Misinformation and Disinformation: Literature Review

Learnings on susceptibility, prevalence, impact and mitigations

Published 27 May 2025

Contents

Executive Summary

This briefing provides an overview of the available academic and grey literature on misinformation and disinformation, with a particular focus on the available evidence for the UK.

It seeks to answer the following questions:

- What is the nature of misinformation and disinformation?
- How do we measure the prevalence of misinformation and disinformation?
- Who is susceptible to believing misinformation and disinformation?
- What interventions and mitigations are effective against misinformation and disinformation?

Key findings

- **Literature on misinformation and disinformation remains largely focused on the US.** The next most common geographies are the UK and other European countries, with very little scholarship focusing on the Global South.
- **Studies on susceptibility to mis and disinformation tend to focus on investigating the short-term aspects of mis and disinformation, including the efficacy of possible interventions.** Samples tend to be US-based, often not nationally representative and drawn from locations like Mechanical Turk, or student-based. Studies tend to utilise survey and experiment-based methodologies and tend not to consider the longer-term impacts on interventions on behaviour and attitudes.
- **There is limited evidence on susceptibility to mis and disinformation in a UK context.** This is largely due to the limited scholarship available on mis and disinformation in the UK. Much of the research in this area focuses on susceptibility to conspiracy theories.
- **There is no clear evidence that any specific groups of people are inherently more or less susceptible to misinformation and disinformation.** While studies have explored lots of different demographics and their potential susceptibility, there is no consensus between studies.
- **Literature on misinformation and disinformation largely focuses on two topics: COVID-19-related mis/disinformation and political misinformation.** There are fewer pieces of research on broader health-related mis and disinformation, and fewer on climate change mis and disinformation.
- **Literature on mitigations against misinformation and disinformation largely focuses on two areas and methods: media literacy (such as source alerts and information panels) and fact-checking.** Other mitigations include content moderation, gamified solutions, automated language classification among others.
- **While establishing a causal link remains challenging, literature identifies several real-world harms associated with misinformation and disinformation.** These include health, social and societal impacts.

Across the literature, definitions of misinformation, disinformation, and related concepts like conspiracy theories differ. However, there are commonalities amongst the differential definitions used, and for the purposes of this literature review, we use the following definitions:

- **Misinformation:** a rhetorical strategy that produce and disseminates false or misleading information, spread unintentionally, that tends to confuse, influence, harm, mobilise or demobilise an audience.[1]
- **Disinformation:** a rhetorical strategy that produces and disseminates false or misleading information in a deliberate effort to confuse, influence, harm, mobilise or demobilise a target audience.[2]
- **Conspiracy theory:** a belief that two or more actors have coordinated in secret to achieve an outcome and that their conspiracy is of public interest but not public knowledge. Conspiracy theories:
    a. are oppositional, which means they oppose publicly accepted understandings of events;
    b. describe malevolent or forbidden acts;
    c. ascribe agency to individuals and groups rather than to impersonal or systemic forces;
    d. are epistemically risky, meaning that though they are not necessarily false or implausible, taken collectively they are more prone to falsity than other types of belief; and
    e. are social constructs that are not merely adopted by individuals but are shared with social objectives in mind, and they have the potential not only to represent and interpret reality but also to fashion new social realities.[3]

Overall, the literature on mis and disinformation remains focused on audiences in the US, with most studies into the manifestation and mitigation of mis and disinformation conducted on US-based samples and focused on US audiences.

[1] Spies, Samuel, 2019, Defining "Disinformation", *MediaWell,* published 22 October 2019, accessed 25 July 2024
[2] Spies, Samuel, 2019, Defining "Disinformation", *MediaWell,* published 22 October 2019, accessed 25 July 2024
[3] Douglas, Karen M.; and Sutton, Robbie M. (2023): "What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication", *Annual Review of Psychology*, 74, pp. 271-298

Understanding susceptibility to misinformation and disinformation

Measuring susceptibility to mis and disinformation

Despite an increase in the attention paid to mis and disinformation by academic and civil society researchers during and following the COVID-19 pandemic, little research has been conducted into what makes people susceptible to believing mis and disinformation. What research exists is often topic-specific, as well as geographically-specific (for example, US-based studies looking at partisan affiliation).

Individual factors

Factors relating to the characteristics of individuals, such as age, gender, ethnicity and others, can impact their susceptibility to mis and disinformation.

However, the literature investigating susceptibility is limited, and often cannot draw definitive, causal conclusions. Studies are often survey-based and identify correlations, such as: "people with characteristic X are more likely to say they believe in mis and disinformation", rather than suggesting a causal link between a specific characteristic and belief or susceptibility to mis and disinformation.

Some studies have also offered more general explanations for why an individual may be more susceptible to mis and disinformation.

For example, an overview of psychological research suggests that people are drawn to conspiracy theories when three types of psychological needs are not being met:

- **Epistemic needs:** the desire for understanding, or accuracy, or subjective certainty.
- **Existential needs:** the desire for control or security.
- **Social needs:** the desire to maintain a positive image of the self or in-group.[4]

The overview notes that psychological research also suggests that belief in conspiracy theories is stronger among people who habitually seek meanings or patterns in their environment. Additionally, belief in conspiracy theories is stronger when especially large-scale, significant events (like a pandemic or large-scale terrorist attack) happen and leave people dissatisfied with mundane and small-scale explanations.[5]

The following sections will discuss the literature on the interactions between susceptibility to or belief in misinformation, disinformation and conspiracy theories and various characteristics, including age, gender, income levels, education levels and more.

[4] Douglas, Karen M.; Sutton, Robbie M., and Cichoka, Aleksandra, 2017, The Psychology of Conspiracy Theories, *Current Directions in Psychological Science*, 26:6, pp. 538-542

[5] Douglas, Karen M.; Sutton, Robbie M., and Cichoka, Aleksandra, 2017, The Psychology of Conspiracy Theories, *Current Directions in Psychological Science*, 26:6, pp. 538-542

Age

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various demographic factors. On age, the overview found that several studies found that older people were more susceptible to misinformation, noting that one theory suggested that older adults tend to rely on their existing knowledge when confronted with new information.[6]

However, another study found that a "notable minority of the public also believe conspiracy theories about a COVID-19 vaccine – with belief especially high among younger people and those who get a lot of information on the pandemic from social media platforms".[7]

The OECD Truth Quest Survey, a large-scale survey measuring the ability of individuals in 21 countries to identify false and misleading content, found that a participant's perceived ability to identify false and misleading content online was uncorrelated with their measured ability to do so. The survey found that confidence in identifying false and misleading content online tended to decrease with age.[8]

