 3

The third and final review considered studies that focused on user-focused strategies to mitigate public-health misinformation. More specifically, studies that focus on health workers, for instance, were excluded. While our initial focus was on WhatsApp as the platform of interest, a decision was made to consider a more inclusive social media focus

that included applications such as Facebook, Twitter and Instagram. A paucity of evidence on Africa also broadened our interest to an international one. Additionally, all video, voice, text and image formats were considered. For this review, studies that describe mitigation strategies and those that evaluate their effectiveness of were eligible for inclusion.

3. SCREENING

The screening process was two-pronged and reviewed academic and grey literature that met the outlined inclusion criteria. First, studies were screened by title and abstract to determine eligibility on the likelihood of their inclusion. Second, the included studies from the first stage of screening were then filtered to a full text screening process in order to determine final includability. The second stage also included an assignment of the included studies (based on primary focus and objective) to one of each review.

In a thorough and meticulous process, two reviewers were tasked with single screening at both stages. Additionally, a third reviewer then screened the final list of included studies to quality assure the screening process as well as ensure that the identified evidence fits the inclusion criteria for each review.

4. ANALYSIS

To understand the different strategies for countering misinformation present in the literature, we went through all the studies that had strategies, in order to identify themes. From this process we identified nine themes for strategies. These are the themes identified:

- Credible information over misinformation
- Self-efficacy to detect misinformation
- Make misinformation illegal
- Infoveillance
- Technical
- Debunking
- Social media companies tackling misinformation in their platforms
- Collective action against misinformation
- Social media campaigns

To understand the effectiveness of strategies in the literature, we went through all the studies that evaluated the effectiveness of strategies, to identify themes. From this process we identified five impact evaluation themes. These are the themes identified:

- Attitude adjustment
- Behavioural change
- Truth discernment
- Responsiveness to correction
- Psychological outcomes

A2. Included studies from each of the three rapid evidence assessments

Potential studies that we could include which were not identified in time to be analysed within these rapid reviews.

Xaudiera, S. & Cardenal A.S. (2020). Ibuprofen Narratives in Five European Countries During the Covid-19 Pandemic. *The Harvard Kennedy School (HKS) Misinformation Review, Volume 1, Special Issue on Covid-19 and Misinformation.*

Barua, Z., Barua, S., Aktar, S., Kabir, N., & Li, M. (2020). Effects of misinformation on Covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science, 8*, 100119.
https://doi.org/10.1016/j.pdisas.2020.100119

Review 1: Studies included in our rapid evidence assessment of risks and harm of public-health misinformation on social media

Ahinkorah, B.O., Ameyaw E.K., & Hagan, J.E. (2020). Rising above misinformation or fake news in Africa: another strategy to control Covid-19 spread. *Frontiers in Communication*, p. 45.

Ahmad, A.R., & Murad, H.R. (2020). The impact of social media on panic during the Covid-19 pandemic in Iraqi Kurdistan: online questionnaire study. *Journal of Medical Internet Research*, 22(5), pp. e19556.

Ahmed, W., Vidal-Alaball, J., Downing, J. & López Seguí, F. (2020). Covid-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of Medical Internet Research*, 22(5), pp. e19458.

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2020). Health-protective behaviour, social media usage and conspiracy belief during the Covid-19 public health emergency. *Psychological Medicine*, pp. 1-7.

Bastani, P., & Bahrami, M.A. (2020). Covid-19 related misinformation on social media: a qualitative study from Iran. *Journal of Medical Internet Research*, 10.2196/18932. Advance online publication. https://doi.org/10.2196/18932

Bendau, A., Petzold, M.B., Pyrkosch, L., Mascarell Maricic, L., Betzler, F., Rogoll, J., Große, J., Ströhle, A., & Plag, J. (2020). Associations between Covid-19 related media consumption and symptoms of anxiety, depression and Covid-19 related fear in the general population in Germany. *European Archives of Psychiatry and Clinical Neuroscience*, pp. 1-9.

Budhwani, H. & Sun, R. (2020). Creating Covid-19 stigma by referencing the novel coronavirus as the "Chinese virus" on Twitter: quantitative analysis of social media data. *Journal of Medical Internet Research*, 22(5), pp. e19301.

Duffy, B., & Allington, D. (2020). Covid conspiracies and confusions: the impact on compliance with the UK's lockdown rules and the link with social media use, https://www.kcl.ac.uk/policy-institute/assets/covid-conspiracies-and-confusions.pdf

Freeman, D., Waite, F., Rosebrock, L., Petit, A., Causier, C., East, A., Jenner, L., Teale, A.L., Carr, L., Bold, L.M., & Lambe, S. (2020). Coronavirus conspiracy beliefs, mistrust, and compliance with government guidelines in England. *Psychological Medicine*, pp. 1-30.

Gao, J., Zheng, P., Yingnan, J., Chen, H., Mao, Y., Chen, S., Wang, Y., Fu, H., & Dai, J. (2020). Mental health problems and social media exposure during Covid-19 outbreak. *PLoS ONE*, pp. e0231924.

Georgiou, N., Delfabbro, P., & Balzan, R. (2020). Covid-19-related conspiracy beliefs and their relationship with perceived stress and pre-existing conspiracy beliefs. *Personality and Individual Differences*, 166, pp. 110201.

Kawchuk, G., Hartvigsen, J., Harsted, S., Nim, C.G., & Nyirö, L. (2020). Misinformation about spinal manipulation and boosting immunity: an analysis of Twitter activity during the Covid-19 crisis. *Chiropractic & Manual Therapies*, pp. 1-13.

Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E.W., & Baddour, K. (2020). Coronavirus goes viral: quantifying the Covid-19 misinformation epidemic on Twitter. *Cureus*, 12(3), pp. e7255.

Kulkarni, P., Prabhu, S., Dumar, S.D., & Ramraj, B. (2020). Covid-19-infodemic overtaking pandemic? Time to disseminate facts over fear. *Indian Journal of Community Health*, pp. 264-268.

Michael, C., Overbeek, D., Papadimoulis, A., Sheroff, A., & Burns, M. (2020). Geospatial correlation between Covid-19 health misinformation and poisoning with household cleaners in the Greater Boston Area. *Clinical Toxicology*, pp. 1-6.

Moukarzel, S., Del Fresno, M., Bode, L., & Daly, A.J. (2020). Distance, diffusion and the role of social media in a time of COVID contagion. *Maternal & Child Nutrition*, pp. e13025.

