Behavioural Research UK

False and Misleading Information: An Evidence Summary and Map for Policy and Practice

May 2026

Commissioned by the Government Office for Science

Authors
Dr Harriet M. Baird, School of Psychology, University of Sheffield, UK
Dr Hui Zhang, School of Psychology, University of Sheffield, UK
Dr Lesley Uttley, School of Medicine and Population Health, University of Sheffield, UK
Professor Kalina Bontcheva, School of Computer Science, University of Sheffield, UK
Professor Thomas L. Webb, School of Psychology, University of Sheffield, UK
Joanna Wright, School of Computer Science, University of Sheffield, UK
Professor Linda Bauld, College of Medicine and Veterinary Medicine, University of Edinburgh, UK

Acknowledgements: We gratefully acknowledge the contributions of the following experts, whose insights through the expert consultation process helped to shape and refine this work: Dr Madalina Botan, Dr John Cook, Paula Gori, Professor Stephan Lewandowsky, Clare Melford, Chris Morris, Stephan Mündges, Pallavi Sethi, Colin Strong, Professor Sander van der Linden, and Bob Ward. Names are listed in alphabetical order and do not indicate the level or nature of contribution.

To cite this report: Baird, H. M., Zhang, H., Uttley, L., Bontcheva, K., Webb, T. L., Wright, J., & Bauld, L. (2026). False and misleading information: An evidence summary and map for policy and practice. Open Science Framework. https://doi.org/10.17605/OSF.IO/8CKDY.

Contents

1 Executive Summary

This report outlines findings from an evidence review (covering the period 2020-2025) informed by expert consultation to respond to a need across government for a robust, accessible synthesis of current evidence on false and misleading information. This includes such information across online platforms, AI systems, mainstream and blended media, private communication channels, web and advertising ecosystems. The report is structured around:

- How such information should be understood and measured
- Its impacts, and
- Which interventions are effective in countering and mitigating harms from such information

1.1 Understanding and Measuring the Problem

False and misleading information is now widely recognised as a systemic national risk, rather than a series of isolated incidents. International and UK assessments identify it as a persistent threat to democratic governance, public trust, national security, and effective policy delivery.

There are conceptual and policy challenges associated with how terms to describe false and misleading information are used and interpreted, but debates in the literature around these should not become a distraction from action to address harms. Research, policy and practice are increasingly moving away from categorising types of content (as mis-, dis- or malinformation, for example) towards strengthening the resilience of the broader information environment

The nature and sources of false and misleading information are evolving. Data from the European Digital Media Observatory briefing series (June 2021 to January 2026) suggests that the phenomenon is no longer episodic or confined to moments of crisis, but has instead become a permanent and embedded feature of the information environment. Whereas COVID-19 once dominated this landscape, false and misleading content now circulates across multiple topics simultaneously. Within this environment, these domains not only co-exist but increasingly interact, with false or misleading narratives crossing boundaries (e.g., public health risks being attributed to migration within immigration debates).

The average number of topics per month nearly doubled over the five-year period, with health, conflict and immigration narratives dominating. The most significant changes to the nature of false and misleading information were driven by AI-generated content.

There are a range of distributional mechanisms for false and misleading information. Harm is driven not by content alone, but by amplification systems, including:

- Engagement-optimised recommender algorithms and monetisation mechanisms
- Repetition effects (illusory truth – where repeated exposure increases the perceived accuracy of a statement - and illusory consensus, where amplification methods give a false impression of prominence to fringe or false viewpoints)
- Blended ecosystems (social media + mainstream media + advertising + AI generated search results)
- Private and personalised channels (messaging apps, AI companions)

These main distribution mechanisms and their effects are reinforced by limited transparency, weak governance, and restricted data access for researchers.

1.2 Assessing the Impacts

Harms operate at three interconnected levels: individual; societal; and organisational.

Individual harms are the most direct and easy to evidence. They include, for example, preventable deaths from false health claims (e.g. about the effectiveness of treatments or vaccines), and financial losses from fraud. However, it is important to take a measured approach when assessing individual impact. While clear examples of harm traceable to false and misleading information exist, this must be set against the broader evidence that it is generally more difficult to persuade people of information that conflicts with their existing beliefs.

Societal harms include erosion of trust in science, media, and public institutions; increased societal polarisation; and a broader breakdown in shared standards of truth. This has economic costs as low levels of trust can increase transaction costs in markets and governance.

Organisational harms affect businesses, public institutions, and democratic systems, particularly during crises. These include disruption to industries, reputational damage, and interference with democratic processes.

Across all levels of harm, health misinformation is well studied but evidence gaps remain significant in areas such as climate, democratic participation, gendered disinformation, and impacts on marginalised communities. This reflects disparities in research attention rather than the relative importance of these harms.

There are also challenges in evidencing causal links between false and misleading information and harms. This is due to: limited data access; methodological gaps (e.g. no standard framework for measurement and many small-scale studies); and personalised AI-generated content making monitoring and attribution particularly difficult.

1.3 Countering and Mitigating False and Misleading Information

There is consensus in the literature and among experts that no single intervention works alone. Effective responses require layered, adaptive strategies, combining individual and system-level actions.

At the individual level:

- Prebunking as a ‘first line of defence’ is supported by strong evidence. It reduces susceptibility to false and misleading information by explaining manipulation techniques before exposure. However, its effects decline without reinforcement - booster interventions are necessary.
- Debunking (correcting false claims) can be effective but is limited by the “continued influence effect” when the claim continues to shape beliefs and behaviour even after an individual accepts a correction. They are most effective when explanatory, timely, and delivered by trusted sources.
- Media and digital literacy builds long-term resilience, but is slow to deliver, doesn’t reach all groups, and cannot necessarily address acute threats.

At the system level:

- Voluntary platform self-regulation is insufficient. Statutory frameworks with independent audit and enforcement are needed
- Researcher access to platform data is a critical bottleneck. Evidence suggests access has deteriorated, undermining accountability and evaluation

1.4 Key policy considerations arising from the evidence

1.4.1 Nature of the Risk

- False and misleading information is a systemic risk, rather than one that focuses on events or issues as they arise
- It is continuously evolving, meriting sustained and continued mitigation
- Targeted attacks represent a strategic risk, not just a communications challenge
- AI-driven, personalised information environments are becoming an increasing source of false and misleading information

1.4.2 Evidence Gap and Scope

- The evidence is relatively weak on false and misleading information relating to climate, democracy, economic decision-making and the effects on marginalised communities, making policy response more difficult
- Research is needed across topics, beyond health to cover climate, democracy, and targeted harms
- The COVID-19 pandemic led to increased funding and a temporary greater openness in data access, acting as a key catalyst for research and response efforts

1.4.3 Drivers and Challenges

- Algorithms and business models contribute to the amplification and impact of false and misleading information
- Platform accountability and data access represent major challenges
- Coordination is key across jurisdictions (including internationally), given the limits of national regulation

1.4.4 Responses and Interventions

- System-level interventions are needed, especially addressing algorithms and business models
- Embedding prebunking (including personalised and identity-based messaging) across education, crisis preparedness and public communications has potential to improve societal resilience, alongside the ongoing value of debunking false and misleading information.

2 Evidence Map

Alongside this report, a visual map has been developed to provide readers with an overview of the key terms, topics, and issues associated with false and misleading information. The map is organised around the three review questions addressed in this report: (i) understanding and measuring the problem, (ii) assessing impacts, and (iii) countering and mitigation (see example outline below). A database of the key literature is provided as a supplementary resource, alongside a publicly available Zotero reference library. The literature has been coded using key terms aligned with both thematic areas (e.g., definitions, actors, harms, interventions) and domains of interest (e.g., health, climate change). This serves as a practical reference tool for officials and policymakers seeking evidence on specific areas of interest (e.g., to enable users to identify studies examining interventions to counter false and misleading information in the context of climate change).

3 Introduction

The spread of false and misleading information, variously described as misinformation or disinformation among other terms, is one of the most pressing challenges facing governments and societies worldwide. The scale of concern is reflected in the World Economic Forum's *Global Risks Report* (2026), which draws on the views of over 1,300 experts, leaders, and policymakers. The report ranks false and misleading information among the most severe risks across all three time-horizons: immediate (2026), short-term (to 2028), and long-term (to 2036), making it one of the few threats to feature persistently regardless of the timeframe considered. This assessment is reinforced by the United Nations Global Risk Report (2024), which identifies false and misleading information as the most prominent "global vulnerability": a risk that is both extremely important and one for which the international community is insufficiently prepared, particularly in terms of reduction and mitigation.

At the national level, the UK Parliament's Foreign Affairs Committee (2026) similarly characterises foreign disinformation as a form of "new warfare," warning that responses remain fragmented, under-resourced, and insufficient to protect open democracies. Moreover, the UK OFCOM report on adults' media use and attitudes (2026) shows that susceptibility and harms vary across age groups (e.g., increased anxiety and body image issues for young adults and social isolation and financial loss for older adults). The challenge is made even more difficult to address effectively since false and misleading information is sometimes hard to differentiate clearly from legitimate opinion, thus creating a risk that aggressive moderation could inadvertently suppress legitimate dissent or protected speech. This tension forces a delicate balancing act where efforts to protect public truth must be weighed against the fundamental right to hold and share ideas without state or corporate overreach (Bontcheva & Posetti, 2020).

These UK, WEF, and UN reports highlight the role of false and misleading information as a meta-risk: one that does not operate in isolation but amplifies other major global threats by distorting public understanding and impeding coordinated responses. Across a wide range of policy domains, from public health and climate change to elections and national security, it has the potential to undermine evidence-based policymaking, erode public trust, weaken democratic processes, increase societal polarisation (Ognyanova et al., 2020; Surjatmodjo et al., 2024), and cause harm to individuals (Bhandari et al., 2026; Lyons et al., 2024), communities (Surjatmodjo, 2026; Saner, 2026), and society as a whole (Surjatmodjo et al., 2024; Spampatti et al., 2025). While the spread of false and misleading information is not a new phenomenon, propaganda, rumour, and

fabrication have shaped public discourse throughout history, the emergence of social media has fundamentally transformed its reach, speed, and impact (Plikynas et al., 2025; Sultan et al., 2024; López-Borrull & Lopezosa, 2025). Platforms designed to maximise engagement have created environments in which misleading content can spread rapidly and at scale, often outpacing corrections and reaching audiences far beyond what traditional media channels allow (Vosoughi, 2018). The problem has become even more acute recently, due to the widespread use of Artificial Intelligence as a creator and amplifier of false and misleading information at scale, cheaply, and in a targeted manner (Bentzen, 2025b; Bontcheva et al., 2024). For example, in the context of Foreign Information Manipulation and Interference alone, a 259% growth in AI-enhanced FIMI activities has been reported in Europe in 2025 by the European Union External Action Service (EEAS, 2026).

The scale and urgency of this challenge have driven significant growth in research on false and misleading information. Progress has been made, predominantly within the health domain, in identifying impacts and exploring strategies for reducing these impacts. However, much of the existing evidence remains fragmented, with many findings emerging from isolated or relatively small-scale studies that can be difficult to translate into clear, actionable insights. As reflected both in the literature and in expert perspectives, there also remains substantial debate regarding the prevalence and severity of the problem, as well as the most effective approaches for preventing or mitigating it at scale (van der Linden et al., 2025; Diaz Ruiz, 2025). These challenges are further compounded by the conditions under which research is conducted, including funding constraints and limited access to platform data.

At the same time, the nature of false and misleading information, its mechanisms for spreading, and impacts continue to evolve, adapting in response to emerging prevention measures and tools, while also exploiting the changing capabilities of Artificial Intelligence (AI), platform governance, and attempts at regulation.

3.1 Scope of This Review

This evidence summary was commissioned by the Government Office for Science (GO-Science) on behalf of the UK Government's Chief Scientific Adviser and departmental Chief Scientific Advisers. It responds to a recognised need across government for a robust, accessible synthesis of the current evidence base on false and misleading information to inform HMG policy and operations across departments affected by the problem. The review was designed to provide an overview of the

published evidence, to highlight where there are issues of contention, and where significant gaps in research remain. The review is structured around three thematic areas, each corresponding to a core policy question:

1. **Understanding and measuring the problem:** including challenges in how concepts related to false and misleading information are defined and distinguished, the range of actors involved in creating and amplifying such information, and the mechanisms through which it spreads across both online and offline environments.
2. **Assessing impacts**: including harms to individuals, societies, and organisations, as well as key debates in measuring these impacts. We need to understand what types of information have what effects, on whom, and in which domains; ideally identifying principles that help understand impacts across domains.
3. **Countering and mitigation**: including the range of individual- and system-level strategies available to manage the impact of false and misleading information, the evidence on their effectiveness, key limitations, and the factors shaping their effectiveness, alongside directions for future research and policy.

3.2 Summary of Approach

We combine a rapid review of the evidence with expert consultation. Given the rapidly evolving nature of the landscape, the review focuses primarily on review-level evidence, including systematic reviews, meta-analyses, and narrative syntheses, published in peer-reviewed academic journals from 2020 onwards. In addition to the academic evidence, reports and reviews have examined aspects of false and misleading information in recent years, including work by Ofcom, the Royal Society, the Alan Turing Institute, the Organisation for Economic Co-operation and Development (OECD), and various parliamentary committees. We integrate insights from these key reports alongside the academic evidence. To complement the synthesis of review-level evidence, and to account for the time lag in the production of such syntheses, we also draw selectively on recent primary studies to identify emerging trends and developments.

The rapid review was combined with consultation with experts working in the field of false and misleading information. The aim was to help ensure coverage and balance, assess the limitations and UK relevance of the evidence, identify areas of convergence

and disagreement, and highlight priority gaps for policy and future research. Full details relating to the approach can be found in Section 9 (*Methods*).