In September 2021, as part of Ofcom's survey into UK individuals' attitudes and consumption of news and information on COVID-19, those aged under 35 were more likely to say that they had come across potentially false or misleading claims than those over the age of 35. (33% of under-35s compared with 20% of those aged 35 and over).[9]

A survey by Adobe of over 6,000 individuals across several countries, including over 2,000 UK-based respondents, found that only 57% of individuals feel confident that they can spot misinformation. However, the survey found that 1 in 4 Gen Z respondents in the UK admitted to sharing misinformation in the past six months, though 73% of Gen Z respondents said they could confidently spot misinformation. 17% of millennial respondents admitted to sharing misinformation, with 70% of millennial respondents saying they could confidently identify misinformation. 10% of Gen X respondents admitted to sharing misinformation, and 53% of Gen X respondents said they felt confident identifying misinformation. 4% of Baby Boomers admitted to sharing misinformation, and 37% of Baby Boomers said they felt confident identifying misinformation.[10]

In a study of 66,242 individuals from 24 countries, which has participants complete the Misinformation Susceptibility Test (MIST) and indicate their self-perceived ability to identify misinformation, found that Generation Z (defined in this study as those born between 1997 and 2012), non-male, less educated, and more conservative individuals were more susceptible to misinformation. However, despite performing worst in the test, the research found that Generation Z were able to perceive their misinformation discernment ability most accurately.[11]

---

[6] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[7] King's College London, 2020: Coronavirus: vaccine misinformation and the role of social media, published 14 December 2020

[8] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

[9] Ofcom, 2021. Covid-19 news and information: consumption and attitudes, Key findings from week 76

[10] Adobe, 2024. Adobe Future of Trust Study Narrative (UK), published 18 April 2024

[11] Kyrychenko, Yara; J. Koo, Hyunjin; Maertens, Rakoen; Roozenbeek, Jon; van der Linden, Sander; Götz, Friedrich M., 2025. Profiling misinformation susceptibility, *Personality and Individual Differences*, 241

Gender

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various demographic factors. The study found conflicting evidence on how gender influences susceptibility to misinformation, with some studies finding that men are more likely to share health information without fact-checking it, and others find that women are more likely to share information from websites that contain fake news.[12]

A survey of 4,343 UK residents found positive correlations between female gender and conspiracy suspicions.[13]

The OECD's Truth Quest survey found that a participant's perceived ability to identify false and misleading content online was uncorrelated with their measured ability to do so. The survey found that across all countries, men had higher confidence in their ability to identify false and misleading content online than women.[14]

Education levels

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various demographic factors. In reference to education levels, several studies find that higher levels of education are associated with decreased belief in conspiracy theories. Additionally, it highlighted several other studies that suggested people with higher cognitive ability are less susceptible to misinformation, with some showing that individuals who engage in more analytical reasoning are less susceptible to misinformation. [15]

The OECD's Truth Quest survey found that a participant's perceived ability to identify false and misleading content online was uncorrelated with their measured ability to do so. The survey found that confidence in identifying false and misleading content online tended to increase with an individual's education level.[16]

A survey of 4,343 UK residents found that factors associated with vaccine hesitancy included low levels of education.[17]

Ethnicity

During the COVID-19 pandemic, Ofcom commissioned regular surveys from March 2020 to September 2021 about the UK population's consumption and attitudes towards COVID-19 news and information. Survey responses between July and September 2021 found that people from a minority ethnic background (31%), including 31% of Asian respondents and 32% of Black respondents, were more likely than White respondents (25%) to say that they had come across news or information

---

[12] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[13] Allington, Daniel; McAndrew, Siobhan; Moxham-Hall, Vivienne; and Duffy, Bobby (2021): "Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the Covid-19 pandemic", *Psychological Medicine*, 53, pp. 236-247

[14] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

[15] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[16] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

[17] Allington, Daniel; McAndrew, Siobhan; Moxham-Hall, Vivienne; and Duffy, Bobby (2021): "Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the Covid-19 pandemic", *Psychological Medicine*, 53, pp. 236-247

about COVID-19 that could be false or misleading. Where respondents had seen claims that could be considered false or misleading, people from a minority ethnic background (43%) were almost twice as likely as White respondents (23%) to agree that seeing these claims made them think twice about the issue.[18]

A study highlighting the need for transnational research on the spread of mis- and disinformation in Asian diasporic communities highlights how first-generation immigrants in the US, who may have limited English proficiency, turn to ethnic media, including print, broadcast and social media, as primary information sources. The study suggests that these channels may be a key location for exposure to misinformation and disinformation.[19]

An overview of psychological research into belief in conspiracy theories suggests that experiences of ostracism can cause people to believe in conspiracy theories, occultism and superstitions, as part of their efforts to make sense of their experiences. Members of groups with objectively low societal status because of their ethnicity, income or other related factors, are more likely to endorse conspiracy theories.[20]

A survey of 4,343 UK residents found that ethnic minority status was positively correlated with conspiracy suspicions and use of social media for information on COVID-19.[21]

Income levels

An overview of psychological research into belief in conspiracy theories suggests that experiences of ostracism can cause people to believe in conspiracy theories, occultism and superstitions, as part of their efforts to make sense of their experiences. Members of groups with objectively low societal status because of their ethnicity, income or other related factors, are more likely to endorse conspiracy theories.[22]

The OECD's Truth Quest survey found that a participants' perceived ability to identify false and misleading content online was uncorrelated with their measured ability to do so. The survey found that confidence in identifying false and misleading content online tended to increase with income level. The survey also found that respondents in the lowest income bracket of most countries consistently performed the worst at identifying false and misleading content, while respondents in the highest income bracket performed the best.[23] A survey of 4,343 UK residents found factors associated with vaccine hesitancy included low-income levels.[24]

---

[18] Ofcom, 2021. Covid-19 news and information: consumption and attitudes

[19] Ngyuen, Sarah; Kuo, Rachel; Reddi, Madhavi; Li, Lan; and Rachel E. Moran; 2022, Studying mis- and disinformation in Asian diasporic communities: The need for critical transnational research beyond Anglocentrism, *Harvard Kennedy School Misinformation Review*

[20] Douglas, Karen M.; Sutton, Robbie M., and Cichoka, Aleksandra, 2017, The Psychology of Conspiracy Theories, *Current Directions in Psychological Science*, 26:6, pp. 538-542