Naeem, S.B., Bhatti, R., & Khan, A. (2020). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information and Libraries Journal*, 10.1111/hir.12320. Advance online publication. https://doi.org/10.1111/hir.12320

Roy, M., Moreau, N., Rousseau, C., Mercier, A., Wilson, A., & Atlani-Duault, L. (2020). Ebola and localised blame on social media: analysis of Twitter and Facebook conversations during the 2014-2015 Ebola epidemic. *Culture, and Medicine and Psychiatry*, 44(1), pp. 56-79.

Sahoo, S.S., Sahu, D.P., & Kankaria, A. (2020). Mis-infodemic: the achilles' heel in combating the Covid-19 pandemic in an Indian perspective. *Monaldi Archives for Chest Disease*, 90(2). https://doi.org/10.4081/monaldi.2020.1405

Soltaninejad, K. (2020). Methanol mass poisoning outbreak: a consequence of Covid-19 pandemic and misleading messages on social media. *International Journal of Occupational and Environmental Medicine*, 11(3), pp. 148-150.

Watt, P.S., & Sala, S.D. (2017). Misinformation lingers in memory: failure of three pro-vaccination strategies. *PLoS ONE*, pp. e0181640.

Review 2: Studies included in our rapid evidence assessment of social media users' responses to public health misinformation

Adebimpe, W.O., Adeyemi, D.H., Faremi, A., Ojo, J.O., & Efuntoye, A.E. (2015). The relevance of the social networking media in Ebola virus disease prevention and control in Southwestern Nigeria. *The Pan African Medical Journal*, pp. 22.

Ahmed, D.B., & Hadiza, U.M. (2019). Misinformation on salt water use among Nigerians during 2014 Ebola outbreak and the role of social media. *Asian Pacific Journal of Tropical Medicine*, pp. 175.

Arkaitz, Z., Wong, G., Liakata, M., & Procter, R. (2015). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS One*, pp. e0150989.

Bessi, A., Coletto, M., Davidescu, G.A., & Scala, A. (2015). Science vs conspiracy: collective narratives in the age of misinformation. *PLoS ONE*, pp. e0118093.

Chua, A.Y.K., & Banerjee, S. (2017). To share or not to share: The role of epistemic belief in online health rumours. *International Journal of Medical Informatics*, pp. 36-41.

Colliander, J. (2019). "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behaviour*, pp. 202-215.

Cronkhite, A., Zhang, W., & Caughell, L. (2020). #FakeNews in NatSec: handling misinformation. *Parameters*, pp. 5-22.

Huang, B., & Carley, K.M. (2020). Disinformation and misinformation on Twitter during the novel coronavirus outbreak. *arXiv preprint arXiv*, pp. 2006.04278.

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

Khan, M.L., & Idris, I.K. (2019). Recognise misinformation and verify before sharing: a reasoned action and information literacy perspective. *Behaviour & Information Technology*, pp. 1194-1212.

Kulkarni, P., Prabhu, S., Dumar S.D., & Ramraj, B. (2020). COVID-19-infodemic overtaking pandemic? Time to disseminate facts over fear. *Indian Journal of Community Health*, pp. 264-268.

Pasquetto, I., Jahani, E., Baranovsky, A., & Baum, M. (2020). Understanding misinformation on mobile instant messengers (MIMs) in developing countries, pp. 1-27.

Samuli, L., Islam, A.K.M., Muhammad, N.I., & Eoin, W. (2020). What drives unverified information sharing and cyberchondria during the Covid-19 pandemic? *European Journal of Information Systems*, pp. 1-18.

Talwar, S., Dhir, A., Kaur, P., Zafar, N., & Alrasheedy, M. (2019). Why do people share fake news? associations between the dark side of social. *Journal of Retailing and Consumer Services*, pp. 72-82.

Tandoc, E.C., Lim, D., & Ling, R. (2020). Diffusion of disinformation: how social media users respond to fake news and why. *Journalism*, pp. 381-398.

Tiago, M., Tiago, F., & Cosme, C. (2019). Exploring users' motivations to participate in viral communication on social media. *Journal of Business Research*, pp. 574-582.

Wasserman, H., Madrid-Morales, D., Mare, A., Ndlovu, K., Tully, M., & Umejei, E. (2019). Audience motivations for sharing dis- and misinformation: a comparative study in five sub-Saharan African countries, https://cyber.harvard.edu/sites/default/files/2019-12/%20 Audience%20Motivations%20for%20Sharing%20Dis-%20and%20Misinformation.pdf

Zollo, F., Novak, P.K., Del Vicario, M., & Bessi, A. (2015). Emotional dynamics in the age of misinformation. *PLoS ONE*, pp. e0138740.

Review 3: Studies included in our rapid evidence assessment of strategies to mitigate public health misinformation on social media

Abbasi, M., & Liu, H. (2013). Measuring user credibility in social media. *Social Computing, Behavioural-Cultural Modelling, and Prediction*, pp. 441-448.

Africa Check & Volume. (2020). Case Study: "What's Crap on WhatsApp?"

Ahinkorah, B.O., Ameyaw, E.K., & Hagan, J.E. (2020). Rising above misinformation or fake news in Africa: another strategy to control Covid-19 spread. *Frontiers in Communication*, pp. 45.

Al-Dmour, H., Masa'deh, R., Salman, A., Abuhashesh, M., & Al-Dmour, R. (2020). Influence of social-media platforms on public health protection against the Covid-19 pandemic via the mediating effects of public health awareness and behavioural changes: integrated model. *Journal of Medical Internet Research*, 22(8), e19996. https://doi.org/10.2196/19996

Aldwairi, M., & Alwahedi, A. (2018). Detecting fake news in social media networks. *Procedia Computer Science*, pp. 215-222.

Alonso, P., & Alemañy-Castilla, C. (2020). Curbing misinformation and disinformation in the Covid-19 era: a view from Cuba. *MEDICC Review*, pp. 45-46.

Alvarez, A., Mejia, C., Delgado-Zegarra, J., Del-Aguila-Arcentales, S., Arce-Esquivel, A.A., Valladares-Garrido, M.J., Portal, M.R., Villegas, L., Curioso, W., Sekar, C., & Yañez, J.A. (2020). The Peru approach against the Covid-19 infodemic: insights and strategies. *The American Journal of Tropical Medicine and Hygiene*, pp. Tpmd200536.

Anderson, K.E. (2018). Getting acquainted with social networks and apps: combating fake news on social media. *Library Hi Tech News*, 35(3), pp. 1-6. Retrieved from https://doi.org/doi:10.7282/T32J6GGK

Armitage, L., Lawson, B.K., Whelan, M.E., & Newhouse, N. (2020). Paying special consideration to the digital sharing of information during the Covid-19 pandemic and beyond. BJGP open, 4(2), pp. 1-4.

Baeth, M.J., & Aktas, M.S. (2019). Detecting misinformation in social networks using provenance data. *Concurrency and Computation: Practice and Experience*, pp. E4793.