3.3 Note on Terminology Used in This Report

This report primarily uses the term “false and misleading information” as an umbrella term to refer to the wider phenomenon of inaccurate or misleading information. While we acknowledge that there are important conceptual distinctions between terms (e.g., misinformation represents sharing false or misleading information ***without*** the intent to deceive, whereas disinformation represents sharing false or misleading information ***with*** the intent to deceive), intent, while important, is often difficult to establish in practice. Where the nature of the information (including the presence or absence of intent) is relevant, this will be explicitly specified. A more detailed discussion of definitions and terminology is provided in Section 4.1 (Definitions and Terminology), where these issues are examined in the context of understanding and measuring the problem.

4 Findings: Understanding and Measuring the Problem

Understanding and measuring the occurrence, spread, and distribution of false and misleading information is a necessary foundation for any response. This section addresses four aspects of the problem: (i) how key concepts are defined and distinguished; (ii) how the content and nature of false and misleading information has evolved in recent years; (iii) who the key actors are in its creation and amplification; and (iv) how false and misleading information spreads.

4.1 Definitions and Terminology

Section Summary: Debates about how to define and categorise false and misleading information are legitimate and serve a purpose, but should not become a distraction from the mechanisms that cause the greatest harm. While intent matters in some contexts, it is impact on audiences that should be the primary lens for policy responses. Established terms such as misinformation and disinformation remain valuable and should not be abandoned in response to political pressure, as doing so risks normalising the very attacks on the field they are designed to deflect. Above all, work on definitions and terminology should serve policy and intervention design (e.g., identifying the most harmful content and the structural conditions that amplify it) rather than becoming an end in itself.

There are a range of terms to describe false and misleading information. These include widely used terms such as *misinformation*, *disinformation*, and *malinformation,* as well as a growing number of related terms that describe specific forms of manipulation or information harm (e.g., *deep fakes*, *conspiracy theories, propaganda,* and *foreign information manipulation and interference*). An overview of the main terms used in research, policy, and public debate is provided in Table 1.1.

Rather than providing detailed definitions of each term, this section focuses on the broader conceptual and policy challenges associated with how such terminology is used and interpreted. This reflects our consultation with experts, who emphasised that an excessive focus on definitional distinctions can detract from understanding how false and misleading information operates in practice, the impacts it has, and how it can be addressed at scale.

Table 4.1 Definition of key terms related to false and misleading information used in research and policy

Term	Definition	Example
Misinformation	False or inaccurate information shared without intent to harm.	Sharing an outdated health claim believing it to be true
Disinformation	Fabricated or deliberately manipulated content created with intent to harm a person, social group, organisation, or country.	State-sponsored fabricated news stories targeting an election
Malinformation	Genuine information used with the intent to harm. Includes leaks, harassment, and hate speech where the content is true, but the purpose is to cause harm.	Leaking genuine private communications to damage a political opponent
Propaganda	The deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behaviour to achieve a response that furthers the desired intent of the propagandist.	Wartime messaging that selectively omits unfavourable facts
Conspiracy theories	Explanations of important events as secret plots by powerful and malevolent groups. Belief is driven by epistemic, existential, and social motives and tends to be resistant to counter-evidence.	Claims that public health measures are coordinated population control
Deepfakes	AI-generated synthetic media (including images, audio and video) that makes it possible to depict real people saying and doing things they never said or did, using machine learning techniques.	Fabricated video of a politician making inflammatory statements
Satire / parody	Content that, while a form of art or commentary, can become misinformation when audiences misinterpret the message and share it as factual.	A satirical news article shared without context as factual
Foreign Information Manipulation and Interference (FIMI)	A mostly non-illegal pattern of behaviour that threatens or has the potential to negatively impact values, procedures and political processes. Manipulative in character, conducted in an intentional and coordinated manner by state or non-state actors.	Coordinated inauthentic social media campaigns targeting foreign elections

4.1.1 Key Terminology Considerations

Ongoing debates about terminology have introduced a number of conceptual and practical challenges. The key issues identified by experts are outlined below:

1. Cognitive intent: Terminology emphasising the assumed human intent behind the spread of false or misleading information (i.e., misinformation; disinformation; malinformation) requires knowledge of whether it was shared intentionally (e.g., in the case of disinformation) or unintentionally (e.g., in the case of misinformation). Experts highlight three important considerations:

- **First, intent is often difficult to establish.** Research shows that in many real-world contexts, the motivations behind the creation and sharing of false and misleading information cannot be reliably determined (Pennycook & Rand, 2021; Littrell et al., 2023). That said, in certain domains intent is well-established (e.g., the deliberate deception practised by the tobacco and fossil fuel industries has been extensively documented legally), and recent work challenged the view that intent is fundamentally unknowable. For example, Lewandowsky et al. (2024) argue that, although intent cannot be observed directly, disinformation can often be distinguished from good-faith disagreement by looking at (in)consistent patterns in statements, behaviour, and available evidence. Where intent can be identified, it may reflect financial, political, or strategic motivations, often shaped by platform dynamics that reward engagement, amplification, and polarisation. The interaction between these incentives and platform design is explored further in Sections 4.3 and 4.4.
- **Second, Intent evolves across contexts.** Intent behind the information may also change as it spreads meaning the boundaries between these terms often become blurred. For example, content created with deceptive intent may later be shared by individuals who believe it to be accurate, meaning disinformation can subsequently circulate as misinformation. Vice-versa, inadvertently shared misinformation may be picked up and purposefully amplified by AI bots and sockpuppet accounts, as well as by individuals who knowingly share falsehoods for strategic or social reasons, often described in the literature as “participatory propaganda” (Lewandowsky, 2022).
- **Third, intent should not be the sole primary focus for policy.** The harm experienced by those who encounter false or misleading information occurs regardless of whether content was created maliciously, shared in good faith, or produced for profit. For this reason, impact on audiences is a critical lens for policy responses. At the same time, understanding intent remains important for identifying the underlying drivers of misinformation and disinformation, and for designing interventions that address their root causes.

2. Limited differences in impacts: In many cases, the impacts of false and misleading information and the strategies required to address it are similar regardless of intent.

- False or misleading claims may undermine public understanding, influence behaviour, or erode trust irrespective of whether they were created deliberately or not. Likewise, many interventions, such as improving information literacy, strengthening transparency, and enhancing platform governance, apply equally to both mis- and disinformation.
- That said, intent is not always entirely irrelevant to impact. Research suggests that people can respond differently depending on perceived source motivation, with a coordinated disinformation attack via automated bot networks landing differently psychologically than an honest error in a news report (Lewandowsky et al., 2024).

3. A shift towards information integrity: Research and policy is increasingly moving away from categorising types of false and misleading content and towards strengthening the resilience of the broader information environment. This shift is reflected in emerging concepts such as *information integrity*, *healthy knowledge ecosystems*, and *resilient information environments*.

- The United Nations (2025) define information integrity as "...*an information ecosystem in which reliable and accurate information is available to all, enabling people to engage meaningfully in public life, make informed decisions and exercise their rights"*
- The core pillars of information integrity are: (i) accuracy and authenticity (information is truthful, fact-checked, and sourced from trusted, independent channels) (ii) consistency & transparency (reliable access to information without censorship, including transparency in how algorithms amplify content), (iii) safety & security (protection against the malicious spread of false or misleading information that threatens public health, safety, and democratic processes, and (iv) fidelity (information is understood as originally intended, not manipulated by being presented outside of its original context).
- Whether 'information integrity' will be less contested or debated as a term than those it seeks to replace remains to be seen.

4. Risk of distracting from core issues: An excessive focus on definitions and terminology can distract from the broader questions of how false or misleading information spreads, the negative impacts it creates, and the most effective strategies to counter and mitigate harm.

- Importantly, this distraction is not always accidental. Definitional ambiguity is sometimes deliberately exploited to undermine the field. Claims that misinformation "cannot really be defined," do not advance conceptual clarity so much as cast doubt on the legitimacy of research and policy efforts. Evidence suggests far greater consensus among academic experts than these critiques imply (e.g., Altay et al., 2023, HKS Misinformation Review). While there remains scholarly debate about the boundaries of key terms (e.g., misinformation), this is not the same as the concept being meaningless (Lewandowsky, 2024).
- Experts emphasised a deeper risk: that preoccupation with definitions can draw attention away from the mechanisms that actually cause harm at scale. False and misleading content does not become a societal problem simply by existing - it becomes one through large-scale amplification. It is the architecture of the internet, and engagement-based algorithms in particular, that drives this amplification, systematically rewarding enraging content because outrage generates engagement, and engagement generates advertising revenue. In this view, it is the plumbing of the internet that most urgently requires regulatory attention, not the content itself.
- Definitions still serve an important purpose: not all content poses equal risk, and distinguishing between types of false and misleading information is necessary for proportionate and effective responses. The key is to ensure that definitional work supports policy and intervention design, rather than displacing it.

5. Weaponisation of terms: Experts highlighted that previously established terminology (i.e., misinformation; disinformation; malinformation; fake news) has become increasingly weaponised.

- 'Fake news' in particular is now widely avoided in research and policy contexts given that it is imprecise and highly politicised. More broadly, for much of 2025, terms such as mis- and disinformation were avoided by major philanthropic funders and government departments alike, in part to avoid attracting criticism from the new US administration.

- However, experts cautioned strongly against overcorrecting. The trend toward abandoning established terminology risks becoming a form of advance compliance with authoritarian pressure rather than a considered conceptual choice. While it can sometimes be useful to refer to 'false and misleading information', adopting it as an exclusive replacement for more specific terms would be a mistake.

4.2 The Nature and Sources of False and Misleading Information between 2021-2026

Section Summary: Between 2021 and 2026, false and misleading information shifted from episodic disruption to a permanent, multi-domain feature of the information environment. The number of concurrent topics nearly doubled, with health, conflict, and immigration narratives persisting at consistently high levels. The emergence of generative AI further accelerated this trend, enabling more scalable and sophisticated forms of manipulation. Together, these developments have produced a complex “polycrisis” landscape.

This section draws on data from fact-checking organisations that has been compiled by the European Digital Media Observatory (EDMO) briefing series, covering June 2021 to January 2026. The dataset provides a longitudinal view of false and misleading information across Europe, enabling analysis not only of ***what*** narratives circulate, but how their ***scale***, ***composition***, and ***persistence*** have changed over time. Instances were coded into 11 thematic categories (see Table 2.1 for an overview; full dataset can be found here).

The rationale for choosing a European source for the data analysis is mainly due to the limited breadth of fact-checked content in the UK, the limited scope, funding and staffing capacity of these UK organisations, and the cross-border spread of false and misleading information on social media. Below the two most active UK fact-checking organisations and their broad topics are elaborated for completeness:

- FullFact is a UK non-partisan fact-checking charity, with a primary mission dedicated to promoting accuracy in public debate and holding powerful institutions accountable. The primary focus is fact-checks on claims made by

politicians, public officials, and the media, mostly limited to the economy, health, immigration, and education.

- BBC Verify is a team within the BBC, also with limited staff capacity, which primarily focuses on high-stakes international conflicts, forensic verification of viral videos, and investigating large-scale disinformation campaigns or conspiracy movements.

Table 4.2 Thematic Categories of False and Misleading Information in EDMO Data

Category	Description and scope
Health	Vaccine misinformation, COVID-19 narratives, miracle cure claims, and pandemic policy. Present in 100% of monitored months, the most persistent category across the full period.
Climate change	Climate denial, weather map manipulation, geoengineering conspiracies (including HAARP and directed energy weapons), and the misattribution of natural disasters to deliberate technological intervention.
Election	Electoral fraud claims, deepfake-enabled campaign interference, and vote manipulation narratives. Active in 76% of months overall, rising to 100% by 2023, indicating a shift from episodic to year-round activity.
Immigration	False claims about migrant entitlements, misrepresentation of the demographic composition of migrant arrivals, and the misattribution of criminal acts to migrant communities. Present in 89% of months.
Security & conflict	War-related false narratives, staged attack claims, and conflict imagery manipulation. Dominated by Ukraine-related content from February 2022, with Gaza narratives emerging from October 2023.
Foreign information manipulation	Coordinated state-linked or state-adjacent influence campaigns, copycat news websites, fabricated statements attributed to political leaders, and narratives designed to amplify domestic divisions in European countries.
Racism	Racially motivated false narratives, xenophobic, Islamophobic and antisemitic content, and ethnically targeted disinformation, including content associating minority communities with crime or cultural threat.

AI-generated	Deepfake images and video, synthetic audio, voice cloning, and AI-generated content used for political manipulation. Not tracked as a distinct category prior to March 2023; present in 97% of subsequent months.
Gender, LGBTQ+ & EDI	Anti-LGBTQ+ narratives, 'gender ideology' claims targeting schools and public institutions, Pride event misrepresentation, and transgender athlete controversies.
Financial	Exploitation of economic anxiety through false narratives about sanctions, EU financial policy, digital currency surveillance, and the misrepresentation of aid flows and public spending.
Miscellaneous	False and misleading narratives not captured by the above categories, including broader institutional delegitimisation (EU, national governments) and event-specific claims not falling within a named thematic area.

4.2.1 Key findings from the EDMO Briefs

Our analysis reveals that the landscape of false and misleading information has grown substantially in both volume and complexity over the five-year period. Below we summarise the key findings:

1. From Episodic to Structural Saturation: False and misleading information is no longer episodic or crisis-bound. Instead, it has become a permanent and structurally embedded feature of the information environment.

- No category fell below 71% monthly prevalence across the 56-month period.
- Health-related false and misleading information appeared in 100% of months, making it a constant baseline.
- Immigration and conflict-related narratives exceeded 89% prevalence, indicating long-term persistence rather than reactive spikes.
- This marks a clear shift from earlier models (e.g. COVID-19 in 2020–21), where attention was concentrated around a single dominant issue.