[21] Allington, Daniel; McAndrew, Siobhan; Moxham-Hall, Vivienne; and Duffy, Bobby (2021): "Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the Covid-19 pandemic", *Psychological Medicine*, 53, pp. 236-247

[22] Douglas, Karen M.; Sutton, Robbie M., and Cichoka, Aleksandra, 2017, The Psychology of Conspiracy Theories, *Current Directions in Psychological Science*, 26:6, pp. 538-542

[23] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

[24] Allington, Daniel; McAndrew, Siobhan; Moxham-Hall, Vivienne; and Duffy, Bobby (2021): "Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the Covid-19 pandemic", *Psychological Medicine*, 53, pp. 236-247

Trust in media

The OECD Truth Quest Survey found that respondents from the UK had the lowest level of trust in news from social media of surveyed countries, with around a quarter of people trusting social media news some or a lot.[25]

Participants in a small, qualitative study into the experiences of those holding minority beliefs in health protection, the Russian invasion of Ukraine, and climate change suggested that a large range of world events were not reported on by traditional, legacy media sources.[26] There is some suggestion that lower trust in media may be related to susceptibility to mis and disinformation.

Trust in government

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various demographic factors. The study found several pieces of research linked a lack of trust in government with greater beliefs in conspiracy. The study also noted that some other studies suggest that greater trust in government in associated with greater adherence to COVID-19-related guidelines.[27]

Literacies

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various types of literacies.

The study found that higher levels of health literacy had been associated with being less susceptible to health-related misinformation, though the evidence base on this is mixed, with some studies finding that people were likely to use unaccredited sources to answer health-based questions.[28]

On numerical literacy, the study suggested that some studies have found that trust in scientists and higher numeracy skills were associated with lower susceptibility to COVID-19-related misinformation. [29]

The study suggested that the evidence on the relationship between various literacies (digital, media and information literacies) tends to suggest that higher levels of these literacies is associated with lower levels of susceptibility to misinformation. However, some evidence has suggested that some media or digital literacy interventions do little to reduce people's susceptibility.[30]

Political ideology

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various demographic factors. The study reported several studies that found conspiracy theories are more likely to be believed by those with extreme political

---

[25] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

[26] Ofcom, 2023, Understanding experiences of minority beliefs on online communications platforms,

[27] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[28] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[29] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[30] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

views. In addition, several studies show that “conservative right-wing" beliefs are associated with misinformation. However, further investigation is needed into “left-wing false information”, as much scholarship has focused on right-wing, conservative misinformation. [31]

Personality traits

A 2021 study examining susceptibility to health misinformation in the context of the COVID-19 pandemic provided an overview of the literature surrounding susceptibility to mis and disinformation and its relationship to various demographic factors. The evidence on the link between susceptibility to believing misinformation and personality traits are mixed. Some studies suggest that those who score lower on ‘agreeableness’ are more likely to interact with online misinformation.[32]

Content factors

Factors relating to content on online services, such as method of presentation, medium, style and others, can also impact people’s susceptibility to mis and disinformation.

A 2021 study examining susceptibility to health misinformation, specifically in the context of the COVID-19 pandemic, identified the following features of content that may impact susceptibility to misinformation:

- **How content is presented:** The study suggests that online content can lack the heuristics (mental shortcuts we use to understand the world) that legacy media and other forms of traditional offline content have. Some studies have shown that adding pictures to content and presenting the content in easy-to-read font increases the perceived trustworthiness of the content.
- **Style and understandability**: The study suggests that the ease with which an individual can process content can affect the believability of that content, with some studies suggesting that easy-to-understand information may be more believable. In a similar vein, some studies argue that misleading content, rather than entirely fabricated content, is more likely to be accepted.
- **The source of content**: The study suggests that the creator of a piece of content can affect its perceived trustworthiness, with some studies finding that messages shared from sources perceived to be trustworthy are more likely to be shared by others.
- **Warnings:** The study suggests that **s**ocial media services are increasingly attaching warnings to false and misleading content, with several studies suggesting that this can be effective in combatting harmful untruths, though others suggest that these warnings can be ineffective. A notable adverse effect of these warnings is the implied truth effect, where false headlines that are not given warnings are considered implicitly validated and therefore accurate – even though they may have just not been reviewed.
- **Information overload:** The study highlights this as another factor that may make people more susceptible to mis and disinformation, defining it as where people find it hard to

[31] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

[32] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to online misinformation, *The Alan Turing Institute*

understand and make decisions about issues when they are faced with too much information.

Some research has found that the source of content and the perspective from which it is presented can make it more persuasive for some audiences. Ipsos and Ofcom's research into the experiences of those self-identifying as holders of minority beliefs found that the availability of eye-witness testimony (or content which appears to be eye-witness testimony) and video footage on online communications platforms held a particular power for some participants. If this footage or testimony was not covered by traditional media outlets, these participants questioned the process by which traditional media outlets decide whether to report on a topic, rather than the veracity of the evidence.[33]

In addition, a study of COVID-19 misinformation by the Reuters Institute for the Study of Journalism found that 59% of this misinformation involved forms of reconfiguration (existing and often true information either spun, twisted, recontextualised or reworked, with 38% of it completely fabricated). The study also identified three sub-types of misinformation:

- Misleading content (29%) – some true information but details reformulated, selected and re-contextualised ways that made them false or misleading.
- Labelling or describing images or videos as being something other than they are (24%) – which the study notes is sometime referred to as 'malinformation'.
- Small number of manipulated images and videos employing low tech non-sophisticated techniques. Referred to as 'cheap fakes' rather than deepfakes.[34]

Disinformation can also be categorised by the ways in which content is presented or manipulated as part of its creation.

Hamleers et al, 2020, highlight four key techniques used in the creation of multimodal disinformation (disinformation that combines visual and textural techniques to create false or misleading content):

1. **De-contextualisation:** pairing real images or videos with false, manipulated or misleading text;
2. **Reframing:** Cropping or de-contextualising videos to make certain aspects of issues more obvious or prominent in pursuit of a specific agenda;
3. **Visual doctoring:** Manipulating images or videos to present a different reality than they contain in their non-edited form;
4. **Multimodal doctoring:** Fabricating content by pairing manipulated images or videos with false, misleading or manipulated text.[35]

The OECD Truth Quest Survey found no differences in people's ability to correctly identify true or false and misleading content across the three main themes it studied (health, international affairs and the environment).[36]

---

[33] Ofcom, 2023, Understanding experiences of minority beliefs on online communications platforms,

[34] Simon, Felix; Howard, Philip N.; and Nielsen, Rasmus Klein (2020): "Types, sources and claims of Covid-19 misinformation", Reuters Institute for the Study of Journalism, published 7 April 2020

[35] Hameleers, Michael; Powell, Thomas E,; Van Der Meer, Toni G.L.A; and Bos, Lieke, 2020, A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated on Social Media, *Political Communication*, 37:2, pp.281-301

[36] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

The impact of AI-generated content

Recent advancements in generative AI technologies have led to the recognition of both the risk and opportunities associated with these technologies. The World Economic Forum's Global Risk Report 2024 highlighted how AI could be used to amplify manipulated and distorted information that may be used to destabilise societies.[37]

Some recent studies have investigated how well individuals do at discerning between content created by a generative AI model and content created by humans, as well as investigating whether misleading AI-generated content is more persuasive, or harder to detect as false, than misleading human-generated content.