Bode, L., & Vraga, E.K. (2018). See something, say something: correction of global health misinformation on social media. *Health communication*, pp. 1131-1140.

Bowles, J., Larreguy, H., & Liu, S. (2020). Countering misinformation via WhatsApp: evidence from the Covid-19 pandemic in Zimbabwe.

Cronkhite, A., Zhang, W., & Caughell, L. (2020). #FakeNews in NatSec: Handling misinformation. *Parameters*, pp. 5-22.

De Keersmaecker, J., & Roets, A. (2017). 'Fake news': incorrect, but hard to correct: The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, pp. 107-110.

Eysenbach, G. (2020). How to fight an infodemic: the four pillars of infodemic management. *Journal of Medical Internet Research*, 22(6), pp. e21820.

Gesser-Edelsburg, A., Diamant, A., Hijazi, R., & Mesch, G.S. (2018). Correcting misinformation by health organisations during measles outbreaks: a controlled experiment. *PloS One*, 13(12), pp. E0209505.

Iosifidis, P., & Nicoli, N. (2019). The battle to end fake news: a qualitative content analysis of Facebook announcements on how it combats disinformation. *International Communication Gazette*, pp. 60-81.

Kabha, R., Kamel, A., Elbahi, M., & Narula, S. (2019). Comparison Study between the UAE, the UK, and India in dealing with WhatsApp fake news. *Journal of Content, and Community and Communication*, pp. 176-186.

Khan, M.L., & Idris, I.K. (2019). Recognise misinformation and verify before sharing: a reasoned action and information literacy perspective. *Behaviour & Information Technology*, pp. 1194-1212.

Krishna Kumar, K.P., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, pp. 1-22.

Kulkarni, P., Prabhu, S., Dumar, S.D., & Ramraj, B. (2020). Covid-19-infodemic overtaking pandemic? Time to disseminate facts over fear. *Indian Journal of Community Health*, pp. 264-268.

Liang, H., Chun-Hai Fung, I., Ho Tse, Z., Yin, J., Chan, C., & Pechta, L.E. (2019). How did Ebola information spread on Twitter: broadcasting or viral spreading? *BMC Public Health*, pp. 438.

Michael, C., Overbeek, D., Papadimoulis, A., Sheroff, A., & Burns, M. (2020). Geospatial correlation between Covid-19 health misinformation and poisoning with household cleaners in the Greater Boston Area. *Clinical Toxicology*, pp. 1-6.

Morley, J., Cowls, J., Taddeo, M., & Floridi, L. (2020). Public health in the information age: recognising the infosphere as a social determinant of health. *Journal of Medical Internet Research*, 22(8), pp. e19311.

Naeem, S.B., Bhatti, R., & Khan, A. (2020). An exploration of how fake news is taking over social media and putting public health at risk. *Health Information and Libraries Journal*. pp.

Ozturk, P., Li, H., & Sakamoto, Y. (2015). Combating rumour spread on social media: the effectiveness of refutation and warning. In *2015 48th Hawaii International Conference on System Sciences* 2015, pp. 2406-2414.

Pasquetto, I., Jahani, E., Baranovsky, A., & Baum, M. (2020). Understanding misinformation on mobile instant messengers (MIMs) in developing countries, pp. 1-27.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., & Rand, D.G. (2020). Fighting Covid-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, pp. 770-780.

Pulido, C.M., Ruiz-Eugenio, L., Redondo-Sama, G., & Villarejo-Carballido, B. (2020). A new application of social impact in social media for overcoming fake news in health. *International Journal of Environmental Research and Public Health*, pp. 2430.

Sahoo, S.S., Sahu, D.P., & Kankaria, A. (2020). Mis-infodemic: the Achilles' heel in combating the Covid-19 pandemic in an Indian perspective. *Monaldi Archives for Chest Disease*, 90(1405), pp. 330-332

Sang, H., Lee, S., & Han, C. (2020). The effects of social media use on preventive behaviours during infectious disease outbreaks: the mediating role of self-relevant emotions and public risk perception. Health Communication, pp. 1-10.

Sojung, C.K., & John, C. (2020). An eye tracking approach to understanding misinformation and correction strategies on social media: the mediating role of attention and credibility to reduce HPV vaccine misperceptions. *Health Communication*, pp. 1-10.

Soltaninejad, K. (2020). Methanol mass poisoning outbreak: a consequence of Covid-19 pandemic and misleading messages on social media. *The International Journal of Occupational and Environmental Medicine*, 11(3), pp. 148-150.

Stephen, B.L. (2020). Fact or fiction? misinformation and social media in the era of Covid-19: an infodemiology study.

Swire-Thompson, B., & Lazer, D. (2019). Public health and online misinformation: challenges and recommendations. *Annual Review of Public Health*, pp. 433-51.

Swire-Thompson, B., Ecker, U.K., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12, pp. 1948.

Van der Meer, T., & Jin, Y. (2019). Seeking formula for misinformation treatment in public health crises: the effects of corrective information type and source. *Health Communication*, pp. 560-575.

Vraga, E., & Bode, L. (2017a). I do not believe you: how providing a source corrects health misperceptions across social-media platforms. *Information, and Communication & Society*, pp. 1337-1353.

Vraga, E.K., & Bode, L. (2017b). Using expert sources to correct health misinformation in social media: science communication. *Science Communication*, pp. 621-645.

Watt, P.S., & Sala, S.D. (2017). Misinformation lingers in memory: failure of three pro-vaccination strategies. *PLoS One*, pp.e0181640.

Zollo, F. (2019). Dealing with digital misinformation: a polarised context of narratives and tribes. *EFSA Journal*, pp. e170720.

A3. Survey methods

Data Collection

In order to understand WhatsApp user responses to Covid-19 Misinformation we designed an electronic self-administered questionnaire survey on google forms. We made use of both closed and open-ended questions, incorporating multiple and single responses to understand responded engagement with Covid-19 Information and misinformation on WhatsApp.

The survey could be completed either in English or French. No personally identifying information was collected. All respondents provided informed consent before completing the survey.

The survey questionnaire was piloted in both English and French and was approved by the Research Ethics Committee at the Faculty of Humanities at the University of Johannesburg. It was 'live' from 25 August 2020 to 25 September 2020, and promoted through a range of social-media platforms and groups.

Analysis

Survey data was cleaned by removing duplicates surveys, checked for completion and de identified using unique survey IDs.

We used both qualitative and quantitative techniques to analyse survey data. descriptive analysis being preferred for analysing closed questions. Frequency distribution was used to ascertain common trends and themes.