Table 4.3 Prevalence of False and Misleading Information by Category

Category	Months present	Prevalence	Classification
Health-related (incl. COVID-19)	56/56	100%	Constant (present every month)
Ukraine conflict	52/56	92.9%	Constant (present >90% of months)
Immigration	50/56	89.3%	Constant (present >90% of months)
Climate change	47/56	83.9%	High (present in >80% of months)
LGBTQ+ & gender issues	43/56	76.8%	High (present in >75% of months)
AI-generated	40/56	71.4%	High (rapidly growing from March 2023)

2. Expansion in Breadth: The breadth of false and misleading information has expanded significantly.

- The average number of active topics per month increased from 3.0 (2021) to 5.9 (2024–26).
- By the end of the period, five to six thematic categories were active simultaneously, each containing multiple high-volume narratives.
- This reflects a transition from a single-issue environment to a multi-topic system, in which different narratives reinforce and compete with one another for attention.
- The most significant change across the period is the shift from a COVID-dominated landscape in 2021 to a polycrisis landscape in which five to six major categories operate simultaneously. The average number of active topics per month nearly doubled over the five-year period.

Table 4.4 Growth in Concurrent False and Misleading Information Topics

Year	Avg active topics/month	Context

2021 (Jun–Dec, partial)	3.0	COVID-dominated; limited multi-topic activity
2022	4.2	Ukraine war catalyst from Feb 2022; topics multiply
2023	5.1	AI misinformation emerges (Mar 2023); polycrisis begins
2024	5.7	Election cycle; AI deepfakes target multiple countries
2025–26	5.9	Polycrisis peak; AI content reaches 16% share

3. Three Inflection Points: This expansion was driven by three key turning points that reshaped the landscape:

- **February 2022; Ukraine war**: Conflict-related false and misleading information shifted from intermittent to permanent. Narratives around staged attacks, alleged biolabs (e.g., claims that the country hosted covert biological weapons facilities), and political corruption became continuously present rather than event-driven.
- **March 2023; Mainstream adoption of generative AI**: The emergence of AI introduced a new production infrastructure:.
    - AI-generated content rose from marginal presence to 97% of months tracked post-2023. It evolved rapidly from novelty images to targeted political manipulation (e.g. voice cloning, synthetic speeches).
- **2024 onwards; Polycrisis convergence**: Previously distinct domains (health, conflict, immigration, climate, gender, AI) began operating simultaneously and at scale, creating what can be characterised as a "polycrisis information environment", in which these domains not only co-exist but increasingly interact, with narratives crossing boundaries (e.g., public health risks being attributed to migration, particularly in immigration debates).

Table 4.5 Key Inflection Points in the Evolution of False and Misleading Information

Date	Event	Impact on misinformation landscape

Feb 2022	Ukraine war	Conflict misinformation surged from 50% to 100% of months. Introduced sustained narratives on staged attacks, biolab conspiracies and Zelensky corruption that persisted through 2026.
Mar 2023	AI mainstream adoption	AI-generated content prevalence jumped from 28.6% to 97.1% of months. Early deepfake images evolved rapidly to election-targeted voice cloning and synthetic political speeches.
2024–26	Polycrisis convergence	Health, conflict, immigration, climate, AI and gender themes operating simultaneously. Average active topics reached 5.9/month, nearly double the 2021 baseline of 3.0.

4. The Emergence of AI: AI-generated false and misleading information represents the most significant qualitative change in the period reviewed:

- Not tracked as a distinct category before 2023, it became near-ubiquitous (97% of months) thereafter.
- Its share of total content rose to 16% by December 2025
- The technology enabled a shift from static manipulation (images, text) to dynamic, personalised, and scalable deception (audio, video, synthetic personas)
- AI should therefore be understood not simply as a new category, but as a force multiplier across all categories.

Table 4.6 Escalation of AI-Generated False and Misleading Information

Date	Development
Mar 2023	First tracked instances - images only (Trump arrest, Pope jacket, Pentagon explosion)
Sep 2023	First deepfake audio of a political leader
Oct 2023	AI content influences Slovak and Polish elections
2024	Voice cloning deployed in electoral campaigns across multiple countries

Jul 2025	10% share of all content - first record peak
Dec 2025	16% share - highest recorded level

4.2.2 Implications for Policy

Taken together, the evidence points to a fundamental transformation in the nature of false and misleading information:

- **Persistence:** activity is continuous, not event-driven
- **Concurrency:** multiple high-volume narratives operate simultaneously
- **Scalability**: AI enables rapid, low-cost production and adaptation
- **Integration:** domestic and foreign narratives increasingly overlap

This has important policy implications. Approaches based on reactive, topic-specific responses are increasingly ineffective in an environment defined by constant, multi-domain, and technologically amplified information. To aid more effective responses, a comprehensive, wide-ranging, and monthly updated database of false and misleading narratives spreading in the UK is needed, ideally cross-linked to the EDMO data from across Europe.

4.3 Distribution Mechanisms

Section Summary: The distribution of false and misleading information is driven by amplification dynamics embedded in digital platforms and emerging AI systems. Repetition and high-velocity spread create an illusion of truth and consensus, making misleading claims appear credible and widely accepted. Platform algorithms and business models systematically prioritise engagement, amplifying polarising and misleading content, while generative AI introduces new, scalable, and less understood distribution pathways. At the same time, the shift towards personalised and private information environments makes detection and intervention more difficult. These dynamics are reinforced by limited transparency, weak governance, and restricted data access, which hinder accountability and research. Policymakers should therefore

focus on the structural drivers of distribution, particularly algorithmic systems and platform incentives, rather than addressing misleading content in isolation.

4.3.1 Illusory Truth and Illusory Consensus

- Distribution mechanisms are used to spread and amplify false or misleading information leading to a phenomenon known as the illusory truth effect, where repeated exposure to a statement increases its perceived accuracy regardless of its factual basis. Recent research confirms that this effect is robust across varying levels of plausibility and even persists when the information contradicts a person's prior knowledge (Pillai & Fazio, 2025).
- High-velocity amplification methods also give a false impression of prominence, where users start to perceive fringe or false viewpoints as the majority consensus, a state of illusory consensus (Fazio et al., 2022). Consequently, reliable, evidence-based information is frequently "drowned out", not only reducing the accurate interpretation of facts but also delaying care provision and increasing hateful and divisive rhetoric (Borges do Nascimento et al., 2022).

4.3.2 Role of Platforms and AI

- While websites and social media platforms (López-Borrull & Lopezosa, 2025) continue to play a central role in the spread of false or misleading information, the meteoric rise in the role of AI in the online information ecosystem has led to newer, more fragmented, less studied, and more opaque distribution mechanisms (e.g. AI-generated search result overviews (Martins-Rodal & López Bolás, 2026); misinformation in AI chatbot responses (Tiller et al., 2026; Meyrowitsch et al., 2023); false or misleading information in microtargeted online ads; increased persuasiveness of tailored AI-generated content (Carrella et al., 2025; Simchon et al., 2024); and AI-driven influence operation campaigns (Goldstein et al., 2023).
- The core issue is not only the publication and organic spread of false or misleading content (López-Borrull & Lopezosa, 2025), but the harmful ways in which platform algorithms and AI are used to disproportionately amplify it (Bontcheva et al., 2024). Crucially, experts noted that this is not merely a matter of misuse; these platform algorithms and AI are designed and function for that

purpose, a double responsibility that platforms cannot evade. Experts emphasised:

- Content production and distribution are increasingly intertwined. As direct traffic to websites declines, they now serve primarily as algorithmic fodder for engagement-driven platforms. Content is deliberately crafted to be polarising or divisive in order to maximise visibility, generate ad revenue, and drive users toward commercial or deceptive ends. At the same time, such material may also be produced to enter AI training datasets, embedding biased or misleading information into future model outputs.
- AI accelerates this dynamic by making the production of highly optimised, engagement-targeted content significantly cheaper and easier. However, its impact on distribution is, for now, less clear. Unlike social media platforms, generative AI chatbots offer users emotion-based private dyadic interactions, where false and misleading information is not public and thus subject to most well-studied mitigation measures such as fact-checking (Flore, 2025; Meyrowitsch et al., 2023). In the short term, AI primarily expands the supply of harmful content; in the longer term, as chatbots evolve into major attention intermediaries and adopt monetisation models, they may replicate, and potentially intensify, the amplification dynamics seen in existing platforms but in harder to mitigate individual AI-to-user settings.

4.3.3 Algorithmic Bias, Ownership, and Governance

- Related to this is the challenge of algorithmic bias, ownership interference, and ineffective algorithmic governance. The governance of online platform algorithms is exclusively profit- and engagement-driven, where platform owners prioritise shareworthy, emotional, and likely to become viral posts over factual content, leading to promotion of false and misleading information (Hastuti et al., 2025).
- The subject matter experts we consulted also emphasised important documented cases of ownership interference, leading to biased information and algorithmic amplification (Waldman, 2025).
- Policy makers have now also started investigations into the transparency and fairness of platform algorithms and their governance. For example, X's recommender systems—including its "For You" algorithm—and its AI chatbot

Grok are currently under assessment to determine whether they effectively mitigate systemic risks, such as the viral spread of sexually explicit deepfakes and foreign information manipulation (European Commission, 2026). Recent research shows that algorithmic ranking and reranking systems on X can enhance partisan animosity and alter affective polarization (Piccardi et al., 2025), while platform-specific studies also find measurable political biases and differential amplification patterns in content exposure prior to major elections (Prama et al., 2025).

4.3.4 Mainstream Media and Blended Ecosystems

- Increasingly, mainstream media is becoming a distribution mechanism for false and misleading information, due to the eroding distinction between social and mainstream media. For instance, politicians now bypass press conferences to post directly on social media, and traditional outlets increasingly embed social media posts in their reporting (Chadwick, 2017). This interdependence means that false and misleading information can now also gain legitimacy and reach through mainstream news media, creating a complex, blended information ecosystem.

4.3.5 Generative AI as a Scalable Threat

- An even bigger threat is Generative AI, which is a powerful new distribution mechanism that operates on a different scale.
- Firstly, it can be used to create networks of fake websites, not necessarily to attract direct visitors, but to poison the training data of other AI models (Bentzen, 2025b). This demonstrates a shift from simply “deceiving humans” to “deceiving the models” that humans rely on.
- Next, Generative AI can be misused to carry out sophisticated automated influence operations at scale and across multiple platforms (European Union External Action Service, 2025; Goldstein et al., 2023), although further research is necessary on the scale and effectiveness of this threat.
- The use of AI-driven chatbots to run targeted, localised campaigns—such as automated systems generating tailored climate misinformation to influence municipal climate action decisions—and the proliferation of false or misleading AI-generated visual content during the conflict involving Iran has demonstrated a

new level of scale and precision in spreading disinformation, including through psychological microtargeting enabled by AI systems that can infer personality traits from users' digital behaviour and tailor persuasive political messaging accordingly (Simchon et al., 2023; Simchon et al., 2024).

- Even when constrained by safeguards, GenAI chatbots can be easily manipulated (Lopez-López et al., 2025; Tian & Rizoiu, 2024). Looking ahead, the importance of addressing AI as a major and little studied distribution mechanism was emphasised specifically by subject domain experts. The business models of AI development companies remain unclear. While Anthropic may generate revenue by selling large language models to enterprises, it is uncertain how business-to-customer generative AI companies such as OpenAI will generate sufficient returns to justify their substantial investments (Hitzig, 2026). These emerging business models will have significant implications for the creation and distribution of harmful content across the web.

4.3.6 Personalised and Private Information Environments

- Powered by the growth of AI is also the ongoing shift from exposure to false and misleading information in public spaces to exposure in personalised, one-to-one information environments. Recent research (confirmed by our experts) has flagged the newly emerging dangers of AI companions, where conversations are private and inaccessible to fact-checkers, journalists, or regulators (Flore, 2025). Unlike misinformation in public social media feeds, user manipulation and exposure to false and misleading information in these private, personalised, and emotion-driven interactions with AI companions is significantly harder to study (Flore, 2025) and counter with traditional methods such as debunking and prebunking (Manor, 2025).

4.3.7 Barriers to Research and Data Access

- A fundamental and currently largely insurmountable stumbling block (especially for UK scientists) in advancing evidence-based research on the distribution mechanisms (but also on actors, belief, and impact) is the lack of effective social media data access provisions for academic researchers, media organisations, and relevant NGOs (Davidson et al., 2023; van der Linden et al., 2025).

- Even for platforms where data where access is currently available via APIs (e.g. YouTube, TikTok), evidence-based research is currently largely small-scale and based on discreet data snapshots, due to the major computational challenges in large-scale, longitudinal monitoring of information integrity across platforms, languages, and modalities (e.g. during election campaigns). Therefore, academic researchers have argued for the need of a cross-platform information integrity monitoring and analysis computational research infrastructure equipped with state-of-the-art open-source data science tools (Bontcheva et al., 2024). These are needed not only for replicability and transparency reasons, but also to avoid duplication of the already scarce time and money resources of publicly funded researchers.
- Experts we consulted noted the loss of capabilities for real-time monitoring, which are essential for identifying viral misinformation as it happens (e.g. for monitoring political discourse during elections, especially on X - previously Twitter - which is widely used by many political candidates which demands exorbitant fees for data access). Similarly, Meta's shutdown of CrowdTangle, a tool used by fact-checkers and researchers to analyse public content, was cited as a direct move that reduced transparency and accountability on Facebook and Instagram (Arney, 2024; Santini et al., 2025). Crucially, experts noted that Meta's alternative has failed to match CrowdTangle's performance, leaving a critical gap in research and fact-checking capabilities. Scholars argue that this shift represents a move toward “privacy washing”, where platforms restrict data access under the guise of protecting users while actually hiding how their internal algorithms work.

4.3.8 Why Distribution Is So Effective: Psychology and Systemic Vulnerability

- Understanding why these distribution dynamics are so powerful requires looking beyond platform design to the psychological conditions they exploit. The spread and uptake of false and misleading information is increasingly understood through a multi-causal framework combining both cognitive and motivational explanations.
- The cognitive deficit model suggests that individuals fall for false and misleading information simply due to a lack of intelligence or scientific literacy (Choi et al., 2023). However, researchers argue that this framing is insufficient: the spread of

false and misleading information is better understood as a systemic phenomenon driven by a combination of algorithmic design, polarised social environments, and identity-based motivated reasoning (Metzler & Garcia, 2024; Brashier & Schacter, 2020; Choi et al., 2023).