The OCED Truth Quest survey found that on average, respondents found AI-generated content easier to identify than human-generated content. Across all surveyed countries, AI-generated disinformation was 10 percentage points easier to correctly identify as false compared with human-generated disinformation. [38]

This finding conflicts with a different study, which found that AI models are better at both producing accurate information that is easier to understand, and at producing more compelling disinformation, than humans. This same study found that participants were not able to distinguish between social media posts generated by an AI model and social media posts produced by humans.[39]

A survey by Adobe of over 6,000 individuals across several countries, including over 2,000 UK-based respondents, found that only 57% of individuals in the UK feel confident that they can spot misinformation.[40]

[37] World Economic Forum (2024): Global Risks Report 2024, published 10 January 2024, accessed 16 May 2024

[38] OECD, 2024, The OECD Truth Quest Survey: Methodology and Findings,

[39] Spitale, Giovanni; Biller-Andorno, Nikola; and Germani, Federico (2023): AI model GPT-3 (dis)informs us better than humans, *Science Advances*, 9:26

[40] Adobe, 2024. Adobe Future of Trust Study Narrative (UK), published 18 April 2024

Prevalence and Impacts

Challenges in estimating prevalence

In the context of online communications, prevalence is one way to conceptualise the proportional exposure of an audience to online content or behaviour. For misinformation and disinformation, we can understand the prevalence as the proportion of online users exposed to a specific piece of mis- and disinformation content, or a specific mis- or disinformation narrative, at a specific point in time.

Establishing the prevalence of mis and disinformation can be challenging, especially in a non-US context. Mis- and disinformation are context and time-specific, and retroactively assessing the prevalence of mis and disinformation can be further complicated by the removal of violative content by services, many of which have policies prohibiting the sharing of mis and disinformation.

In addition, there are disagreements about what mis and disinformation are, and who gets to decide what is and is not mis and disinformation. Some people feel that the mainstream media is also a key location for the spreading of mis and disinformation, as several participants highlighted in Ipsos and Ofcom's qualitative research with those who held minority beliefs.[41]

However, there are some estimates of the prevalence of mis and disinformation about specific issues in the UK. These are outlined in the following section.

Prevalence estimates

Ofcom's Online Experience Tracker (OET) found that 68% of respondents had a high level of concern about misinformation, with the same proportion (68%) expressing concern about fake or deceptive images and videos.[42]

A study of 2,244 UK residents aged 16-75 found that 1 in 3 people and nearly half (46%) of all 16-34-year-olds said that they had seen or heard messages discouraging the public from getting a COVID-19 vaccine.[43]

In September 2021, an Ofcom survey on UK individuals' attitudes and consumption of news and information about the COVID-19 pandemic found that 24% of people said that they had encountered claims about COVID-19 that could have been false or misleading. The most common claims seen by respondents, from a prompted list, were:

- Face masks/coverings offer no protection or are harmful (seen by 22%)
- The flu alone is killing more people than COVID-19 (19%)
- The number of deaths linked to COVID-19 is much lower than is being reported (17%)
- COVID-19 does not exist and was genetically engineered (15%)

[41] Strong, Colin; Owen, Katy; and Mansfield, Jill. 2023. Understanding experiences of minority beliefs on online communication platforms, *Ipsos*, published September 2023
[42] Ofcom, 2024. Online Experiences Tracker Wave 5, accessed 26 July 2024
[43] Duffy, Bobby; Beaver, K.; and Meyer, C. (2020): "Coronavirus: vaccine misinformation and the role of social media", The Policy Institute, King's College London

In addition, the survey found that over half (58%) of social media users said they had seen posts on social media with warnings or notices that the information may be untrustworthy or untrue.[44]

A Reuters Institute study of COVID-19-related misinformation found that misinformation from politicians, celebrities and prominent public figures made up 20% of claims in their sample but account for 69% of total social media engagement with COVID-19-related misinformation.[45] The study also found that 80% of misinformation claims came from ordinary social media users and these generated less engagement. However, the report noted a few instances of bottom-up misinformation generating a large reach but the true spread, which includes through private groups and encrypted messaging applications, was not captured.[46]

A survey of over 6,000 individuals across several countries, including over 2,000 UK-based respondents by Adobe, found that 37% of UK respondents reported that they had seen someone they know share what they believe to be misinformation in the past six months. Additionally, the survey found that 13% of UK respondents said they had shared content in the past six months that turned out to be misinformation, with 62% of these respondents trying to correct the misinformation once they discovered it was misleading.[47]

Prevalence on different services

A study into vaccine misinformation and the role of social media by King's College London, carried out prior to the approval of vaccinations against COVID-19 by UK medical authorities, found that 1 in 3 people in the UK said they had seen or heard messages discouraging social media users from getting a COVID-19 vaccine, should one become available. Of these, 58% reported seeing them on Facebook – by far the top source cited, which amounts to 1 in 5 people (20%) in the UK saying they've seen such messages on Facebook.

This high proportion will partly reflect the fact that Facebook has a larger user base than other social media companies. There are far fewer people who report seeing anti-COVID-19-vaccine messages on other platforms. For example, 19% of those who say they have seen this kind of content say Twitter[48] was a source, and 17% say Instagram. This is equal to around 6% of the UK population, and a similar proportion of the UK public say a friend or family member was a source of such messages.[49]

Impacts of mis and disinformation

Health impacts

Several studies have suggested a link between vaccine hesitancy[50] and holding COVID-19-related conspiracy beliefs.