The open-ended questions were analysed using a framework analysis and inductive reasoning to make sense of the feedback we had received through the survey. We derived themes from the quotes. Themes with more than three quotes were approved. The development of this framework necessitated numerous attempts to review the themes and quotes to ensure they made analytical sense.

A4. Survey tool

Covid-19 information on WhatsApp

We want to know more about how people respond to WhatsApp messages about Covid-19 (also known as 'the coronavirus' or 'the pandemic'). Responding to this survey should take you no more than 10 minutes.

Before you get started, please read the information below about this survey.

To take part in the survey you must be 18 years or older, a WhatsApp user, and live in Africa.

This survey is anonymous and your identity and details will not be used for any other purposes outside this study.

By submitting your answers to this survey, you are giving consent for your answers to be included in our research.

You do not have to answer any question that you feel uncomfortable with.

Your participation is voluntary. You are free to decide if you want to take part in the research. You can refuse to participate, or stop at any time without giving any reason.

There is no direct benefit to you. There will be no payment to participate in the survey.

We may quote what you tell us but you will not be identified.

The survey is being conducted by the Africa Centre for Evidence at the University of Johannesburg in partnership with Africa Check. It has been approved by the University of Johannesburg's Humanities Research Ethics Committee.

If you want to know about the researchers conducting this survey, click here: www.africacentreforevidence.org

1. What country do you live in? (Note that we are focusing on Kenya, Nigeria, Senegal and South Africa, but welcome answers from anywhere in Africa)

   ☐ Kenya

   ☐ Nigeria

   ☐ Senegal

   ☐ South Africa

   ☐ Other

2. Have you ever received information about Covid-19 via WhatsApp?

   ☐ Yes

   ☐ No

3. Do you belong to any Covid-19 specific groups on WhatsApp?

   ☐ Yes

   ☐ No

Covid-19 information on WhatsApp

How do you respond to Covid-19 messages on WhatsApp?

Different people respond in different ways to Covid-19 messages on WhatsApp. We want to know more about how you respond.

4. Think about WhatsApp messages about Covid-19 that you have received over the last month and tell us how you have responded (tick all that apply)

☐ I have forwarded a message to individual contacts

☐ I have forwarded a message to one or more WhatsApp groups

☐ I have asked the sender of the message about its accuracy

☐ I have deleted a message because I thought it was false

☐ I have reported a message (for example, via Africa Checks' misinformation line)

☐ I have acted on the information changing my behaviour

☐ I have done nothing

5. If you ticked 'I have forwarded a message' (to either individuals or a group), please explain why you forwarded it.

6. If you ticked 'I have acted on the information changing my behaviour', please explain what you did differently in response to the information.

What shapes how you respond to Covid-19 messages on WhatsApp?

Different people respond in different ways to Covid-19 messages on WhatsApp. The following questions ask about what influences your responses.

7. Which of the following influence how you respond to a Covid-19 message on WhatsApp? (tick all that apply)

☐ My response depends on which person sends it to me

☐ My response depends on whether it comes from an organisation that I have heard of

☐ My response depends on whether it gives good news or bad

☐ My response depends on whether I think the message is true or not

☐ My response depends on which language the message is in

☐ My response depends on what format it takes (for example, text or image)

☐ My response depends on other factors (please give details below)

8. If you ticked 'my response depends on whether I think the message is true or not', please explain how this influences your response:

9. If you ticked 'my response depends on other factors', please explain which other factors influence your response:

10. Which of the following makes you trust the message? (tick all that apply)

    ☐ I trust a message when it comes from someone whom I respect

    ☐ I trust a message when the message comes from a legitimate news source

    ☐ I trust a message when it comes from an organisation that I know

    ☐ I trust a message when it comes from a government source (for example, Ministry of Health)

11. If you ticked 'I trust a message when it comes from an organisation that I know', please list the organisations that you trust:

12. Which of the following do you trust the most?

    ☐ Messages about Covid-19 in my home language which is not English or French

    ☐ Messages about Covid-19 in French

    ☐ Messages about Covid-19 in English

13. Which of the following message formats do you trust the most? (please tick all that apply)

    ☐ Text

    ☐ Image/Picture

    ☐ Voice

    ☐ Video

    ☐ Link to an article

14. Have you ever received a message that you thought was false or inaccurate?

    ☐ Yes

    ☐ No

    ☐ Maybe

15. What makes you suspicious that a message about Covid-19 is false or inaccurate?

    ☐ Spelling mistakes

    ☐ Too good to be true

    ☐ I cannot tell the origins of the message

    ☐ Has a double arrow which tells me it has been forwarded many times

    ☐ The sender of the message is someone who often forwards false or inaccurate messages

    ☐ The message is from an organisation I have never heard of

    ☐ If the message has a link the seems strange

    ☐ If the aim of the message seems to be to get me to click on a link to a website

16. Have you ever done any of the following to check whether a message about Covid-19 was true or false? (tick all that apply)

    ☐ Googled for more information

    ☐ Asked someone I trust

    ☐ Looked up a fact checking service or website

17. If you concluded that the message was false, what did you do? (tick all that apply)

    ☐ I ignore it

    ☐ I delete it

    ☐ I told the person who sent it to me that it was false

    ☐ I reported it to a fact checking organisation

    ☐ I wanted to report it but I did not know how

18. If you said 'I reported it', please explain more about the message you reported and why you decided to report it.

19. If you have reported a message about Covid-19 for being false, how did you report it?

    ☐ I told the sender that the message was false

    ☐ I told my cellphone provider

    ☐ I reported it to my health ministry

    ☐ I reported it to a fact checking organisation (for example, Africa Check)

20. Please use this space to tell us anything else you want to share about Covid-19 messages on WhatsApp.

Please tell us a bit more about yourself

Your answers to the following questions will help us analyse the information you have shared with us.

21. How old are you?

    ☐ 18-20

    ☐ 21-30

    ☐ 31-40

    ☐ 41-50

    ☐ 51-60

    ☐ 61 or older

22. What gender are you?

☐ Female

☐ Male

☐ Prefer not to say

23. What level of formal education did you reach?

☐ I did not finish primary school

☐ I finished primary school

☐ I started high school but I did not finish

☐ I finished high school

☐ I attended further education after high school

☐ I prefer not to say

24. Have you, or a member of your family, suffered from Covid-19?

☐ Yes

☐ No

☐ I don't know

☐ Prefer not to say

We do not have any more questions for you, but if there is anything that you would like to share about any of the questions covered above; or that you think would be important for us to know, please explain here.

A5. Interview methods

The Interviewees came recommended by Africa Check, which has had a working relationship with fact-checking organisations in Africa and beyond through the Africa Facts Network. In addition, at the start of the project, Africa Check and the Africa Centre for Evidence hosted a stakeholder meeting that included these fact-checking organisations. The outcome of this meeting had the team realise the importance of conducting interviews with fact-checkers from these organisations – this was not initially part of our project plan. Ultimately, we wanted to find out how African fact-checking organisations can respond with effective mitigating strategies in as far as misinformation on WhatsApp.