- According to Ecker et al. (2022, 2025), the digital information ecology prioritises engagement over accuracy, creating a "systemic vulnerability" where high-velocity sharing and echo chambers amplify false narratives regardless of an individual's cognitive ability.
- This is further compounded by what Wardle and Derakhshan (2017) describe as "information disorder," a complex framework where social and technical infrastructures interact to erode institutional trust. Rather than a failure of individual critical thinking, the problem is viewed as a mismatch between human evolutionary psychology, which favours group belonging and rapid information processing, and the current high-tech information environment (Lewandowsky et al., 2020). This systemic framing has direct implications for how policymakers approach the actors who operate within these environments.

4.4 Actors in the Creation and Amplification of Misleading Information

Section Summary: The production and amplification of false and misleading information is driven by a wide and expanding range of actors, from state and political figures to corporations, individual content creators, and AI systems, whose financial, political, and psychological incentives frequently reinforce one another. Importantly, the internet's engagement-based business model directly rewards the creation of divisive and misleading content, meaning the problem is as much economic as it is ideological. Policymakers should therefore look beyond regulating individual bad actors and focus on the underlying incentive structures, particularly the algorithmic systems that systematically amplify the most enraging content for profit. Without addressing these structural drivers, interventions targeted at individual creators or pieces of content are unlikely to achieve meaningful impact at scale.

Consultations with subject matter experts highlighted a complex picture of actors whose roles are evolving, driven by a confluence of financial, political, and psychological incentives that often reinforce one another.

- **Organised Denial and Vested Interests as Key Drivers**: A significant portion of science false and misleading information is not accidental but is driven by

organised denial campaigns (e.g., in the case of climate change (Lamb et al., 2020). Experts emphasised that these are often sophisticated operations funded and executed by vested interests, such as the fossil fuel, tobacco, and technology industries, and political actors. These actors are paid to create doubt, using coordinated networks to push a deliberate agenda for corporate financial or political gain (Supran & Oreskes, 2017, 2021; Cook, 2024).

- **State and Political Actors as Creators**: Internationally, some current political administrations and politicians were identified as primary creators of disinformation through Foreign Information Manipulation and Interference (FIMI) (EEAS, 2025). Their ability to amplify the salience of information which is potentially false or misleading grants them immense authority and reach, blurring the line between political communication and disinformation. Moreover, false narratives from foreign propaganda campaigns are now amplified via AI to reach millions of views on social media (Huet et al., 2025) and are also repeated to citizens by leading generative AI chatbots (Huet et al., 2025).

- **The Emergence of Non-Traditional and Non-Political Actors**: The ecosystem of actors is expanding beyond politicians and corporations. A key insight was that a significant volume of misleading content is generated by individuals seeking attention (Morosoli & Humprecht, 2025), a powerful and not always political incentive. This includes online content creators who have learned that polarising or outrageous content is the most effective way to build an audience and monetise their presence (Hiaeshutter-Rice et al., 2021; Combrink & Mkungeka, 2025).

- **AI is a major actor through chatbots, AI Agents and AI-generated search results:** Please see section 4.3.

- **Incentive Structures of the Digital Economy**: A critical insight is that the business model of the internet directly incentivises the creation of false and misleading information, due to a so-called *Disinformation Economy* (Scholtens et al., 2024). Research has estimated that nearly US$6 billion worldwide in digital advertising budgets are redirected to fake news websites (Skibinski, 2021). Monetisation programmes on platforms reward engagement with advertising revenue, creating a market for hate speech (Center for Countering Digital Hate, 2024a), AI slop (i.e., low-quality, mass-produced content generated by AI), deliberately misleading content, and low-quality content that prioritises speed and

engagement over factuality. In some cases, highly emotive or polarising content, such as misogynistic narratives (e.g., as featured in the Louis Theroux documentary: Inside the Manosphere), functions as an attention-grabbing gateway, building large audiences that can then be monetised through subscriptions, scam training programmes, or more overt scams (e.g., investment or cryptocurrency schemes). This highlights how misinformation is often embedded within broader commercial strategies designed to capture and convert attention. This financial incentive system is so effective that platforms such as X (Roeloffs, 2023) adopted demonetisation policies, signalling the severity of the problem. Evidence has emerged however that the community notes demonetisation approach adopted by X falls short on addressing the problem effectively, due to being too slow or inconsistent (Center for Countering Digital Hate, 2024b).

- **Interplay and Reinforcement of Incentives**: The incentives are not isolated but often work in synergy. A content creator seeking attention (a psychological incentive), may produce politically charged content (a political incentive), that generates high engagement, which in turn yields advertising revenue (a financial incentive). Algorithms then amplify this highly engaging content, creating a feedback loop that rewards and reinforces the spread of the most divisive and misleading material.

5 Findings: Assessing the Impacts

This section synthesises review-level evidence and expert consultation on the impacts of false and misleading information. The review evidence provides a broad overview of the landscape, while the expert insights offer more specific, in-depth perspectives. It maps the landscape of reported harms, identifies key gaps in the evidence, and highlights the principal challenges and priorities for research and policy.

Section Summary: The evidence and expert insights highlight that false and misleading information produces multi-level, interconnected harms affecting individuals, society, and organisations. Existing review-level research is heavily skewed toward health misinformation, particularly COVID-19, leaving significant gaps in understanding non-health domains such as climate, politics, economic decision-making, and impacts on marginalised groups. Harms manifest as behavioural, reputational, financial, and systemic effects, often amplified by algorithmic environments, polarised social contexts, and motivated reasoning, rather than individual cognitive deficits alone. Experts emphasise that trust, both in institutions and information channels, is central to mitigating harm, with failures in communication, transparency, and local journalism exacerbating vulnerabilities. Emerging technologies, particularly AI-generated content, and adaptive misinformation actors complicate detection and policy responses, highlighting the need for flexible, evidence-informed strategies. Methodological and data limitations constrain measurement of harm, making prioritisation of policy interventions challenging but critical, particularly where harms are severe, measurable, and linked to tangible outcomes such as public health, political violence, or economic fraud.

5.1 Insights from Review-Level Evidence

Of the 228 reviews and evidence syntheses identified via our search, 159 (70%) focused on ‘Assessing Impacts’, making it the most extensively covered theme in the review. Below we summarise evidence according to individual, societal, and organisational harms.

It is important to note that much of the research described below on harms comes from health and COVID-19 contexts. As such, what has been studied is not necessarily the same as what matters most. The concentration of review-level research in these areas means that policymakers working in non-health domains, including climate, democratic

integrity, economic harms, and impacts on marginalised communities, are operating with a substantially thinner evidence base. This reflects disparities in research attention rather than the relative significance of these harms. These limitations are discussed further in Section 5.2.

5.1.1 Individual-Level Harms

- At the individual level, the most consistently documented harms are psychological and behavioural. A scoping review by Delgado et al. (2021), covering 33 articles on infodemic effects during COVID-19, found that the most common mental health consequences were anxiety, depression, and stress, with young adults and women most affected. This was supported by a systematic review by Dastani and Atarodi (2022), examining 17 studies across eight countries, found that people more exposed to COVID-19-related (mis/dis)information were more prone to psychological consequences including anxiety, depression, and stress.

- Beyond mental health, Schmid, Altay, and Scherer (2023), in a systematic review of 64 randomised controlled trials (N = 37,552), found that in 49% of cases, exposure to health misinformation damaged the psychological antecedents of health behaviours, including knowledge, attitudes, and behavioural intentions.

- In clinical contexts, Bariş et al. (2024) identified that health misinformation causes harm by leading to vaccine hesitancy and the adoption of unproven disease treatments, with secondary societal consequences including increased hate speech towards ethnic groups and medical experts.

- Specific clinical examples are stark: Birkun (2024), in a review of 44 studies on CPR and first aid misinformation, found that 43 studies (97.7%) reported omissions or errors in publicly available guidance, and 11 studies (25%) identified advice that could cause direct injury or death. Similarly, Finnegan et al. (2023) found that topical corticosteroid misinformation creates a complex phenomenon that can obstruct successful treatment, with patients abandoning evidence-based care in favour of unproven alternatives promoted by commercial interests.

- Vaccine hesitancy represents one of the most extensively documented individual-level harms. The WHO named vaccine hesitancy one of the top ten threats to global health in 2019 (Löffler, 2021). Zhao et al. (2023), synthesising

91 observational and 11 interventional studies, found that COVID-19 vaccine misinformation prevalence ranged from 2.5% to 55.4% in the general population, rising to 6.0% to 96.7% among vaccine-hesitant groups, illustrating how misinformation exposure concentrates among already-vulnerable populations.

- The HPV vaccine offers a concrete illustration of the downstream consequence: Wang et al. (2023) note that the CDC estimates that almost 92% of HPV-attributable cancers in the United States between 2013 and 2017 could have been prevented by the available vaccine, yet uptake remains low compared with other childhood vaccines, a gap in which misinformation plays a documented role.

5.1.2 Societal-Level Harms

- At the societal level, the review-level evidence documents harm to institutional trust, democratic functioning, and community cohesion.
- Ries (2022), in a review of reviews on the COVID-19 infodemic, found that the worldwide infodemic is recognised as a multifaceted problem extending beyond health and human rights into global political spheres, causing stress, deception, violence, and harm, and that building and maintaining trust is the most important countermeasure.
- Arcos et al. (2022), reviewing evidence on disinformation as a component of hybrid threats, found that purposely deceitful content targeting political aims or economic purposes is increasingly recognised as a major security threat, particularly following evidence of hostile foreign interference in democratic processes.
- Alkhair et al. (2023), applying a socio-ecological model to COVID-19 misinformation across 11 studies, found documented impacts spanning individual, interpersonal, organisational, community, and policy levels simultaneously, underscoring the systemic rather than isolated nature of harm.
- Al Arfaj et al. (2025) identify five categories of misinformation affecting minority youth specifically; health-related, educational, political, cultural/identity-based, and financial, and find that cultural factors including trust dynamics, language barriers, and digital literacy disparities create distinct vulnerability patterns not captured by generic research.

- Kemei et al. (2022) document that approximately 96% of Canadians and 80% of Americans reported encountering COVID-19 disinformation on social media, with Black communities disproportionately affected and misinformation contributing to both higher infection rates and lower vaccination uptake.

5.1.3 Organisational-Level Harms

- At the organisational level, the evidence is thinner than for individual and societal harms, but points to a range of documented impacts spanning health system disruption, corporate reputational damage, financial market instability, and threats to democratic institutions.
- Within health systems, Borges do Nascimento et al. (2022) identify misallocation of health resources as one of the most negative consequences of health-related false and misleading information, alongside mental health impacts and increased vaccination hesitancy. Abuhaloob et al. (2024), reviewing infodemic management responses across 29 studies, found that infodemics can impact people's risk perceptions, trust, and confidence in health systems and health workers, with knock-on effects for organisational capacity to respond effectively.
- Beyond health, the evidence base (while still developing at the review level) points to significant organisational harms across multiple sectors. Malhotra et al. (2025), in a systematic literature review examining how false and misleading information affects workplaces, found impacts on organisational outcomes including worker productivity, engagement, and job satisfaction.
- On climate, Gertrudix et al. (2024) emphasise the need to understand disinformation's financial, economic, and political roots and its disruptive effects through hidden networks of influence.
- Küçükkocaoğlu and Özden (2025) found that information disorders, particularly rumours spread via social media, lead to market fluctuations and irrational investment decision-making, a harm that remains largely unexamined at the review level.

5.1.4 Limitations of the Evidence Base

- Reviews of studies assessing the impacts of false and misleading information are heavily skewed toward health misinformation (61% of all reviews), and within

health, toward COVID-19 specifically (36%), likely due to the global nature of the problem and also reflecting a surge in research during the pandemic.

- Methods and studies assessing the impact of false and misleading information in non-health domains (11%), including climate, organisational, democratic, and economic contexts, as well as among vulnerable demographic sub-groups (3%), such as ethnic minorities, LGBTQ+ populations, and scientists, are substantially understudied.
- Individual-level behavioural harms are also underrepresented (10%), including offline impacts and economic decision-making, alongside open methodological challenges (5%), such as impact measurement, restricted data access, and algorithmic opacity. These areas represent key priorities for future research.
- The lack of research on the impact of reputational falsehoods, political/event-driven misinformation, crisis misinformation, climate misinformation, and gendered disinformation suggests either a lack of review-level synthesis in these areas or a genuine gap in research.

5.2 Insights from Expert Consultation

Discussions with experts reinforced that misinformation is not a single, isolated problem but a systemic challenge with multi-level, interconnected harms (Bentzen, 2025b). Effective policy responses must therefore adopt a "whole-of-society approach" (Bentzen, 2025a), be adaptable, combining robust evidence with strategies to rebuild trust in information systems.

5.2.1 Harms Are Multi-Level and Interconnected

Harms from false or misleading information span three distinct but interconnected levels, each requiring different policy responses:

- **Individual harms** are the most direct and easy to evidence. It's important to take a measured approach that identifies when and among whom such information does have (negative) impacts (Ecker et al., 2025). Individual harms include preventable deaths from false health claims (e.g. about the effectiveness of cancer treatments or vaccines), and financial losses from fraud. However, while there are clear examples of harm that can be traced back to false and misleading information, this needs to be taken in the context of extensive evidence that it is

more difficult to persuade people of information that is inconsistent with their beliefs (i.e., the confirmation bias; e.g., Altay et al., 2023).

- **Societal harms** of false and misleading information include erosion of trust in science, media, and public institutions (Hyzen, 2026, Essien, 2025); increased political polarisation (Bennett & Livingston, 2018); and a broader breakdown in shared standards of truth (described by experts as 'epistemic chaos') (Terzian, 2025). Declining trust also carries economic costs, as low-trust environments increase transaction and governance costs in markets.

- **Organisational harms** affect businesses, public institutions, and democratic systems, especially in times of crisis. These include disruption to industries (for example, renewable energy projects cancelled due to misinformation), reputational damage, and interference with democratic processes.