---

[44] Ofcom, 2021. Covid-19 news and information: consumption and attitudes

[45] Simon, Felix; Howard, Philip N.; and Nielsen, Rasmus Klein (2020): "Types, sources and claims of Covid-19 misinformation", Reuters Institute for the Study of Journalism, published 7 April 2020

[46] Simon, Felix; Howard, Philip N.; and Nielsen, Rasmus Klein (2020): "Types, sources and claims of Covid-19 misinformation", Reuters Institute for the Study of Journalism, published 7 April 2020

[47] Adobe, 2024. Adobe Future of Trust Study Narrative (UK), published 18 April 2024.

[48] This study was conducted prior to the service changing its name to X

[49] Duffy, Bobby; Beaver, K.; and Meyer, C. (2020): "Coronavirus: vaccine misinformation and the role of social media", The Policy Institute, King's College London

[50] Vaccine hesitancy is defined as: the delay in acceptance of or refusal or vaccination despite availability of vaccination services. Source: Allington, Daniel; McAndrew, Siobhan; Moxham-Hall, Vivienne; and Duffy, Bobby (2021): "Coronavirus

A survey-based study of 4,343 UK adults, stratified for representativeness of the UK population for age, gender, region and working status between November and December 2020 found that the most powerful predictors of vaccine hesitancy were conspiracy suspicions and general vaccine attitudes.[51]

In another study investigating factors that impact vaccine uptake or hesitancy, of the factors that were associated with increased likelihood of vaccine willingness, age, and trust in health organisations such as the NHS and the WHO had the strongest bivariate associations. The oldest respondents were over 20 times more likely to express willingness to get the vaccine compared to the youngest.[52] Of factors that decreased the likelihood to get the vaccine, conspiracy beliefs had the largest effect, followed by distrust of vaccines, belief in COVID-19 misinformation and 'lockdown scepticism'. Users of Instagram, YouTube, Snapchat and TikTok were all less likely to express a willingness to be vaccinated. Only YouTube users were significantly less willing to be vaccinated, with a two-thirds likelihood of vaccine willingness compared to non-users.[53]

Another study found that a "notable minority of the public also believe conspiracy theories about a COVID-19 vaccine – with belief especially high among younger people and those who get a lot of information on the pandemic from social media platforms".[54]

Misinformation can mask itself as credible infection and prevention control strategies and have serious implications if someone prioritises it over evidence-based guidelines. Research suggests that a popular myth that consumption of highly-concentrated alcohol could disinfect the body and kill the virus was circulating across the world. Following this misinformation, approximately 800 people have died, 5,876 have been hospitalised and 60 have developed blindness after drinking methanol as a cure for COVID-19.[55]

Social impacts

In a small, qualitative study Ipsos and Ofcom conducted into those who hold minority beliefs and their experiences on online platforms, participants reported facing relationship challenges with their spouses and partners, alongside broader challenges with their relationships with family, friends, and workplace colleagues because of their beliefs. However, some participants also reported that their engagement with minority beliefs had resulted in new friendships and a sense of community with like-minded others.

In the same study, some participants reported their minority beliefs having impacts on their workplace relationships, such as feeling under pressure, having difficult conversations, and feeling the need to self-censor to avoid negative repercussions. This was especially in reference to health protection and the COVID-19 vaccine.

---

conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the Covid-19 pandemic", *Psychological Medicine*, 53, pp. 236-247

[51] Allington, Daniel; McAndrew, Siobhan; Moxham-Hall, Vivienne; and Duffy, Bobby (2021): "Coronavirus conspiracy suspicions, general vaccine attitudes, trust and coronavirus information source as predictors of vaccine hesitancy among UK residents during the Covid-19 pandemic", *Psychological Medicine*, 53, pp. 236-247

[52] Jennings, Will; Stoker, Gerry, Bunting, Hannah; Valgarosson, Viktor Orri; Gaskell, Jennifer; Devine, Daniel; McKay, Lawrence; and Mills, Melinda C. (2021): "Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy", *Vaccines*, 9:6, 593

[53] Jennings, Will; Stoker, Gerry, Bunting, Hannah; Valgarosson, Viktor Orri; Gaskell, Jennifer; Devine, Daniel; McKay, Lawrence; and Mills, Melinda C. (2021): "Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy", *Vaccines*, 9:6, 593

[54] King's College London; and University of Bristol (2021): "Coronavirus conspiracies and views of vaccination", published 31 January 2021

[55] Islam, Saiful, et al.; 2020, COVID-19-Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis, *The American Journal of Tropical Medicine and Hygiene,* 103:4, pp. 1621-1629

Some of the participants in the study saw themselves as marginalised, due to their beliefs, and expressed anger over their perception of being negatively perceived in wider society due to their minority beliefs. However, other participants found that engaging with minority beliefs was a rewarding experience and gave them a sense of empowerment.[56]

A survey by Abode of over 6,000 individuals across several countries, including over 2,000 UK-based respondents, found that nearly 1 in 3 people (29%) have cut their social media activity based on the frequency of misinformation being spread on platforms.[57]

Societal impacts

Recent research has suggested that COVID-19 conspiracy theories have negative consequences for people's intentions to comply with government recommendations. They can also lead to people developing support for alternative remedies, a willingness to commit vandalism or violence, and stockpiling goods.[58]

A different study of reactions to the announcement of the UK's first national lockdown on Twitter noted that panic buying of food and household items, criminal damage and vandalism of 5G masts, and riots in Ukraine over the evacuation of citizens from China during the early stages of the pandemic were impacts of misinformation around COVID-19.[59]

A survey of over 6,000 individuals across several countries, including over 2,000 UK-based respondents by Adobe, found that 81% of UK respondents agree that misinformation is one of the biggest threats facing society. 78% of UK respondents said that they feared misinformation and deepfakes would impact upcoming elections and interfere with the democratic process. [60]

[56] Ofcom, 2023, Understanding experiences of minority beliefs on online communications platforms,
[57] Adobe, 2024. Adobe Future of Trust Study Narrative (UK), published 18 April 2024
[58] Douglas, Karen M.; 2021, Covid-19 Conspiracy Theories, *Group Processes & Intergroup Relations,* 24:2
[59] Green, Mark; et al.; 2021, Identifying how Covid-19-related misinformation reacts to the announcements of the UK national lockdown: An interrupted time-series study, *Big Data & Society*
[60] Adobe, 2024. Adobe Future of Trust Study Narrative (UK), published 18 April 2024

Mitigations: Options, Costs, Efficacy

Fact-checking

What is it?

Fact-checking is a process that verifies information. Fact-checking is one of the most well-studied solutions for reducing spread of mis- and disinformation online and in other mediums, such as in print and broadcast media.