The study anticipated having eight interviews and we ended up conducting nine. The nine interviewees were from nine fact-checking organisations working in Africa. We conducted and recorded the interviews using an online platform (Zoom), and notes were also taken.

We used a framework analysis approach to make sense of the data collected from the fact-checkers. With the open-ended questions, we came up with themes based on the responses we received. Thereafter, the quotes from interview respondents were assigned to the corresponding theme. The themes were approved where more than two quotes had been allocated. We interrogated the data further in order to draw lessons to inform our research questions.

A6. Interview Guide

Covid-19 information on WhatsApp and Fact-checking organisations

ID Information

Unique ID

Your Organisation and Fact-Checking Covid-19 (Mis) Information

Thank you for consenting to be part of this interview that will explore how fact-checking organisations have been dealing with (mis)information around Covid-19. The study is being conducted by the Africa Centre for Evidence at the University of Johannesburg in partnership with Africa Check. It has been approved by the University of Johannesburg's Humanities Research Ethics Committee. The interview process will gather information on how fact-checking organisations in Africa are dealing with misinformation on WhatsApp (including strategies and challenges). We hope that the data will inform initiatives that shape effective fact checking strategies to help counter the impacts of Covid-19 on WhatsApp in Africa and further afield. We will document and share the results with the broader community of fact-checking organisations and other relevant stakeholders.

If you want to know about the researchers conducting this survey, click here:
www.africacentreforevidence.org

This interview is expected to take 30-45 minutes.

We will begin with obtaining an understanding of how your organisation is dealing with (mis) information around Covid-19.

1. What is the name of your organisation?

2. Is your organisation involved in fact checking information from various sources about Covid-19, including those shared through WhatsApp?

    ☐ Yes

    ☐ No

3. How does your organisation collate/ receive (mis)information, rumours, concerns, etc. about Covid-19 on WhatsApp?

4. What do you think were some of the WhatsApp Covid-19- related harmful claims that were shared with your organisation for fact-checking? Please give examples.

5. What systems have you put in place to track misinformation on WhatsApp?

6. What are the current strategies within your organisations for fact-checking content on WhatsApp?

7. How do you disseminate feedback to target audiences based on what has been verified-relating to WhatsApp content on Covid-19?

8. What are the challenges that your organisation faces in fact-checking content on WhatsApp?

9. What would make it easier for you to do fact checking on WhatsApp more effectively?

10. What else do you think can be done to counter/tackle misinformation on WhatsApp?

11. In which country is your organisation based?

    ☐ Kenya

    ☐ Nigeria

    ☐ Senegal

    ☐ South Africa

    ☐ Other

12. In which country (ies) does the organisation operate? (check all that apply)

13. What is your self-assessment of how long your organisation has been working in the field of Fact-Checking?

    ☐ New to this work (0-3 years)

    ☐ Have some experience (3-8 years)

    ☐ Established (8-15 years)

    ☐ Well established (15+ years)

Respondent Overview

The following questions will endeavour to get an understanding of the interviewee's leadership aspects within the organisation.

14. What is your role in the fact-checking organisation that you are representing in this interview?

Documentation

15. I would be very interested in reading and learning more about the organisation and the fact-checking work that you do. Would you be willing to share key documents related to these with me? [I intend to use this as a means to collect documents for documentary analysis in as far as organisations and fact-checking is concerned]
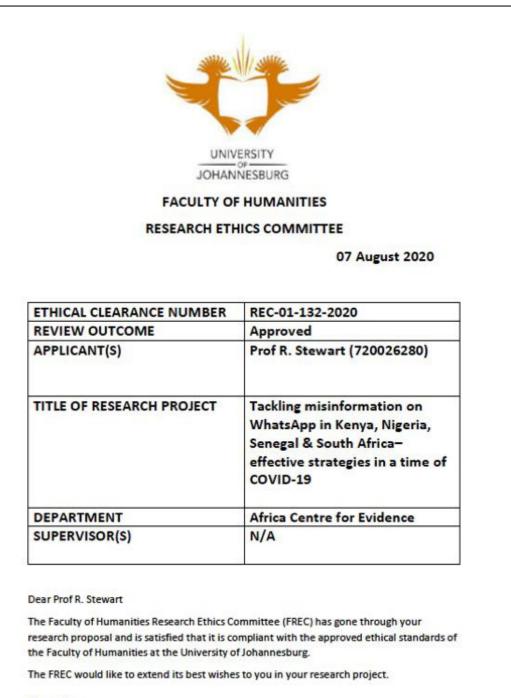
Closure

16. I do not have any more questions for you but is there anything that you would like to share with me about any of the topics we covered above or that you think would be important for me to know?

17. Do you have any questions for me?

Interviewers Reflections

Reflections of interviewer on the interview

Annexures

A7. Ethics approval letter

UNIVERSITY OF JOHANNESBURG

FACULTY OF HUMANITIES

RESEARCH ETHICS COMMITTEE

07 August 2020

ETHICAL CLEARANCE NUMBER	REC-01-132-2020
REVIEW OUTCOME	Approved
APPLICANT(S)	Prof R. Stewart (720026280)
TITLE OF RESEARCH PROJECT	Tackling misinformation on WhatsApp in Kenya, Nigeria, Senegal & South Africa– effective strategies in a time of COVID-19
DEPARTMENT	Africa Centre for Evidence
SUPERVISOR(S)	N/A

Dear Prof R. Stewart

The Faculty of Humanities Research Ethics Committee (FREC) has gone through your research proposal and is satisfied that it is compliant with the approved ethical standards of the Faculty of Humanities at the University of Johannesburg.

The FREC would like to extend its best wishes to you in your research project.

Sincerely,

Prof Tapiwa Chagonda
Interim Chair: Faculty of Humanities REC
Tel: 011 559 3827
E-mail: tchagonda@uj.ac.za

A8. Methodology & examples of WhatsApp claims for analysis

Methodology

In 2019, Africa Check and Volume launched What's Crap on WhatsApp – a fact-checking voice-note show distributed on WhatsApp itself. Through this WhatsApp line, Africa Check was able to solicit and catalogue claim submissions from audiences in South Africa, offering a first window into the nature of misinformation circulating on the platform. In April 2020, Africa Check extended this process by recording all the claims and requests that WhatsApp subscribers sent on five different WhatsApp subscription lines: two in South Africa, one in Nigeria, one in Kenya and one in Senegal, with the hope that it could shed more light on the nature and associated risks of Covid-19 misinformation.