5.2.2 Key Domains and Underserved Harms

While health misinformation is well studied, several domains require greater policy attention:

- **Climate and environment**: Misinformation is increasingly focused on undermining strategies for mitigating climate change rather than denying climate change outright (Lamb et al., 2020). Media narratives often decouple policies (e.g. net zero) from climate science, creating public confusion. Targeted misinformation campaigns against technologies (e.g., EVs, heat pumps) and the use of AI tools to influence local decision making represent emerging risks. However, while climate delay has become a dominant framing, there are signs of a resurgence of outright denial. Also, the Trump administration has emboldened denial arguments that were sidelined.

- **Gender and political participation**: Female public figures, particularly politicians and journalists, are disproportionately targeted by gendered disinformation (Gehrke & Amit-Danhi, 2025; Posetti et al., 2021; Rheault et al., 2019). This contributes to self-censorship, discourages political participation, and reinforces structural underrepresentation.

- **Offline harms**: Online misinformation can translate into real-world consequences, including health, violence, harassment, and illegal behaviour, particularly in relation to migration and gender issues (Purnat et al., 2025; Hameleers, 2026).

- **Targeting of marginalised groups:** Migrants and LGBTQ+ communities are targeted by false and misleading information, which can fuel discrimination, hate speech, and physical violence, further marginalising already vulnerable populations (Szakács & Bognár, 2021; Irfan et al., 2021).
- **Targeting of experts**: Scientists, academics, and fact-checkers are frequent targets of abuse and intimidation. This leads to self-censorship and may undermine research quality and public communication (Brysse et al., 2013; Lewandowsky et al., 2015)
- **Economic harms to individuals**: Fraud and scams linked to disinformation represent a significant and tangible harm, often serving as a key entry point for public concern. For example, when Full Fact conducted focus groups with Ipsos in 2024 and asked what misinformation meant to people, fraud and scams was by the far the most cited type of example (Mortimer, 2024).

5.2.3 Evidence and Measurement Challenges

A central policy challenge emphasised by experts is demonstrating causal links between false and misleading information and harm at scale.

- **Evidentiary and regulatory tension:** While some harms are well supported, others rely on limited or anecdotal evidence. Some experts cautioned against assuming that misinformation is necessarily more harmful today than in the past, and noted that the scale and nature of impact remain debated. However, a growing body of empirical research provides robust causal evidence that exposure to misinformation and partisan media content can meaningfully shape attitudes and behaviours, including increases in hate crimes, shifts in politically relevant beliefs, and changes in public health compliance (Lorenz-Spreen et al., 2022; Ecker et al., 2024). Instrumental-variable and quasi-experimental studies have shown that exposure to partisan media and misinformation can increase hate crimes (Müller & Schwarz, 2021), alter factual beliefs and policy-relevant opinions (Bursztyn et al., 2023), and affect behavioural compliance during the COVID-19 pandemic (Simonov et al., 2020).
- Some policymakers and regulators dismiss harms due to lack of definitive evidence. It is important to clarify that gaps in evidence do not necessarily mean harms do not exist; rather, they can be difficult to measure; either because

access to data is limited, constraining what can be proven, or because ethical and practical constraints limit what can be studied.

- **Data access is a major constraint**: Restricted access to platform data limits the ability to measure the exposure, spread, and impact of false and misleading information—particularly across languages and platforms (Davidson et al., 2023; van der Linden et al., 2025), as does the lack of a computational research infrastructure for large-scale, cross-modal information integrity monitoring (Bontcheva et al., 2024).
- **Methodological gaps persist:** There is no standardised framework for measuring the negative impact of false and misleading information, and many studies are small scale or ad hoc. For example, researchers cannot randomly expose half the population to disinformation during a real election to test its effects or to ask social media users about their actual voting decision. Therefore, understanding harms requires triangulating different forms of evidence, often from incomplete information.

- **Emerging opacity**: The rise of personalised, AI-generated content reduces visibility, making monitoring and attribution increasingly difficult. See section 4.3 for details.

5.2.4 Trust and Communication as Central Mechanisms

False and misleading information can erode trust and undermine the conditions for effective collective decision-making.

- **Institutional trust is a critical asset** (Parr, 2025): Attacks on science, media, and government institutions can undermine the foundations of evidence-based policymaking. In some countries, particularly those with authoritarian governments, this extends to a broader erosion of trust in civil society organisations, non-governmental organisations, and other independent institutions that play a vital role in information integrity and accountability.

- **Democratic systems at risk**: Beyond discrete policy impacts, false and misleading information often poses a systemic risk to democracy by undermining shared standards of truth and the possibility of informed public deliberation (Bentzen, 2025b; Sato & Wiebrecht, 2024).

- **Crisis communication failures amplify harm**: Lack of transparency, inconsistent messaging, and dismissal of public concerns can contribute to distrust (Sauer et al., 2021). Such failures can also create information voids (Purnat et al., 2021; Combrink & Mkungeka, 2025), gaps in reliable, authoritative information that are rapidly filled by false or misleading content, further eroding public confidence and complicating response efforts.
- **Engagement over coercion**: Policies that fail to address public concerns risk reinforcing mistrust, even when compliance is achieved.
- **Decline of independent journalism**: Economic pressures on journalism (including paywalls) are increasingly reducing the availability of high-quality information, enabling the proliferation of false and misleading information which instead spreads and gains prominence in such information voids and news deserts (Torre et al., 2024; Wouters & Opgenhaffen, 2024). This is particularly evident in the decline of local journalism (Neff & Pickard, 2023), including BBC local radio. If people are to trust information about events on the other side of the world, the best place to start is by ensuring they trust reporting on what is happening on their own street (Wouters & Opgenhaffen, 2024).

5.2.5 Evolving Threat Landscape

The information landscape is adaptive and increasingly complex:

- **Artificial Intelligence:** AI systems enable the creation of personalised, scalable misinformation across text, audio, and video—each with distinct risks (Bontcheva et al., 2024). Text-based AI can produce high volumes of plausible content at low cost. AI-generated audio can impersonate trusted figures, enabling fraud and manipulation. AI-generated video (deepfakes) presents acute risks due to its high perceived credibility. Together, these shift harms from public, observable spaces to private, opaque environments, complicating detection and response. It can also disrupt the economic model of the information ecosystem, particularly advertising-based funding.
- **Coordinated and cross-domain actors:** Actors spreading misinformation are highly adaptive, often shifting across topics (e.g. from COVID-19 to climate or geopolitical conflicts). This challenges siloed policy and research approaches.

5.3 Open Questions and Debates

5.3.1 Prioritising Policy Responses: A Harms-Based Approach

Some subject matter experts are concerned about whether and how to prioritise policy responses to the diverse harms arising from false and misleading information. While there was broad agreement that the evidence base and expert consultation offer a descriptive map of harms, experts differed on the feasibility and utility of ranking them for intervention.

- **Challenges in ranking harms**. Some experts cautioned that creating a hierarchy of harms is inherently subjective. They noted that policymakers' priorities will inevitably vary depending on their specific portfolios and responsibilities, and that harms are often interconnected, as discussed in Section 5.1 and 5.2, making neat categorisation difficult.
- **The case for prioritisation**. Other experts argued that failing to distinguish between harms by severity and impact risks diffusing policy attention and resources. While acknowledging the existence of grey areas, they contended that this should not prevent differentiation where possible. Some harms, they noted, are clearly more severe and measurable than others, making prioritisation both possible and necessary (Ecker et al. 2025). For example, politically charged misinformation has been linked to threats of violence or actual harm (Full Fact Report, 2025), and false health information has resulted in life-threatening outcomes, such as the Paloma Shemirani case (RCNi, 2025). In these instances, the case for prioritised intervention is compelling.

6 Findings: Countering and Mitigating False and Misleading Information

Of the 228 reviews and evidence syntheses identified via our search, 116 articles address some aspects of countering and mitigating false and misleading information. The evidence spans a wide range of intervention types, from individual-level, psychological approaches through to platform governance and automated detection systems. Below, we organise evidence and insights on countering and mitigating false and misleading information into:

- **Individual-level interventions**, which focus on strengthening people's ability to critically evaluate information and make informed decisions about what they engage with and share.
- **System-level interventions**, which focus on changing the structures, platforms, and supply chains through which false and misleading information travels.

Table 6.1 (provided at the end of section 6) provides a summary of interventions for countering and mitigating false and misleading information, alongside key summaries from the evidence, limitations of the approaches discussed, factors that may moderate their effectiveness, and specific design features that have been considered.

6.1 Summary of Findings on Countering and Mitigation

This section synthesises evidence from the review and expert consultations, identifying key findings, implications for policy, and priorities for future research.

6.1. Individual- and System-Level Interventions: Interactions and Trade-offs

Section Summary: Policy responses should adopt a portfolio approach that combines individual- and system-level interventions, while rebalancing effort toward addressing structural drivers. At the individual level, this includes continued investment in prebunking, debunking, and media and digital literacy. At the system level, stronger regulatory and governance measures targeting platform design and algorithmic recommender systems. In particular, policymakers should prioritise interventions that address engagement-driven incentives, including requirements for platforms to incorporate and transparently operationalise independent quality signals in content

ranking. Without such system-level measures, interventions focused solely on individual users are unlikely to achieve meaningful impact at scale.

- **No single intervention is sufficient:** There is strong agreement that no single approach can effectively counter false and misleading information at scale. Effective responses require layered, coordinated strategies, operating across both the individual level (how people access, interpret and share information) and the system level (how platforms and information ecosystems are governed) (van der Linden et al., 2025; Aghajari et al., 2023; Sanfilippo et al., 2025; Hartwig et al., 2024). Research has also shown differences in opinion on mitigation interventions, namely academics place more responsibility on individual media literacy while fact-checkers prioritize platform accountability (Weikmann et al., 2026).
- **Overemphasis on individual-level approaches:** Research and policy responses have largely focused on demand-side interventions, such as media literacy, public awareness campaigns, and fact checking, which target how individuals evaluate false and misleading information. While important, these do not address key structural drivers, including algorithmic amplification, weak platform accountability, and the ongoing, coordinated production of false and misleading information by both state and non-state actors. Aghajari et al. (2023) highlight this imbalance, noting that research disproportionately targets individual-level solutions despite growing evidence that systemic factors are primary drivers. Expert consultations reinforced this view, framing false and misleading information as a systemic issue rooted in platform incentives, the attention economy, and engagement-driven design.
- **Interaction between individual- and system- level approaches:** While sometimes presented as distinct, the boundary between individual- and system-level interventions is not always clear-cut. Many approaches operate across both levels or are mutually reinforcing. For example, platform-based fact-checking programmes (e.g., on Meta (in the process of being discontinued and replaced by a community approach) and TikTok) rely on individual-level assessments but are implemented at the system level through labelling, downranking, or content moderation. Similarly, community-driven initiatives such as crowd-sourced verification systems (e.g. community notes on X (Drolsbach et al., 2024; Center for Countering Digital Hate, 2024b) draw heavily on fact-checking outputs and user contributions (Razuvayevskaya & Bontcheva,

2026), blurring the distinction between individual input and system-level impact. More broadly, efforts to improve platform accountability, investigate platform dynamics, and disrupt the supply of false and misleading information often depend on data, expertise, and signals generated at the individual level.

- **Comparing effectiveness and trade-offs:** System-level interventions may be more effective at addressing false and misleading information at scale, but raise concerns around freedom of expression and require cooperation from technology companies (van der Linden et al., 2025). In contrast, individual-level interventions are easier to implement and pose fewer risks, but are unlikely to keep pace with the scale and speed at which false and misleading information is produced (Roozenbeek et al., 2023; Sanfilippo et al., 2025; Aghajari et al., 2023).
- **Need for stronger system-level interventions:** There was strong agreement among experts that more robust system-level interventions are necessary to address underlying incentive structures. This includes regulatory approaches targeting: (i) design features that promote compulsive or high-frequency engagement (e.g., infinite scroll, where content loads continuously without a stopping point, and “streak” features that reward repeated daily use), and (ii) algorithmic recommender systems in consumer-facing products. In particular, experts emphasised the need to require platforms that rely on engagement-based ranking systems to incorporate independent or third-party quality signals into content ranking and recommendation processes. To ensure these measures have substantive impact rather than resulting in symbolic compliance, such requirements would need to go beyond voluntary approaches and include obligations to disclose how quality signals are weighted and operationalised in practice. Without this, platforms could nominally adopt external indicators while assigning them negligible influence. Strengthening transparency and enforceability in this area is therefore critical to counteracting structural incentives that prioritise engagement over content quality.

6.1.2 Areas of Stronger Evidence and Consensus

Despite the overall complexity of the landscape, the evidence supports a number of conclusions with reasonable confidence.

Section Summary: No single intervention is sufficient to address false and misleading information; effectiveness depends on combining complementary approaches.

Prebunking is a robust individual-level strategy and should serve as a first line of defence, but it cannot anticipate all narratives, meaning debunking and fact-checking remain essential backstops despite their limits. Media literacy builds long-term resilience but operates too slowly to address acute harms and remains difficult to scale and evaluate. At the system level, voluntary platform self-regulation is insufficient, requiring enforceable governance, independent auditing, and meaningful data access. Overall, an effective response must be sustained, adaptive, and layered across prevention, correction, education, and regulation, in a whole-of-society approach.

1. Prebunking as a first line of defence, supported by debunking

- Prebunking refers to a broad category of strategies aimed at stopping people from accepting false and misleading information before they encounter it. The most widely employed of these strategies is psychological inoculation, the core idea is that presenting someone with a weakened form of a misleading claim builds psychological resistance against future manipulation. Psychological inoculation typically consists of two elements: alerting someone in advance that their beliefs may come under attack (e.g., "warming: someone might try to influence you by claiming X"), and offering a preemptive counter-argument that undermines the false claim before it takes hold (e.g., "this is misleading, because Y"; van der Linden, 2025)

- Van der Linden et al. (2025) describe prebunking as a “first line of defence,” on the grounds that preventing belief formation is generally more effective than attempting to correct beliefs once established (i.e., prevention is better than cure).