Efficacy evidence

While most studies find that fact-checks do improve platform users' abilities to discern between real and fake news[61], there is disagreement over whether the format of a fact-check affects its effect. Some studies have found that videos are a much more effective format for fact-checks[62], while others have found no difference in effect between video-based and text-based fact checks.[63]

Studies have also found that while fact-checks improve users' ability to detect false news, they do not reduce the sharing of this fake news by users.[64] Other studies have identified a "continued influence effect," whereby individuals continue to rely on misinformation even after it has been debunked.[65][66]

Other studies have suggested more novel uses of fact-checking, such as creating databases of fact-checked content for services to check against as part of efforts to prevent the spread of mis and disinformation.[67]

---

[61] Bor, Alexander; Osmundsen, Mathias; Rasmussen, Stig Hebbelstrup Rye; Bechmann, Anja; and Petersen, Michael Bang, 2021, "Fact-checking" videos reduce belief in, but not the sharing of fake news, *PsyArxiv*,
[62] Courchesne, Laura; Ilhardt, Julia; and Shapiro, Jacob N., 2021, Review of Social Science Research on the Impact of Countermeasures against Influence Operations, *Harvard Kennedy School Misinformation Review*,
[63] Hameleers, Michael; Powell, Thomas E,; Van Der Meer, Toni G.L.A; and Bos, Lieke, 2020, A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated on Social Media, *Political Communication*, 37:2, pp.281-301,
[64] Bor, Alexander; Osmundsen, Mathias; Rasmussen, Stig Hebbelstrup Rye; Bechmann, Anja; and Petersen, Michael Bang, 2021, "Fact-checking" videos reduce belief in, but not the sharing of fake news, *PsyArxiv*,
[65] Ecker, U.K.H; Lewandowsky, S.; and Tang, D.T.W (2010): Explicit warnings reduce but do not eliminate the continued influence of misinformation, Memory & Cognition, 38, pp. 1087-1100
[66] Lewandowsky, S.; Ecker, U. K. H.; Seifert, C. M.; Schwarz, N.; and Cook, J. (2012): Misinformation and Its Correction: Continued Influence and Successful Debiasing, Psychological Science in the Public Interest, 13:3, pp. 106-131
[67] Reis, Julio C. S.; Melo, Philipe; Garimella, Kiran; and Benevenuto, Fabricio, 2020, Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation?, *Harvard Kennedy School Misinformation Review,*

Content labelling: state media labels and source alerts

What is it?

Source alerts are a type of content label that platforms apply to either accounts or posts that are suspected or confirmed to be from state-sponsored or state-linked entities, or from locations known for sharing unverified information.

Generally, this form of labelling is intended to provide information about the editorial independence (or lack of) of media sources on services, rather than about the funding model they operate under. Several services have published their criteria for this labelling. For example, TikTok labels accounts run by entities whose editorial output or decision-making process is subject to control or influence by a government[68], and Meta defines state-controlled media as media outlets it believes may be partially or wholly under the editorial control of their government.[69]

These labels are distinct from verification labels, which are used by services to authenticate an account owner (for example, an elected representative, or official government department), as opposed to providing an indication of the editorial angle from which they post content.

Efficacy evidence

Scholarship has found that the success of these source labels is mixed. The literature focuses mostly on the use of labels on popular social media services and mostly uses US samples in the research.

Studies have found that the success of these alerts is highly dependent on their format, with their efficacy reduced, as would be expected, if the label is hard to see or easy to ignore.[70] A different study found that the success of these alerts differs across political affiliations and services, with source alerts found to be more effective on X (then Twitter) than on Facebook.[71] In addition, studies have suggested that the corrective effects of these labels is dependent on the label being noticed and the information they contained being absorbed.[72]

Content labelling: warnings and publisher information

What is it?

Online services are increasingly making it easier for their users to identify the author or publisher of online news. Additionally, platforms are providing contextual information to users so they can better judge the accuracy of content themselves.

Efficacy evidence

Research on this mitigation is mixed. One study investigating whether this intervention helped users distinguish between accurate and inaccurate content found that publisher information had no significant impact on whether participants perceived the headline as accurate or expressed an intent

[68] TikTok (Erlich, Justin) 2023, TikTok's state-affiliated media policy, published 18 January 2023

[69] Meta (Gleicher, Nathaniel), 2020: Labeling State-Controlled Media on Facebook, published 4 June 2020

[70] Nasseta, Jack; and Gross, Kimberly; 2020, State Media Warnings Can Counteract the Effect of Foreign Misinformation, *Harvard Kennedy School Misinformation Review*

[71] Arnold, Jason Ross; Reckendorf, Alexandra; and Wintersieck, Amanda L.; 2021, Source Alerts Can Reduce the Harms of Foreign Disinformation, *Harvard Kennedy School Misinformation Review*

[72] Nasseta, Jack; and Gross, Kimberly; 2020, State Media Warnings Can Counteract the Effect of Foreign Misinformation, *Harvard Kennedy School Misinformation Review*

to share it.[73] Research has found that a notable adverse effect of these warnings is the implied truth effect, where false headlines that are not given warnings are considered implicitly validated and therefore accurate – even though they may have just not been reviewed.[74] However, a study by The Alan Turing Institute found that these warnings had a significant, but very small, effect on participants, reducing the amount of error by 0.03 (of their total error score[75]) on average.[76]

Downranking and algorithm amendments on search engines

What is it?

Search engines use an algorithm to decide which results are most relevant and rank the results that they show users accordingly. *Downranking* means that a search engine has made the deliberate choice to move a domain further down the list of results shown for a particular search query. This practice has been applied to domains hosting health misinformation, alleged foreign influence operation domains, domains hosting CSAM content and domains hosting terror content.

Efficacy evidence

In 2019, Bing, was found to return at least 125 sources of disinformation and misinformation, while Google returned 13, across the top 50 results for 12 separate queries (a total of 600 results).[77] This research suggests that the latter's algorithmic adjustments have impacted the prevalence of disinformation in its results.

Costs and risks

On search engines with a publicly-stated commitment to user privacy like DuckDuckGo, some users have criticised their decision to downrank Russian disinformation content as part of their broader opposition to social media content moderation more broadly.[78]

Election-specific measures

What are they?

Following the Internet Research Agency's well-publicised foreign interference activities across a variety of social media services targeting the 2016 US Presidential elections, and wider concerns about mis and disinformation in elections, many social media platforms implemented specific

[73] Dias, N., Pennycook, G., & Rand, D. G.; 2020, Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media, *Harvard Kennedy School Misinformation Review*

[74] Pennycook, Gordon; Bear, Adam; Collins, Evan T; and Rand, David G.; 2020, The Implied Truth Effect: Attaching Warning to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings, *Management Science*, 66:11,

[75] The error score is the difference between the participants' assessment and the correct assessment.