All claims were recorded in a cloud-based spreadsheet in the four different countries. The Africa Check teams in South Africa, Nigeria, Kenya and Senegal recorded 283 suspicious claims reported to them over the three-month April to June period. Out of these, 222 claims were fit for analysis according to their risks (the 61 claims not selected for analysis were either questions, or not Covid-19 related).

Of the 222 claims, 60% (n=133) were from South Africa; 19% (n=42) were from Kenya; 17.5% (n=39) were from Nigeria; and 3.5% (n=8) were from Senegal.

The claims were analysed and discussed using the established risk framework *(Table 2: An evidence-informed risk framework of health misinformation)*. They were categorised in four categories: Harm to physical health; Economic harm; Social harm and Political harm.

Loosely categorised, 30% (n=67) of claims were associated with harm to physical health; 18% (n=39) of claims fell in the sphere of economic harm; 19% (n=42) of claims were categorised in the sphere of social harm; and 33% (n=74) of claims analysed were associated with political harm.

There is a complex overlap in categorising the risks associated with misinformation. Each claim was categorised according to the sphere where it was perceived to cause the most harm.

Harm to physical health

- Here are examples of the series of home remedies and herbal 'cures' were flagged on the different WhatsApp lines":

  Good News: Nigerian Finally Gets Coronavirus Herbal Cure

  This was discovered by a Nigerian living in Germany and married to a German medical doctor. Many doctors including his wife were infected by a virus which has all the symptoms of the much dreaded corona virus. All treatments using orthodox approach were ineffective. He decided to use his Nigerian experience. All those who took the mixture were cured completely. Here is the mixture:1- Pineapple peels 2-Lime 3-Ginger.

And:

> *This is one way to prevent Covid19 from spreading - If you can drink as much warm water as possible or gargle with salt water*

Or in this message, someone's brother from Hong Kong, with his own 'sources of info from China' gives the following advice:

> *Use steam inhalation everyday - morning and evening if possible, gargle with warm water and salt daily - morning and evening if possible. This is a good daily practice which will improve your overall well-being and can continue even after COVID19 ends.*

In one video 'undeniable proof' was found that "corona is the common cold".

- This is the example of the claim that could easily lead to behaviour in which people increase their aspirin dosage when presenting with Covid-19 symptoms. It was sent to us by 24 different users:

> *A Mexican family in the United States were cured with a home remedy was documented: three 500 mg aspirins dissolved in lemon juice boiled with honey, taken hot. The next day they woke up as if nothing had happened to them! Well, the scientific information that follows proves they are right!. According to valuable information from Italian pathologists, ventilators and intensive care units are not necessary.*

- Example of the claim detected with the exact same wording in Kenya and Nigeria claims that Covid-19 is a bacterial infection:

> *IN ITALY THE CURE FOR THE CORONAVIRUS IS FINALLY FOUND.*

> *Italian doctors disobeyed the world health law WHO, not to do autopsies on the dead of the Coronavirus and they found that it is NOT a VIRUS but a BACTERIA that causes death. This causes blood clots to form and causes the death of the patient.*

- Example of a claim showing how the use of medical terms creates a sense of pseudo-accuracy, that could mislead users into believing it's credible information:

> *Autopsies Prove that Covid-19 is a Disseminated Intravascular Coagulation (Pulmonary Thrombosis)*

> *It is now clear that the whole world has been attacking the so-called Coronavirus Pandemic wrongly due to a serious pathophysiological diagnosis error.*

> *According to valuable information from Italian pathologists, ventilators and intensive care units were never needed.*

> *Autopsies performed by the Italian pathologists have shown that it is not pneumonia but it is Disseminated Intravascular Coagulation (Thrombosis) which ought to be fought with antibiotics, antivirals, anti-inflammatories and anticoagulants.T*

- Examples of the helpful, friendly tone in some of the messages below:

  Hey everybody, as an infectious disease epidemiologist, at this point I feel morally obligated...

  My advice is as follows: Wherever you are (markets, hospitals, etc ...) if they distribute the masks, please do not take them.

  Please my fellow Nigerians. Stay away from all the CoronaVirus kits sent from China government

  Military Assistance Corps released this information for ur safety. Please pass this message to your family and friends NOW.

  My daughter. 19 yrs old. Healthy. Frontline worker at a huge grocery store chain...

- Some of the messages recorded contained hybrid advice of half-fact and half-fiction, which only makes it more difficult for users to discern accuracy, for example:

  Heating kills coronavirus

  Heat is very effective at sanitising and disinfecting objects from coronavirus. If anyone tells you that coronavirus is resistant to heat, they're wrong. You should not soak N95 or surgical masks in disinfectants or other liquids as this can compromise their integrity and fit. Note that washing cloth face masks in a washing machine should suffice to disinfect them, according to the CDC. The agency also cautions not to touch your eyes, nose, and mouth when removing cloth coverings, and to and wash hands immediately after removing.

- An interesting trend in Nigeria was how three classic conspiracies (the whistle-blower Chinese doctor, coronavirus being a bio-weapon, and the Madagascar fabrication) were used as a pretext to promote natural remedies like fever grass or the inhalation of hot water vapour:

  Breaking News from CNN :- #CORONA_VIRUS_IS_DEAD:
  Dr. Li Wenliang, China's hero doctor who was punished for telling the truth about Corona Virus and later died due to the same disease, had documented case files for research purposes and had in the case files proposed a cure that would significantly decrease the impact of the COVID - 19 Virus on the human body. The chemical Methylxanthine, Theobromine and Theophylline stimulate compounds that can ward off these virus in a human with at least an average immune system. Whats more shocking is that these complex words that were so difficult for people in China to understand is actually called Fever Grass Tea(African Lemon grass tea), YES, our Fever Grass Tea has all these chemicals already in it. Who would have known that all the solution to these virus would be a simple cup of FEVER GRASS TEA. and that is the reason so many patients in China are being cured.

and:

> *Finally, the FBI arrested a professor at Boston University who was in contact with the Chinese University and the Wuhan research laboratory, and who was very well paid by China .... Now, he is very clear that the corona virus is a bio-attack planned and led by China.*
> *- A Chinese expert assures everyone that inhaling hot water vapour kills the Corona virus 100%. Even if the virus has entered the nose, throat or lungs. The Corona virus does not support hot water vapour ...*
> *PLEASE DISTRIBUTE THIS INFORMATION TO ALL OF YOUR FAMILY MEMBERS AND FRIENDS.*

and:

> *US to fund Madagascar Covid-19 herbal cure with $2.5m.*
> *The plant has been proven effective in cure for covid-19.*
>
> *What is it called in your language?*
> *English - Artemisia*
> *Yoruba - Ewe Egbin*
> *Zulu - Umhlonyane*
> *Ibibio - Mkpatat*
> *Hausa - Tazargade (Baaba)* https://t.co/A6GKnKd7Rn