- The evidence for prebunking has strengthened considerably in recent years and represents one of the more promising individual-level approaches. Lu et al. (2023), in a systematic review and meta-analysis of 42 studies (42,530 participants), found that psychological inoculation significantly reduces misinformation credibility assessment, improves real information credibility assessment, and enhances both credibility discernment and sharing discernment. However, it is important to note that psychological inoculation did not significantly influence misinformation sharing intention, suggesting its effects

on stopping people from actively spreading misinformation are less established than its effects on belief and discernment.

- Similarly, evidence from a meta-analysis of over 37,000 participants shows that these interventions improve veracity discernment (i.e., the ability to distinguish truth from falsehood) without increasing response bias, meaning they do not make individuals uniformly more skeptical of all news (Simchon et al., 2026).
- Game-based inoculation interventions also show promise. Kiili et al. (2024), reviewing 15 studies, found that games grounded in inoculation theory, designed to expose players to weakened doses of manipulation techniques, reported positive outcomes, though the authors caution that findings cannot yet be generalised given the immaturity of this research area.
- However, prebunking is not a complete solution. Its effects decay over time, with evidence suggesting they can persist for up to three months when reinforced with brief booster reminders (Maertens et al., 2025; Van der Linden et al., 2025). This has clear implications for programme design: one-off interventions are insufficient; sustained, periodic reinforcement is required to maintain protective effects.
- Crucially, not all misinformation can be anticipated. Even the most comprehensive prebunking strategies will leave significant volumes of false and misleading content unaddressed. For these cases, particularly the most harmful or fast-moving narratives, debunking remains essential. There is a risk that strong emphasis on prebunking could inadvertently sideline debunking, but this would be a mistake.
- Rather than competing approaches, prebunking and debunking are mutually reinforcing. Effective prebunking depends on identifying emerging manipulation techniques and narratives, which in turn requires a substantial and continuously updated corpus of debunking. At the same time, debunking provides a necessary backstop, addressing misinformation that escapes pre-emptive efforts. Together, they form a complementary toolkit: prebunking to reduce susceptibility at scale, and debunking to correct what inevitably gets through (Bruns et al., 2024).

2. Debunking and corrections are commonly used, but cannot eliminate misperceptions

- Post-hoc correction is broadly effective and remains the most widely deployed intervention (Francis et al., 2025).
- However, corrections do not fully eliminate misperceptions. The "continued influence effect", where misinformation continues to shape beliefs and behaviour even after individuals have understood and acknowledged a correction, is robust and well-replicated (Ecker et al., 2022). Lewandowsky et al. (2012) document this, noting that retractions rarely if ever fully eliminate reliance on misinformation.
- Chan et al. (2017), in a meta-analysis of 52 experimental samples (N = 6,878), found that corrections are more effective when they provide a replacement explanation, giving recipients something accurate to believe in place of the false claim, rather than simply labelling the misinformation as incorrect. Lewandowsky et al. (2012) suggest why: without a replacement, people are left with a gap in their understanding that the original false belief continues to fill. They also note that corrections are more likely to succeed when they come from a credible source and fit with what the audience already believes. Complementary research also highlights that, where appropriate, debunking can be effective by discrediting unreliable or misleading sources, thereby reducing the influence of misinformation (Ecker et al., 2024).
- Martel and Rand (2023) review the evidence on warning labels and conclude they are broadly effective at reducing belief in and sharing of misinformation, with effects generally consistent across party lines. However, the specificity and design of labels matters: general warnings (e.g., "It is important to remain skeptical when reading headlines") can inadvertently reduce belief in true as well as false content, whereas more precise tags such as 'Rated false' outperform vaguer ones like 'Disputed.' Combining warning labels with more detailed fact-checks can also improve effectiveness, though coverage remains a key challenge, unlabeled false content may actually be perceived as more credible in the presence of labeled content (the 'implied truth effect'). Similar findings have been observed in community notes (Drolsbach et al., 2024).
- However, labelling should not be understood as independent of fact-checking. In practice, fact-checking is the precondition for labelling, providing the evidentiary basis for claim classification, matching, and intervention. In platform contexts (e.g., Meta’s third-party fact-checking programme), labels are typically

accompanied by additional measures such as downranking and claim matching, which reduce the visibility and recirculation of flagged content. There is also emerging evidence of second-order effects: exposure to fact-checks can reduce the likelihood that accounts share misinformation in the future (e.g., Cagé et al., 2026).

- At the same time, expert consultations highlight important limitations of detection-and-warning approaches, though these vary by label type. For accuracy-based labels (e.g., content labelled as false or misleading), there is strong evidence that they can reduce belief and sharing, particularly when paired with explanations or alternative information. However, more persistent challenges arise for other forms of labelling, especially disclosures about synthetic or manipulated media (e.g., “this video is AI-generated” or a deepfake). In these cases, the persuasive or emotional impact of the content may not be fully mitigated by the label, particularly if users engage with the content before noticing the warning or if the label does not directly address the implied claim.
- This evidence suggests that while corrections, fact-checking, and labelling are necessary components of the response, they are not sufficient on their own. Detection and warning systems should be understood as one layer in a broader strategy, rather than a primary line of defence.

3. Media literacy works when designed well, but reach and long-term impact remain significant challenges.

- Media and digital literacy education shows consistent benefits on knowledge and critical evaluation skills, with the strongest evidence for detection ability. Droog et al. (2025), in a systematic review of 80 experimental studies, found that most interventions successfully improved users' ability to detect misinformation. However, their effects on more distal persuasive outcomes, such as attitudes, were more inconsistent, suggesting that changing deeper beliefs and behaviours may require different or additional strategies beyond detection training alone. Notably, effectiveness depended more on the outcome variables targeted than on specific intervention characteristics, which has important implications for how programmes are designed and evaluated.
- Susceptibility to misinformation is itself shaped by literacy. Nan et al. (2022), reviewing 47 publications, found that subject knowledge, literacy and numeracy,

and analytical thinking confer strong resistance to health misinformation, providing a rationale for literacy investment as a protective strategy. Cultural and contextual factors also influence outcomes, and interventions must be tailored accordingly (Bhattacharya & Singh, 2025).

- A recent review by van der Linden et al. (2025) picks out several important nuances in the media literacy evidence base:
    - Longer interventions work better. Bergsma and Carney (2008) found that programmes of five hours or more consistently outperformed shorter exposures, with brief interventions largely failing to produce measurable change.
    - Gains fade quickly. Studies with extended follow-up, including Guess et al. (2020) and McGrew et al. (2019), found that improvements declined after the immediate post-intervention period, and Stassen et al. (2020) found no measurable effect at all by six months.
    - It is also key to be able to measure the impact of media literacy initiatives: most studies rely on short-term assessments of knowledge or evaluation skills, while robust and consistent measures of long-term behavioural change remain limited.

- Literacy programmes are essential for building long-term societal resilience but should not be treated as a fast-acting counter to acute false and misleading information threats. Further research is also needed to demonstrate their effectiveness for AI-generated false or misleading information (Feuerriegel et al., 2023).

4. Platform self-regulation underdelivers, and the regulatory picture is worsening

- The evidence on platform governance is now fairly unambiguous: voluntary self-regulation and platform commitments, without meaningful enforcement mechanisms, have failed to deliver consistent or verifiable outcomes.

- The most direct empirical evidence comes from Mündges & Park (2024), who examined compliance reports submitted by Google, Meta, Microsoft, TikTok, and X (Twitter) under the EU's Strengthened Code of Practice on Disinformation. The average compliance score was 1.9 out of 3. Over half of qualitative reporting requirements were incomplete or irrelevant, and nearly two-thirds of quantitative data points were missing or methodologically unsound. The section requiring

platforms to provide data access for independent researchers was the worst-performing area of all. Crucially, the indicators intended to measure whether the Code was actually reducing disinformation in practice have never been developed, meaning there is no mechanism for assessing real-world impact. The authors also note that the Code is better understood as a co-regulatory instrument than a purely voluntary one, given it was driven by strong institutional pressure from the European Commission, and that the direction of travel in EU policy is clearly toward enforceable obligations.

- D'Andrea et al. (2025) trace this regulatory evolution and note that binding EU legislation has begun making transparency reporting more systematic. However, a key enforcement mechanism, independent audit of platform algorithms, was not yet operational at the time of writing. The authors also note that the closure of a major research data platform without a viable replacement prompted a Commission investigation into whether one of the largest platforms had violated its legal obligations. This illustrates a wider problem: the data infrastructure needed to assess compliance can itself be withdrawn by the platforms being regulated.

- This is consistent with independent monitoring by EDMO, which identifies uneven implementation of commitments on media literacy, research access, and fact-checking, as well as persistent gaps in the quality, comparability, and verifiability of platform reporting (EDMO, 2024).

- More recent EDMO evaluations of Very Large Online Platforms (VLOPs) further highlight limitations in both compliance and effectiveness: while platforms report high levels of activity, there is limited evidence of measurable impact, and key data necessary for independent scrutiny remain restricted (EDMO, 2025). Taken together, this suggests that without stronger oversight, standardised reporting, and enforceable obligations, voluntary frameworks are unlikely to deliver consistent or accountable outcomes.

- Sanfilippo et al. (2025), reviewing nearly three decades of governance literature, conclude that effective governance must address the design of information systems, including algorithmic ranking and amplification, not just the content they carry. They find that platform policies are frequently non-enforceable in practice, and that relatively few successful governance interventions yet exist at scale.

- Stieglitz et al. (2025) add that very few prevention measures in the research literature are designed specifically for government agencies, with most targeted at platform operators, which limits public sector capacity to act independently of platform cooperation.

- Expert consultations reinforce these findings in two important respects. First, there is broad consensus that while regulation is a necessary lever, current enforcement remains weak and real-world impact limited. Second, and more concerning, experts report that researcher access to platform data has deteriorated rather than improved in the period following the Digital Services Act. Since 2022–23, major platforms have tightened access to their APIs (i.e., the tools researchers use to retrieve platform data systematically) making it much harder for independent researchers to study online harms and assess whether regulation is working. This creates a clear paradox: regulatory obligations are expanding at the same time as the evidence base required to assess them is being constrained.

5. Platform regulation efforts in the UK lag behind those in the EU and elsewhere

- While the EU has built a comprehensive regulatory framework to address Gen AI-enabled information manipulation, the UK lags behind. The UK's main tool, the Online Safety Act 2023 (OSA), only came fully into force in mid-2025, and contains no provisions specifically targeting AI-generated content. The UK Parliament's own Science, Innovation and Technology Committee warned that the Online Safety Act was “riddled with gaps - including its failure to explicitly regulate generative AI” (UK Parliament, 2026). By contrast, the EU's AI Act requires deepfakes to be labelled, and the DSA Code of Conduct on Disinformation (European Commission, 2025) obliges major platforms to assess and mitigate risks from disinformation.

- In addition, the Online Safety Act fails to mandate access for vetted researchers to data from Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs), unlike the DSA which has established dedicated contact points for researcher data access called Digital Services Coordinators (DSCs). The EU regulatory frameworks also benefit from significant investment in monitoring and enforcement mechanisms, such as the European Digital Media Observatory (EDMO), a dedicated EU AI Office, a comprehensive network of European funded fact-checking organisations, independent auditors, a Transparency

Centre, and service level metrics. In June 2026, the EU will also publish the final version of a voluntary Code of Practice on Transparency of AI-Generated Content (Bontcheva et al., 2026), which is currently scheduled to come into force in December 2026. The UK has no direct equivalents to any of these provisions. Some remedial steps have been taken, the Data (Use and Access) Act 2025 criminalises the creation of non-consensual intimate deepfakes, but the overall approach remains piecemeal and reactive, lacking the systematic, AI-centric regulatory framework the EU has put in place.

- Another example is California, which has taken an aggressive, multi-pronged approach to regulating AI and misinformation, passing many AI-related bills in late 2024 and 2025, related to election integrity, preventing deception in online advertising, content transparency, protections for individuals, and safety and governance of AI models (Williams, 2024). Even though many of these laws are currently facing intense legal challenges at the federal level, they are sending a strong message towards the US social media and AI technology companies on the need to introduce strong safeguards.

6.1.3 What Factors Moderate the Effectiveness of Strategies and Interventions

Section Summary: The effectiveness of interventions to counter false and misleading information depends on context, audience, and delivery. Evidence shows that individual factors such as analytical reasoning, identity, literacy, and socioeconomic status shape both susceptibility to misinformation and responsiveness to corrections. Interventions are generally less effective where misinformation aligns with political or social identity, particularly in highly polarised environments, while crisis conditions further reduce effectiveness by increasing uncertainty, emotional responses, and information overload. Trusted messengers and sustained, proactive communication approaches are therefore critical. The evidence also suggests that longer, interactive, and repeated interventions outperform one-off or purely informational approaches, especially where misinformation is emotionally charged or visually compelling. Overall, effective responses require tailored, context-sensitive strategies rather than universal solutions.

The effectiveness of strategies and interventions that seek to counter and mitigate false and misleading information is not fixed. It varies systematically depending on who delivers the intervention, to whom, about what, and under what conditions. Understanding these moderating factors is essential for designing interventions that work in practice, not just in principle. Below, we summarise the key moderating factors drawn from the evidence base and expert discussions:

1.Individual Characteristics:

- **Analytical reasoning ability and motivation** moderate both susceptibility to misinformation and receptivity to correction; strategies that demand effortful reflection are less effective among those less equipped or inclined to engage in it (Nan et al., 2022; van der Linden et al., 2025).
- **Willingness to acknowledge the limits of one's knowledge** is associated with lower misinformation belief and greater engagement in evidence-based behaviours (Bowes and Fazio, 2024).
- **Motivated reasoning and confirmation bias** limit the effectiveness of corrections, particularly when misinformation aligns with identity or ideology; when people value identity over accuracy, they are more likely to believe and spread misinformation. However, outright backfire effects (i.e., where corrections cause people to double down) are relatively rare; the more common outcome is that corrections work partially but less well for identity-aligned content (Aghajari, Baumer and DiFranzo, 2023; Ziemer, Rothmund and Altay, 2024; van der Linden et al., 2025).
- **Age** moderates susceptibility and intervention response in complex ways. Older adults are paradoxically better than younger adults at discerning true from false headlines, yet more likely to encounter and share misinformation on social media, a pattern that may reflect poor digital literacy, greater trust in news sources, and different communication goals rather than belief susceptibility per se. Both older people and adolescents therefore require tailored approaches, but for different reasons (Nan et al., 2022; Hartwig, Doell and Reuter, 2024; van der Linden et al., 2025).
- **Lower socioeconomic status and limited health literacy** compound vulnerability by creating barriers to accessing reliable information and understanding health content (Choukou et al., 2022). Equity considerations should therefore be embedded in intervention design from the outset, rather than treated as an afterthought.