[76] Vidgen, Bertie; Taylor, Harry; Pantazi, Myrto; Anastasiou, Zoe; Inkster, Becky; and Margetts, Helen; 2021, Understanding vulnerability to onlineWhi misinformation, *The Alan Turing Institute*,

[77] Bush, Daniel; and Zaheer, Alex; 2019, Bing's Top Search Results Contain an Alarming Amount of Disinformation, *Stanford Freeman Spogli Institute for International Affairs*,

[78] https://mashable.com/article/duckduckgo-search-engine-russian-disinformation

policies to attempt to mitigate the spread of misleading election content and added features that allow users to report it.

Efficacy evidence

Meta's *Voting Alerts* was used to send over 80 million election-related notifications in 2022.[79] However, there is very little visible scholarship investigating the efficacy of these election-specific measures.

Sensitive events policies

What are they?

Some services have introduced specific policies around "sensitive events," which can include elections, terrorist attacks, and natural disasters, among others.

What platforms use it?

Google introduced a *Sensitive events policy* as part of its broader advertising policies, which prohibits ads that potentially profit from or exploit "sensitive event with significant social, cultural and political impact", which it suggests includes "civil emergencies, natural disasters, public health emergencies, terrorism and related activities, conflict or mass acts of violence".[80] Following the Russian invasion of Ukraine in February 2022, Google updated this policy to note that due to the invasion, it paused ads from and for Russian state-funded media.[81] Twitter (now X) has a *Crisis Misinformation Policy* that is applied during situations of armed conflicts, public health emergencies and large-scale natural disasters.[82]

Efficacy evidence

There is very little scholarship investigating the efficacy of these measures.

Account take-downs

What are they?

This is when accounts are removed from the platform, often due to cumulative violations of terms of service. This is sometimes referred to as 'deplatforming'. This mitigation is used in a variety of contexts and in response to a variety of harms and violative content on services – the efficacy evidence discussed below does not focus on account takedowns resulting from mis and disinformation specifically.

Efficacy evidence

Based on an analysis of Facebook's own Coordinated Inauthentic Behaviour reporting, the most common way that Facebook detected coordinated inauthentic behaviour stemmed from known

[79] How Meta Is Planning for the 2022 US Midterms | Meta (fb.com)
[80] https://support.google.com/adspolicy/answer/6015406?hl=en-GB#:~:text=Sensitive%20events&text=Ads%20that%20potentially%20profit%20from,or%20mass%20acts%20of%20violence.
[81] Policy update: Sensitive events - Google Merchant Center Help
[82] https://blog.twitter.com/en_us/topics/company/2022/introducing-our-crisis-misinformation-policy

networks and actors, including previous investigations by Facebook. This could suggest that bad actors tend to reoffend, and that content or account removal alone is not a sufficient deterrent.[83]

Costs and risks

More generally, account take downs can push users and networks onto smaller, more extreme platforms. For example, after the events of 6 January 2021 in the US, Gab gained over 2 million new users in January 2021, the biggest monthly growth in the service's history.[84] A study examining the effects of deplatforming on social networks found that while permanently suspending users did assist in safeguarding the service that did the deplatforming, suspended users that moved to alternative platforms had smaller audiences on the new platform but were more active and became more toxic.[85] Another study identified other unintended consequences from deplatforming accounts as including the spawning of 'minion accounts' that perform the role of spreading the disinforming material being produced by a de-platformed influencer off-platform and increasing the resilience of a group targeted by deplatforming by encouraging them to diversify their cross-service presences.[86]

Automated language classification tools

What are they?

In recent years, the development of automated tools based on machine learning and other computer science techniques, to detect dis- and misinformation on social media platforms has significantly expanded. Many studies explore models and tools that are trained on pre-existing datasets from services themselves.

Efficacy evidence

Many automated and AI-based tools face several critical limitations: firstly, they require a substantial number of data examples to learn specific tasks, they lack the "world context" required to understand deliberately misleading content[87], and this lack of context can lead to a "lack of common sense" in their decision-making.[88]

Costs and risks

Many automatic fact checking systems are not mature enough to operate without human oversight, and many have been trained on databases that are already out-of-date as the news cycle moves

[83] https://www.isdglobal.org/wp-content/uploads/2020/10/Hoodwinked-2.pdf

[84] Thiel, David; and McCain, Miles; 2022, Gabufacuturing Dissent: An in-depth analysis of Gab", *Stanford Cyber Policy Review*, pg. 2-7

[85] Ali, Shiza; Saeed, Mohammad Hammas; Aldreabi, Esraa; Blackburn, Jeremy; De Cristofaro, Emiliano; Zannettou, Savvas; and Stringhini, Gianluca; 2021, Understanding the Effect of Deplatforming on Social Networks, 13TH *ACM Web Science Conference 2021, 21-25 June*

[86] Innes, H.; and Innes, M., 2021. De-platforming disinformation: conspiracy theories and their control, *Information, Communication & Society,* 26:6, pp. 1262-1280

[87] Islam, Rafiqul MD; Liu, Shaowu; Wanf, Xianzhi; and Xu, Guandong; 2020, Deep learning for misinformation detection on online social networks: a survey and new perspectives, *Social Network Analysis and Mining*

[88] NATO Strategic Communications Centre of Excellence, 2022, The Role of AI in the Battle Against Disinformation,

quickly.[89] Problems can also arise when the data used to train these systems suffers from quality issues, like missing, biased, corrupted, or incorrectly labelled data.[90]

Meta-data based forwarding limits

What is it?

Forwarding limits restrict the number of times a message can be forwarded between accounts and sometimes label messages that have been forwarded over a set number of times.

Efficacy evidence

However, a lack of researcher access to the kinds of data that are only held by services has made it hard to gauge the efficacy of this mitigation and whether it is a viable strategy more broadly.[91] A study assessing the efficacy of these forwarding limits in India, Indonesia and Brazil found that they significantly reduce the spread of information, but that they do not block the spread of misinformation through public groups when the content has a high level of virality.[92]

Gamified solutions

What is it?