- Examples of messages that discourage good health-seeking behaviour:

> *DANGER OF FACEMASK:Mask is supposed to be used for a limited time. If you wear it for a long time: Oxygen in the blood reduces.Oxygen to the brain reduces. You start feeling weak. May lead to death.*

Or this text message recorded in South Africa from "Dr. Dennis A Castro B":

> *Wearing a mask for prolonged periods creates hypoxia. Breathing over and over exhaled air turns into carbon dioxide, which is why we feel dizzy. This intoxicates the user and much more when he must move, carry out displacement actions. It causes discomfort, loss of reflexes and conscious thought.*

Or an out-of-context photo with the description:

> *Joggers (including children jogging for gym) had been jogging with masks on and all of them had died due to the lack of oxygen because of the mask*

- There are also several claims about masks and test kits being infected with Covid-19 such as the two below text claims from Nigeria:

> *China supplies 2 million masks for Africa.*
> *The most suspicious is that the WHO says that Africa must prepare for the worst.*
> *My advice is as follows: Wherever you are (markets, hospitals, etc ...) if they distribute the masks, please do not take them.*
> *These are infected brands.*

Save lives.
Share to everyone..... DO NOT ACCEPT FREE MASKS FROM ANYONE, NOT
EVEN FROM YOUR OFFICE.

In Spain: over 640 000 #Covid19 testing kits from China didnt work. Testing
kits from China tested positive for coronavirus Czech, Ukraine and Turkey.
Netherlands returned 600 000 mask from China. China proudly killing the
world. Please my fellow Nigerians. Stay away from all the CoronaVirus kits
sent from China government through their agent called Jack Ma. Those kits
from China are the real CoronaVirus.

- In South Africa claims about masks being "doused with chemicals" as a criminal strategy to rob citizens also circulated:

  People are going door to door handing out masks, they say it's a new
  initiative from local government. They will always ask you to please put
  it on to see if it fits you. It has been doused with chemicals which knocks
  you out cold and once you're knocked out they proceed to rob you. Please
  do not accept masks from strangers. Remember, we are living in critical
  times and people are desperate to take advantage with the aim of making
  money. Crime rate has skyrocketed, so please be cautious and play safe!

Economic harm

- 85% (n=28 of 33) of the claims categorised under economic harm, were labelled 'scams'. Even though the weight of claims that were recorded were from Africa Check's South African WhatsApp line, most of the scams were recorded in Kenya (43%, n=12) and Nigeria (36%, n=10). Examples of some of these claims are listed below:

  All Nigrian Citizens are Entitled to 8500 per week to stay at Home in a bid
  to control the spreed of COVID 19, proceed now to apply fiil the provided
  form and choose your bank name, your account will be credited as soon
  as possible, this is real dont ignore. The government grant is available
  to everyone starting From Monday 23 March 2020.CLICK TO APPLY ==>
  https://bit.ly/2UCS4Hu

  Covid Cash Relief from Safaricom
  Safaricom PLC will be giving out KES 2500 to all its users during this Corona
  Pandemic to help people while staying at home. Claim your share now
  https://bit.ly/SafCovidRelief

  Urgent, in support of all Nigerians
  Obtain N4000 Balance credit and 1000GB Free Internet from here:
  http://danllex.com/Nigeria

  Forwarded a link that stated that Woolworths is giving away free groceries
  worth R 5000 to support the nation during Corona pandemic. Hurry up!
  Collect your FREE voucher here:

- In some cases, convincing job opportunities at USAID or the WHO were advertised:

    JOB AT WORLD HEALTH ORGANISATION
    Help us fight CORONAVIRUS by working from home - No experience
    required SMS sending JOB
    Work 2-3 hours daily on mobile
    and earn $5-$100 daily
    Click Here And Apply Now https://bit.ly/3emUkei

    Vacancy till 31st july,2020

It's also interesting to note that in the two examples below – one from Nigeria and one from Kenya – the wording of the claim stayed almost exactly the same, but the currency and link were changed:

FG has finally approved and have started giving out free N30 000 Relief Funds to each citizen

Below is how to claim and get yours credit Instantly as I have just did now http://ngr.freeinternetz.com/

Note : You can only claim and get credited once and it's also limited so get your now Instantly.

FG has finally approved and have started giving out free KSh10 000 Relief Funds to each citizen

Below is how to claim and get yours credit Instantly as I have just did now

https://bit.ly/Ksh-fund

Note: You can only claim and get credited once and it's also limited so get your now Instantly.

The claim below was flagged to us by Kenyan and Nigerian WhatsApp users:

TO FIGHT AGAINST CORONA VIRUS — We are Giving You FREE 1000GB INTERNET to Stay Safe at Home and Enjoy Free Internet.
Click Below to Activate.
https://Covid19.Internet-Offers.site

Social harm

- Examples of claims flagged to us played into the right-wing Afrikaner narrative:

    WE WILL KILL YOUR FAMILY AND BURN YOUR HOUSE DOWN".Such is the wording used by ANC supporters in the small town of Venterspos to Mrs. Demi van Wyk, a local resident, while she distributed food parcels delivered to her from her private friends on 23 April.

- In another text-based Afrikaans claim falling into the right-wing Afrikaner narrative, someone claims that his Malawian worker and their family decided to go back to Malawi because the Economic Freedom Fighters (EFF) party has been handing out pamphlets with instructions to kill white people and foreigners on 16 June 2020 (South Africa's Youth Day). The message provides a long explanation of how his Malawian worker walked through the bushes to escape police detection and how he's been "crying like a child" in fear of what's to come.