2. Topic Domain:

- **Health misinformation appears more correctable than political misinformation**, in part because health beliefs are less tightly bound to partisan identity and therefore less resistant to updating (van der Linden et al., 2025). Related, polarised political environments present a particularly significant challenge. High levels of political polarisation consistently reduce the effectiveness of individual-level interventions. Partisan-motivated cognition is the single strongest predictor of misinformation sharing: when identity stakes are high, accuracy nudges, fact-checks, and corrections all show weaker effects (van der Linden et al., 2025). In highly polarised contexts, trusted community messengers and technique-based inoculation, targeting manipulation tactics rather than specific partisan claims, are likely to be more effective than content-specific fact-checking.

- **Crisis and acute threat conditions create a further distinct challenge**. During infodemics, information voids are filled rapidly by false content, trust in official sources may already be eroded before corrections can be issued, and the volume of new claims outpaces fact-checking capacity (Abuhaloob et al., 2024). Crisis conditions also heighten emotional arousal, including fear, anxiety, and uncertainty, which can accelerate the spread of false and misleading information and reduce individuals' capacity for critical evaluation (Martle & Pennycook, 2020; Han, Cha & Lee, 2020). Evidence suggests that proactive communication (i.e., filling information voids with accurate content before false claims take hold) outperforms reactive correction in these conditions (Gentili et al., 2024). This makes a strong case for pre-positioning prebunking strategies and trusted source communications before crises emerge, rather than attempting to deploy them under pressure once misinformation has already spread.

3. Source Credibility and Messenger Effects

- **Perceived source credibility** does not consistently predict correction effectiveness across studies; conceptual and methodological factors contribute to this variation. However, debunking has been found to be effective regardless of tone, whether corrections appear to come from an algorithm or a human, and regardless of presentation order, suggesting that simply getting people to engage with a correction may matter more than who delivers it (Mang, Fennis and Epstude, 2024; van der Linden et al., 2025).

- **Trusted influencers and community figures** are important messengers, particularly on platforms where algorithm-driven content amplifies unverified claims; corrections are most successful when delivered by trusted sources and representatives, including religious, political, and community leaders (Ruyang and Hedi, 2025; van der Linden et al., 2025).
- **Declining trust in institutions** is a significant barrier to effective counter-messaging; the most important countermeasure identified in one review was "building and maintaining trust," and exposure to conspiracy theories has been shown to further erode institutional trust, creating a self-reinforcing cycle (Ries, 2022; Surjatmodjo et al., 2024; van der Linden et al., 2025).

4. Intervention Design

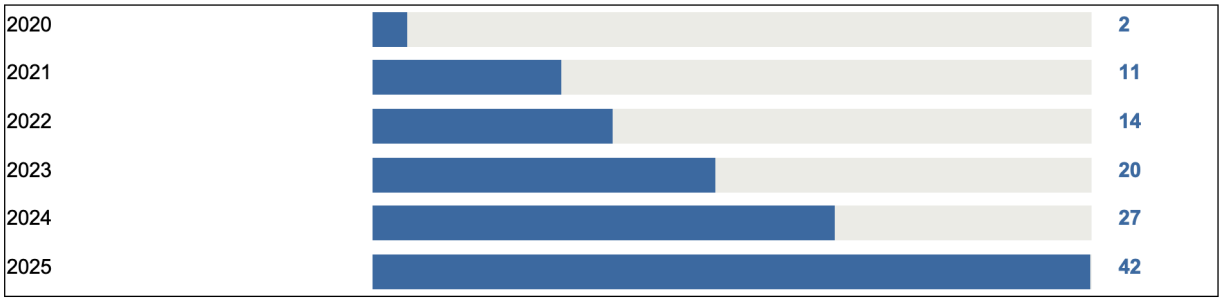
- **Intervention duration** is a critical but frequently overlooked moderator of literacy programme effectiveness. Short interventions of less than sixty minutes are often ineffective, while longer programmes of five hours or more are more likely to produce meaningful gains in critical appraisal skills. Studies with delayed follow-up also find that effects can fade over time, suggesting a need for sustained rather than one-off delivery (van der Linden et al., 2025).
- **Interactive and game-based formats** show stronger evidence than passive instruction; effectiveness is further enhanced when programmes explicitly address emotional targeting and algorithmic governance, not just content accuracy (Boler et al., 2025).
- **Visual misinformation** is more effective than text alone at capturing attention and eliciting emotional responses; text-only counter-strategies are therefore likely to be less effective against visually compelling false content (Liu and Kuru, 2025).
- **The "continued influence effect" is well documented**: corrections reduce but rarely eliminate misperceptions, either because people fail to integrate the correction into their mental model or because they later forget the correction or its source. A related phenomenon, "belief regression," describes the gradual fading of debunking effects over time; repeated fact-checks are therefore likely to be more effective than single corrections (Eva, Sakura and Li, 2021; van der Linden et al., 2025).

6.1.4 Coverage and growth of the evidence base

The volume of literature on strategies to counter and mitigate false and misleading information has expanded rapidly since 2020. This growth has been widely attributed to

the COVID-19 “infodemic,” which acted as a major catalyst for research in the field (Alvarez-Galvez et al., 2025), and, more recently, to emerging concerns surrounding AI-generated content (Bondielli & Marcelloni, 2025). As shown in Table 4.2, 42 of the 116 reviews (36%) were published in 2025, with a further 27 (23%) appearing in 2024. Together, these figures indicate that the evidence base has more than doubled within the past two years alone. ***Policymakers should therefore approach this as a dynamic and rapidly evolving field, rather than a settled body of knowledge.***

Table 6.1 Reviews tagged Countering & Mitigation by year of publication



In terms of policy domains, the evidence on countering and mitigating false and misleading information spans a wide range of areas, although health, particularly COVID-19, dominates the literature. Other important domains, including climate change, nutrition, elections, and mental health, are represented but are supported by comparatively thinner evidence bases. Table 4.3 summarises the review-level evidence on countering and mitigation strategies, coded by the policy domains they address.

Table 6.2 Reviews tagged Countering & Mitigation by policy domain

Domain	Reviews (n)
Social media platforms (general)	44
Health (general)	39
COVID-19 / pandemic & infodemic	36
Education & youth	20
Vaccination	13
Vulnerable / marginalised groups	13
AI-generated content	9
Democratic resilience / elections	8
Mental health	8
Climate & environment	4
Nutrition / food	5

6.1.5 Limitations and priorities for future research

Section Summary: Addressing these gaps requires a reorientation of the research agenda toward the following priorities: large-scale field evaluations embedded in live systems; longitudinal assessment of intervention durability; measurement of observable behavioural outcomes aligned with policy objectives; evidence on combined and sequenced interventions; cross-cultural validation across diverse institutional contexts; evaluation of structural and platform-level approaches; updated assessment of effectiveness in AI-mediated information environments; and greater attention to implementation factors including uptake, reach, and fidelity. Without progress on these fronts, the evidence base will continue to inform what interventions are theoretically promising while providing limited guidance on how they should be designed, targeted, and delivered in practice.

Importantly, the call for further research should not be read as a basis for deferring action. The harms associated with false and misleading information are sufficiently well-evidenced to justify policy responses now, particularly on structural and platform-level drivers where the case for intervention is already strong. The research

agenda outlined above is intended to improve the precision and effectiveness of those responses over the short, medium, and long term, not to replace them.

Existing evidence on strategies to counter and mitigate false and misleading information presents a structural imbalance: it is strong on assessing short-term efficacy in controlled settings, but weak on assessing real-world effectiveness, scalability, and policy applicability. For policymakers, this creates a gap between ***"what works in principle"*** and ***"what works in practice".*** Drawing in particular on two recent reports by Bang et al. (2025) and van der Linden et al. (2025), alongside the wider literature, several interconnected limitations characterise the current literature:

1. **Real-world effectiveness at scale:** Most primary evidence is derived from laboratory or short-term online experiments, using convenience samples. These designs cannot capture platform dynamics, network effects, or variation across populations and information environments. Large-scale field evaluations, embedded within live systems such as social media platforms, public health campaigns, or regulatory interventions, are essential to determine whether efficacy translates into real-world effectiveness (e.g., as in early examples by Farrelly et al., 2002; Holtgrave et al 2009).
2. **Durability and sustainability of effects:** Short-term changes in attitudes, beliefs, or intentions may provide a limited basis for policy decisions, particularly when long-term outcomes are uncertain. Although interventions such as debunking, prebunking, and literacy approaches can be effective, evidence suggests that their effects often decay over time or require reinforcement (Carey et al., 2022; Maertens et al., 2021, 2025; Pennycook et al., 2021; Roozenbeek et al., 2021; Stassen et al., 2020). For example, corrections can fade due to belief regression and memory decay, while prebunking effects may diminish without follow-up "booster" interventions, and literacy interventions frequently show limited or short-lived impacts. Moreover, relatively few studies assess long-term outcomes, highlighting a gap in the evidence base (Dias & Sippitt, 2020; Nordheim et al., 2016). Future research should prioritise longitudinal designs, examine mechanisms of persistence and decay, and evaluate the cost-effectiveness of intervention strategies over time.
3. **Behavioural outcomes and decision-relevant metrics:** The current literature focuses heavily on the effects of false and misleading information on beliefs, attitudes, and intentions, rather than actual behaviour or other outcomes, which

limits policy relevance. Although misinformation has consistently large effects on beliefs (Chan et al., 2017; Chan & Albarracín, 2023), its effects on behaviour are generally smaller (Bierwiaczonek et al., 2022; Stasielowicz, 2022), vary across studies (Greene & Murphy, 2021; Pummerer et al., 2022), and relatively few studies directly assess real-world behavioural outcomes (Schmid et al., 2023). Research should prioritise observable outcomes, such as sharing behaviour, service uptake, or compliance with public guidance, and align evaluation metrics with concrete policy objectives. There were also suggestions from experts that tangible economic effects should be considered.

4. **Combined and sequenced interventions:** Interventions are typically evaluated in isolation, despite often being deployed in combination in real-world settings. There is little evidence on how approaches such as prebunking, debunking, literacy interventions, nudges, and platform-level changes interact, whether they reinforce each other, duplicate effort, or sometimes work against each other. This has practical consequences: modelling of viral misinformation during the 2020 US election found that commonly proposed interventions were unlikely to be effective when implemented individually, but that a combined approach could achieve reductions of over 50% in misinformation volume (Bak-Coleman et al., 2022). Understanding these interaction effects, including optimal sequencing and layering, is a critical and underdeveloped area for policy design.

5. **Contextual adaptation and equity:** The majority of studies have been conducted in high-income, Western populations, which limits generalisability and risks ineffective or inequitable policy. Whether and how interventions can be tailored to different cultural, linguistic, and institutional contexts is rarely addressed, and cross-cultural validation remains limited. Future research should prioritise diverse settings, including low- and middle-income contexts and rural and urban areas, and examine how interventions perform across different populations and levels of institutional trust.

6. **System-level and structural interventions:** There is a mismatch between where policy action is concentrated, such as on governance and regulation of social media platforms, the design of algorithms, and where evidence is strongest, namely, on the effects of individual-level interventions. That is, structural approaches may have substantial impact but are comparatively under-evaluated, in part due to limited data access and reliance on platform cooperation. Generating robust evidence in this area is essential for informing effective regulation and platform accountability.

7. **An evolving technological environment:** The rapid development of generative AI is changing both the scale and characteristics of false and misleading information. Much of the existing evidence base predates these developments and may not generalise to current conditions. Features such as emotional amplification, rapid diffusion, and algorithmic reinforcement are likely to be intensified in AI-mediated environments, requiring updated evaluation of both detection tools and intervention effectiveness. Of particular concern is the emergence of content deliberately engineered to influence AI systems themselves, including large-scale AI-generated propaganda and material designed to be ingested during model training, which risks embedding misinformation into the foundations of future AI outputs and requires dedicated research attention.

8. **Implementation processes: uptake, reach, and fidelity:** A critical shift is needed from asking ***whether*** interventions work to understanding ***how*** they are implemented in practice. Key factors such as uptake, reach, fidelity, and adaptation are largely absent from current research, limiting its applicability for policymakers. Embedding implementation considerations into study design, including barriers to adoption and real-world delivery constraints, will be essential for translating evidence into effective action.

6.1.6 Emerging Policy Challenges

Section Summary: Emerging policy challenges highlight that efforts to counter false and misleading information are constrained by structural factors not fully addressed in the current evidence base. These include limited digital sovereignty over global platforms, weak regulatory enforcement combined with restricted data access, and the dominant role of platform business models and the attention economy in driving the problem. Together, these findings suggest that effective responses will require systemic, rather than primarily individual-level, policy solutions.

Expert consultations highlighted several emerging challenges related to countering and mitigating false and misleading information that are not yet well captured in the evidence base. These gaps point to structural constraints and limitations with significant implications for policy.

Importantly, online information environments are not neutral or unmediated: platform algorithms already curate what users see by ranking, promoting, and demoting content

at scale. As a result, the question is not whether information should be curated, but who exercises this curatorial power and under what conditions. At present, these decisions are largely made by private platforms, often with limited transparency or accountability, raising concerns about the concentration of informational power. This reframing has important implications for policy debates, suggesting that efforts to address false and misleading information must also engage with questions of governance, including the role of regulatory oversight, democratic accountability, and user choice in shaping algorithmic systems.