Gamified anti-misinformation interventions are games that are developed to build cognitive resistance against common forms of manipulation that people may encounter online. Some examples include the games *Bad News, Go Viral!* and *Harmony Square,* with the latter focusing on exposing common tactics used in election misinformation.[93]

Efficacy evidence

A study found that playing the game *Harmony Square* significantly reduces the perceived reliability of fake news, significantly increases players' confidence in their ability to spot fake news, and significantly reduces participants' self-reported willingness to share fake news.[94] A similar study on the game *Bad News* also found that playing it increased participants' ability to spot misinformation and increased their level of confidence on their own judgement of whether content is misinformation or not.[95] A different study used a news literacy game, *Fakey,* to examine the effect social engagement metrics (likes, shares and comments) had whether individuals like and share

[89] Caled, Danielle; and Silvia, Mario J.; 2021, Digital Media and Misinformation: An Outlook on Multidisciplinary Strategies Against Manipulation, *Journal of Computational Social Science*, 5, pp. 123-159,
[90] Ofcom (Winder, Phil; Marsden, Luke; and Rotundo, Enrico), 2023: Automated Content Classification (ACC) Systems, published January 2023
[91] Gorksy, Jacob; and Woolley, Samuel; 2021, Countering Disinformation and Protecting Democratic Communication on Encrypted Messaging Applications, Brookings Institute,
[92] Melo, Philipe Vieira; Carolina, Garimella; Kiran Vaz de Melo, Pedro; and Benevenuto, Fabrício; 2019, Can WhatsApp Counter Misinformation by Limiting Message Forwarding?, *Political Communication*,
[93] Roozenbeek, Jan; and Van Der Linden, Sander; 2020, Breaking Harmony Square: A game that "inoculates" against political misinformation, *Harvard Kennedy School Misinformation Review*,
[94] Roozenbeek, Jan; and Van Der Linden, Sander; 2020, Breaking Harmony Square: A game that "inoculates" against political misinformation, *Harvard Kennedy School Misinformation Review*,
[95] Basol, Melissa; Roozenbeek, Jan; and Van Der Linden, Sander; 2020, Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News, *Journal of Cognition*, 3:1,

questionable content, and whether they will fact-check less questionable sources.[96] Research into the development of a Misinformation Susceptibility Test found that while *Bad News* does decrease individuals' susceptibility to fake news and reduces general naiveté, the game can generate increased general distrust, or hyper-scepticism.[97] Furthermore, a pre-print study re-examining some studies of efficacy of the games *Bad News* and *Go Viral!* found that the two games did not improve participants discrimination against mis and disinformation, but instead increased the number of "false" responses made to all news items, meaning that the games simply made participants more conservative in their assessment of the veracity of news items.[98]

Content provenance measures

What is it?

Content provenance measures, including content credentials and watermarking, aim to provide content creators and consumers with information about the content they create and consume.

One of the most popular measures, the C2PA standard, is a technology that records digitally signed information about content, which shows where the piece of media has come from and how it has been edited. These systems use strong cryptographic binding to attach this context to content, but the technology is not intended to serve as a determinant of whether the image itself is real or fake. Rather, the technology enables users and creators to see transparently what has been done to an image prior to publication.[99] The C2PA standard combines cryptographic watermarking and metadata embedding.

Whilst some content provenance measures provide a visual record of the provenance of an image or piece of content, some content provenance measures, such as types of watermarking, do not necessarily have to be easily publicly identifiable. Forms of watermarking can include embedding information into digital multimedia content that can be detected or extracted by machine for a variety of purposes, with techniques including metadata watermarking, frequency component watermarking, cryptographic watermarking, and statistical watermarking.[100]

Efficacy evidence

The efficacy and vulnerability to attack of these techniques often depend on the format of the watermark and the location of it. For example, metadata watermarks can have limited efficacy, as many online services and file-sharing methods can strip the metadata from files, and it is comparatively easy to edit, remove or otherwise modify metadata. Cryptographic watermarks can be difficult or impossible to remove from content without damaging the content and are often invisible or undetectable by normal sight or sound, preserving the original value of the content. However, this means that users and creators are not necessarily presented with information about

[96] Avram, Mihai; Micallef, Nicholas; Patil, Sameer; and Menczer, Filippo; 2020, Exposure to social engagement metrics increases vulnerability to misinformation, *Harvard Kennedy School Misinformation Review*

[97] Maertens, Rakeon, et al.; 2023, The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment, *Behaviour Research Methods,* 56, pp. 1863 - 1899

[98] Modirrousta-Galian, A. and Higham, P. A., 2022, Gamified Inoculation Interventions Do Not Improve Discrimination Between True and Fake News: Reanalyzing Existing Research With Receiver Operating Characteristic Analysis, *PsyArXiv*

[99] Halford, Charlie, 2024. Mark the good stuff: Content provenance and the fight against disinformation, *BBC Research and Development,* published 5 March 2024

[100] Vasse'i, Ramak Molavi; and Udoh, Gabriel, 2024. In Transparency We Trust? Evaluating the Effectiveness of Watermarking and Labelling AI-Generated Content, *Mozilla Foundation*, published 26 February 2024

the origin of content that has been watermarked in this way, meaning that they can still be deceived or manipulated. [101]

[101] Vasse'i, Ramak Molavi; and Udoh, Gabriel, 2024. In Transaprency We Trust? Evaluating the Effectiveness of Watermarking and Labelling AI-Generated Content, *Mozilla Foundation*, published 26 February 2024

Summary

Research Gaps

Through the course of conducting this literature review, we have identified several gaps in the literature.

UK-specific evidence

As a UK-based regulator, we are particularly interested in evidence that focuses on the UK context.

While there is some evidence from a UK context available on the prevalence of misinformation, disinformation, and conspiracy theories, and on factors that may make individuals more susceptible to these phenomena, there is limited UK-specific evidence on the efficacy of mitigations.

In general, there is limited evidence on the efficacy of interventions against mis and disinformation outside of a US context. A systematic review of interventions against COVID-19 misinformation found that 72% of interventions included in the study were tested on US participants, and only 7% of the interventions included were tested on populations outside of the US, Canada, and Europe. After the US, the UK was the next most-common population to have interventions tested upon it, with 14% of interventions included in the study tested on UK populations.[102]

Causation

While many studies suggest links between various demographic, personal or content factors and misinformation, disinformation and conspiracy theories, there is very little evidence for any causal links between these factors. This means that we have very little idea of *why* a particular factor may be linked to belief in misinformation, disinformation, and conspiracy theories.

Whilst this research gap likely arises from the difficulty in determining causation, it remains a significant gap in the evidence base.

[102] Smith, Rory; Chen, Kung; Winner, Daisy; Friedhoff, Stefanie; and Wardle, Claire; 2023, A Systematic Review of COVID-19 Misinformation Interventions: Lessons Learned, *Health Affairs, 42:12*