  > *Mense, ek weet nie hoeveel waarheid daar is, in die berig híér ónder nie, ek stuur dit uit, alléénlik ter waarskuwing.*

  > *My een Malawi werker, het nou vir my 'n Boodskap gestuur uit Dunoon. Hy het gegroet & vir my totsiens gesê. Hy sê, hulle (Die Malawi ouens, wie vir my werk) gaan almal huistoe. Hulle gaan deur die bosse loop sodat die ander of polisie, hulle nie kán sien en pla nie. Hy sê daar word óral EFF pamfletjies uitgedeel, met instruksies oor die 16de van Junie 2020. Hy sê hulle sê, op die pamfletjies, dat die mense (deelnemers) hulle wapens vanaf die 14de moet gaan afhaal by die afgespreekte plekke. Hulle sé, op die ander pamflet dat Die 16 de gaan álle Wittes en Buitelanders dood gemaak word. Ek het hom gebel. Hy huil soos n kind. Hulle is so bang. Hy sê, hulle het nét na donker begin loop, Malawi toe terug. Hulle het op hulle Malawi whatsapp groepies dit wyd versprei. Hulle vlug nou al. Hy sê hulle het die gerugte al lank gehoor. Maar gedink dit is praatjies. Ek gebruik net Malawi werkers. So van môre af, staan álles stil in my besigheid.*

  Translated:

  > *People, I don't know how true the below report is, but I'm sending it solely as a warning.*

  > *My one Malawian worker just sent me a message from Dunoon. He greeted me and said farewell. He says, they (the Malawian guys who work for me) are all going home. They are going to walk through the bush so that they cannot be seen or bothered by the others, or police. He says EFF pamphlets are being handed out everywhere with instructions about the 16th of June 2020. He says they say, on the pamphlets, that the people (participants) must collect their weapons from the 14th at agreed places. They say, on the other pamphlet, that on The 16 all Whites and Foreigners would be killed. I phoned him. He cries like a child. They are so scared. He says they started walking just after dark to go back to Malawi. They disseminated it widely on their Malawi whatsapp groups. They are fleeing already. He says they have heard the rumours for a while. But thought it was just hearsay. I only use Malawian workers. So from tomorrow, everything in my business is coming to a halt.*

- On the theme of racial polarisation, there is also a claim of a 64-year old "white male" that were assaulted by "black police officers" in the Umkomaas district in KwaZulu-Natal. In a so-called media release, it's stated that:

> *This kind of racist brutality on the part of the SAPS black policemen has now become symptomatic and epidemic during the lock down in South Africa. In particular, it is the vulnerable white minority that is now under the influence of power-laden, black officials' brutality and racism.*

- And yet another forwarded text message claiming that Black South Africans plan to invade white populated areas:

  > *Alexandria will invading Sandton; Soweto invading Johannesburg South; Vosloorus invading Alberton, Boksburg etc; Mamelodi invading Silver Lakes and Attridgeville invading Harties and other supposedly well to do area.*

- We received fewer claims of racially polarising claims from other countries (noting that we received fewer claims from other countries *overall*). One claim accompanied by a photo, states that:

  > *After new legislation prohibiting using US prisoners in experiments was passed, pharmaceutical corporation Pfizer Inc. tested the volatile antibiotic Trovan on scores of Nigeria children, the tests left 11 children dead and many disabled, and resulted in a hefty out-of-court settlement between Pfizer and the Nigerian people.*

A video from Senegal that was flagged to us, claims that a vaccination team is seeking to spread the Covid-19 virus in a village in Casamance, in southern Senegal. It was categorised in the sphere of social harm, because many of the vaccination conspiracies that surfaced appeared to be led by a social belief that vaccination was used as a cruel strategy by the West "against Africans".

- On the same point, several WhatsApp messages were forwarded to us in which Obama reportedly asked Africans not to accept a vaccine from the West. The below claim flagged to us supports how the claims could fuel anti-West or anti-white sentiments with loaded phrases such as 'evil act [by] white people':

  > *I will be an accomplice if I don't denounce this evil act white people want to do to Africans, first of all I was born in America but I'm African blood, I'm not going to allow white people to kill Africans with their toxic vaccines, I ask Africans to be smart, and to ensure that coronavirus vaccines do not enter African territories, there is a Machiavellian plan they invent, saying we come to help Africans, or that they will come to kill you, I will let this message be shared everywhere, to awaken African minds so that the vaccines do not arrive in Africa.*

Political harm

- The Madagascar claim featured strongly on WhatsApp in several countries supporting narratives that could potential be anti-West

  > *World Health Organisation (WHO) has refused to acknowledge the corona virus treatment from Madagascar because "a vaccine is very unlikely to come from Africa", numerous international media have reported.*

- It's also interesting to note how the Madagascar claims 'got legs' on WhatsApp and turned into different variations, as per the text-based claim below:

  WICKEDNESS IN HIGH PLACES
  This message has been deleted, but share.

  Malagasy President Andry Rajoelina declares that WHO offered him $20,000 000 to put a little toxic in their remedy for coronavirus as the Europeans hacked their Remedy.

- The claim that the new coronavirus is a 'bio-chemical weapon' from China also surfaced frequently, in all four countries. Again, these claims were mapped at political level, but they could also sow anti-Chinese sentiments (social harm) in Africa:

  Whatever these things are, they only point to one thing that the corona is a bio-chemical weapon of China, which China has left for destruction in the world! After getting some people killed, China has now controlled this virus! Perhaps he also has medicine, which he is not sharing with the world!

- Several "classic" Covid-19 theories around 5G, a new world order, Bill Gates depopulating the world, or general anti-vaccination theories emerged. Below is one rather intricate example of such a theory:

  The pandemic is a construction of the new world order and 5G would create a monopoly of progress in the world as we would be heavily dependant on it. This was orchestrated by the Illuminati which consits of individuals like Bill Gates and Jeff Bezos including some NGO's for example by implanting microchips into human beings. "5G will connect you to everything you can think of. In fact, it is referred to as internet of everything, but it is also a gateway to where you don't want to go. The good news is that the rapture must take place before they can be in full charge of the world. The microchip implant in the body is what the Holy Bible describes as the "Mark of the Beast" "666" in the Book of Revelations Chapter 13 and once you take it you are doomed forever.

- Examples of South African WhatsApp claims illustrating how misinformation bred growing paranoia during a time when citizens felt vulnerable to government power:

  PLEASE CHECK ON YOUR PHONE TO KNOW IF SECURITY AGENTS ARE MONITORING YOUR LINE. THE STEPS ARE VERY SIMPLE.
  This is Very very informative!!
  They can be monitoring your calls and data information.
  *Please dial *#61# on your phone to know if your phone number(s)/line(s) is(are) being monitored!*
  When you dial the code (#61#), it will show either line forwarded or not forwarded.*
  If it shows "Call Forwarded" that confirms that your phone number/line is being monitored!.
  If it shows "Not forwarded" it means u are safe

Mandate To All Residents. Tonight 12 (midnight) onwards Disaster Management Act has been implemented across the country. According to this update, apart from the Govt department no other citizen is allowed to post any update or share any forward related to Coronavirus and it being punishable offence. Group Admins are requested to post the above update and inform the groups

From tomorrow onwards there are new communication regulations. All calls are recorded, All phone call recordings saved, WhatsApp is monitored, Twitter is monitored, Facebook is monitored, All social media and forums are monitored. Inform those who do not know. Your devices are connected to ministry systems. Take care not to send unnecessary messages,inform your children, Relatives and friends about this to take care
Don't forward any posts or videos etc, you receive regarding politics/ present situation about Government/PM etc.