1. The digital sovereignty challenge: A key issue identified by experts is the lack of digital sovereignty.

- Major platforms are US-based corporations and therefore subject to US law, including the CLOUD Act, which can compel disclosure of data regardless of where it is stored (Wire, 2025).
- This creates a structural constraint on national and regional governance: frameworks such as the UK's Online Safety Act or the EU's Digital Services Act operate within a jurisdiction that may be overridden by US legal authority. This means that efforts to counter false and misleading information are shaped by a foreign legal system whose priorities may diverge from domestic policy objectives.
- This constraint remains underexplored in the academic literature, which tends to focus on platform regulation without fully engaging with questions of legal jurisdiction, trade policy, or digital infrastructure sovereignty. Experts emphasised that addressing this issue will require policy approaches that extend beyond the current scope of research on false and misleading information.

2. Regulation and data access constraints: Experts also highlighted a growing gap between regulatory ambition and practical enforcement.

- While regulation is seen as a necessary tool for addressing false and misleading information, its effectiveness is limited by weak enforcement and restricted access to platform data.
- UK regulatory provisions are also lagging behind those in the EU and more globally (see point 5 in Section 6.1.2).
- There was strong consensus that researcher access to data has worsened, even following the introduction of the Digital Services Act. This significantly constrains

the ability to study the dynamics of false and misleading information and to evaluate the impact of interventions.

3. Systemic drivers over individual interventions:

- Experts consistently emphasised that the spread of false and misleading information is primarily driven by the attention economy and platform business models. As a result, individual-level interventions, however well-designed, are insufficient to address what is fundamentally a system-generated problem. This supports existing findings, but extends their implications: effective policy responses must move beyond content moderation and media literacy to include business model reform and stronger algorithmic accountability.

- A key structural driver identified by experts is the engagement-based algorithm. By amplifying the most emotionally charged and extreme content to maximise ad revenue, these algorithms create the conditions in which false and misleading information spreads at abnormal scale. In this framing, the content itself is not the primary problem, it is the systematic over-amplification of that content that constitutes the harm. This distinction has significant policy implications: rather than regulating content, regulation should target the algorithmic mechanisms that drive amplification. Complementary measures include restrictions on addictive platform design features, such as infinite scroll, which the European Commission has identified as a factor in harmful content exposure (European Commission, 2026), and which US courts are increasingly scrutinising.

4. **AI-accelerated information environments:** Experts highlighted that the rapid growth of generative AI since 2022 has fundamentally altered the landscape in ways the existing evidence base has yet to fully capture.

- Kronhardt et al. (2025) identify a ‘proactive gap’, whereby detection tools remain largely reactive, identifying false and misleading information only after it has spread rather than anticipating or preventing it.

- López-Borrull and Lopezosa (2025) further show that generative AI introduces new challenges for fact-checking, detection, and the assessment of source credibility.

- As a result, much of the pre-2022 evidence, particularly on AI-based detection and estimates of exposure, may no longer be fully applicable.

- Experts therefore emphasised the need for updated research and policy approaches that reflect the speed, scale, and evolving capabilities of AI-generated content.

5. Information Integrity and Fundamental Rights:

Efforts to address false and misleading information must also contend with the need to mitigate harms while fully upholding core democratic freedoms. These objectives should not be treated as competing priorities, but as principles that must be pursued together by governments and digital platforms alike.

- Efforts to counter false and misleading information often rely on content moderation, algorithmic demotion, or legislative mandates, which researchers argue can inadvertently ‘chill’ legitimate speech or be weaponised by authoritarian regimes to suppress dissent. However, this debate often conflates distinct issues: for instance, fact-checking itself constitutes an exercise of free speech rather than its suppression (Lewandowsky, 2025). Moreover, the public tend to favour content moderation in principle (Kozyreva et al., 2023), and emerging evidence suggests that such measures do not necessarily chill expression and may even yield societal benefits, including reductions in offline hate crimes (Jiménez Durán et al., 2022).

- Moreover, regulatory frameworks designed to tackle false information must be grounded in international human rights law to ensure that the pursuit of information integrity does not come at the expense of the right to hold and express diverse opinions (Cipers et al., 2023).

- Furthermore, the lack of a standardised legal definition for "disinformation" remains a significant challenge, where overly broad or vague definitions in national laws risk chilling legitimate expression and facilitating the removal of otherwise lawful communications (Ó Fathaigh et al., 2021).

Table 6.3 Summary of Interventions for Countering and Mitigating False and Misleading information

Intervention	Definition & mechanism	
Individual-Level Interventions		
Prebunking / Psychological Inoculation	**What:** Proactively building cognitive resistance before exposure to false and misleading information by using weakened forms of misleading arguments, alongside preemptive refutations or clear guidance on how to identify propaganda. **How:** A forewarning that manipulation will be attempted (e.g., "*warning: people may try to manipulate you by saying X*"), plus a preemptive counter-argument explaining why the false claim fails (e.g., *"this is not true, because Y*"). **Types:** ● ***Issue-based*:** targeting specific false claims or stories vs. ***Technique-based****:* targeting common manipulation tactics (e.g., emotional manipulation, logical fallacies, conspiratorial reasoning). *NB this is sometimes referred to as "fact-based" vs. "logic-based".* ● ***Passive inoculation:*** counterarguments are provided to the participant (e.g., via text or video) vs. ***Active inoculation***: participants generate their own counterarguments (e.g., via games or quizzes)	**Summary of Evidence** ● **Lu et al. (2023). Meta-analysis, 42 studies (42,530 participants):** Psychological inoculation significantly reduces misinformation credibility assessment (*d*=−0.36) and improves credibility discernment (*d*=0.20) and sharing discernment (d=0.18), with effects documented for health and climate misinformation specifically. However, the effect on misinformation sharing intention was not statistically significant (*d*=−0.35, 95% CI −0.79 to 0.09; *p* =.120). ● **Simchon et al. (2026).** Evidence from a meta-analysis of over 37,000 participants confirms that these interventions improve veracity discernment (the ability to tell truth from lies) without increasing "response bias," meaning they do not make citizens more uniformly skeptical of all news. ● **van der Linden et al. (2026).** A recent large-scale Instagram field experiment exposed over 375,000 users to a brief prebunking video embedded in a social media feed, resulting in a 21% improvement in users' ability to identify manipulative content, with effects persisting over several months. ● **Kiili et al. (2024). Systematic review (15 papers):** Reviewed papers on game-based inoculation interventions reported positive outcomes, with most grounded in inoculation theory and focused on exposing players to manipulation techniques. Research has mainly been conducted in informal settings with adult participants. The evidence base is not yet mature enough to generalise findings. ● Game-based delivery (active inoculation) is a promising intervention that supports the development of critical reading skills by allowing users to interactively "role-play" as misinformation creators, which is theorised to be more effective than "passive" reading campaigns (Cook et al., 2023). ○ The approach has also been implemented in applied policy contexts: for example, *Go Viral* was developed in collaboration with the UK Government Communication Service as part of a public health misinformation response to COVID-19. **Design Requirements** ● **Prebunking is recommended as a first line of defense** to build public resilience by identifying misinformation techniques in advance rather than relying solely on reactive corrections (van der Linden et al., 2025).

		• **Longevity and Reinforcement:** The protective effects of inoculation are not permanent and typically fade within weeks (van der Linden et al., 2025; Lewandowsky & van der Linden, 2021). Regular "booster shots" (brief reminders or repeated exposure) are necessary to sustain resilience and can extend protection to around three months (van der Linden et al., 2025; Maertens et al., 2021). • **Technique-Based Approaches:** Strategies that target the common logic of manipulation (e.g., fake experts, emotional manipulation, or false dichotomies) provide "broad-spectrum" immunity (van der Linden et al., 2025; Lewandowsky & van der Linden, 2021). This is more efficient for policy, as it helps protect against many potential false claims without requiring constant updates (van der Linden et al., 2025; Lewandowsky & van der Linden, 2021). • **Scalability:** While field studies on platforms like YouTube demonstrate that video-based prebunking can be scaled to millions of users, there remains a gap in practical guidance for delivering these tools across more platform-diverse and real-world settings (van der Linden et al., 2025; Smith et al., 2023). ○ Emerging practitioner-oriented resources have begun to address this gap: a practical handbook developed by the University of Cambridge, BBC Media Action, and Jigsaw provides guidance on designing and deploying prebunking interventions, explicitly aiming to translate inoculation theory into scalable, real-world applications; and the prebunking tool developed by Full Fact and Maldita for the European Fact-Checking Standards Network, which forecasts narratives in short-form social media videos across 20+ European languages, illustrate how such approaches are beginning to scale in practice. **Limitations** • **Calibration of Trust:** If interventions are poorly designed, there is a minor risk of encouraging blanket skepticism toward all information (van der Linden et al., 2025). Policy should focus on "calibrating" trust, ensuring citizens can identify specific deceptive tactics while maintaining appropriate trust in legitimate scientific and news content (Simchon et al., 2026; van der Linden et al., 2025). • **Not all misinformation can be anticipated**. Even the most comprehensive prebunking strategies will leave significant volumes of false and misleading content unaddressed.

Debunking	**What:** Post-hoc correction of misinformation after exposure, explaining why a claim is false and providing accurate alternatives. **How:** Corrections are delivered by trusted sources (e.g., independent fact-checkers, health authorities), with sufficient detail about why the claim is false and what is true instead (e.g., via related articles sections, dedicated fact-check pieces, or direct social media corrections). There is also value where the original source of a false or misleading claim issues the correction themselves, as this carries additional credibility and accountability (e.g., in political contexts, where a politician or public official corrects the record on a claim they have made). **Types:** ● ***Reactive corrections:*** corrections issued in direct response to a specific false claim already in circulation ● ***Social media community corrections*** (peer-to-peer correction across platforms, where users directly challenge misinformation in their networks). ● ***Political fact-checking*****:** Verification of claims made by politicians, public officials, or during electoral campaigns, often focused on speeches, debates, manifestos, and policy statements.	**Evidence on effectiveness:** ● Reported as one of the most extensively studied intervention types in the literature (Francis et al., 2025). ● Meta-analyses generally show that debunking is effective at reducing, but not eliminating, misperceptions across ages, cultures, and topics (Chan & Albarracín, 2023; Chan et al., 2017). ● Debunking is most effective when it includes a clear explanation of why the misinformation is wrong, rather than simply flagging it as false (Chan & Albarracín, 2023; Ecker et al., 2010; van der Meer & Jin, 2020) ● Variations in how corrections are delivered, including tone, source type, placement, and ordering, make little difference to their effectiveness; the key factor is whether people engage with the correction at all (van der Linden et al., 2025) ● The effectiveness of debunking fades over time through "belief regression", primarily because people forget the correction or forget that the source was credible. Repeated fact-checks are therefore particularly important (Albarracín et al., 2017; Carey et al., 2022). ● The continued influence effect means that even after a successful correction, people who were exposed to misinformation tend to retain higher levels of false belief than those who were never exposed. This occurs because people fail to fully integrate the correction or do not retrieve it when needed (Chan et al., 2017; Ecker et al., 2022; Lewandowsky et al., 2012). **Design requirements:** ● **Accurate information should be featured more prominently than the information being corrected** so it is properly processed and later retrieved; repeating a false claim is only warranted when actively correcting it, and even then should be stated briefly (Löffler, 2021). This is sometimes called the “truth sandwich”. ● **Corrections must be repeated.** Effects fade over time as people forget the correction or that the source was credible (van der Linden et al., 2025) **Limitations** ● The continued influence effect means corrections reduce but do not eliminate the impact of false information, even when people accept the correction. ● Corrections frequently fail to reach those who most need them, as people already inclined to believe misinformation tend to avoid information that challenges it (Hameleers & van der Meer, 2019; Zollo et al., 2017)

		● Insufficient as a standalone strategy at scale. Because debunking addresses each false claim individually, there is an inherent asymmetry between the speed at which misinformation spreads and the pace at which corrections can be produced and disseminated
Warning labels & accuracy nudges	**What:** Platform-level labels and prompts that encourage accuracy before sharing, without restricting content. **How:** Warning labels (e.g., "disputed," "missing context") applied by fact-checkers or algorithms; accuracy prompts shown before sharing to make accuracy more salient; and source credibility indicators directing users toward higher-quality outlets. **Types:** ● ***Specific vs. generic warning labels*** (flagging individual claims or broadly signalling low-quality content) ● ***Accuracy prompts*** (pre-sharing nudges that shift attention toward accuracy before content is shared) ● ***Source credibility labels*** (indicators highlighting outlet quality or reliability) ● ***Social norm nudges*** (emphasising either what most people find acceptable (injunctive) or how others typically behave (descriptive)	**Evidence on effectiveness** ● **Martel & Rand (2023). Review of warning label research:** Warning labels are widely effective at reducing belief in and spread of labelled content. The size of effects depends on how labels are implemented and the characteristics of the labelled content being labelled. Despite some individual differences, recent evidence indicates that warning labels are generally effective across party lines and other demographic characteristics. ● **Marecos et al. (2024). Systematic review of source credibility labels and nudging interventions:** Results are mixed. Some interventions (content labels identifying misinformation, icon arrays) proved capable of influencing behaviour in specific contexts. Strategies specifically targeting source credibility signals failed to produce significant effects in the tested circumstances. ● **Smith et al. (2023). Systematic review of COVID-19 misinformation interventions:** Found evidence supporting accuracy prompts, warning labels, and overlays in mitigating belief in or spread of misinformation, but identified considerable variation across studies and called for standardised outcome measures. ● Accuracy nudges , prompts that make the concept of accuracy more salient before people share content, have been shown to improve the quality of news-sharing decisions, though effects are generally small (Pennycook & Rand, 2022; Pennycook et al., 2020, 2021) **Design requirements** ● Labels should explain why content is false and link to accurate alternatives, not simply flag it as disputed. ● Nudge effects appear stronger when delivered as multiple prompts in close succession, and weaker when delivered as a single one-off prompt (Pennycook & Rand, 2022) ● Deployment at scale requires platform cooperation to implement and maintain (Mündges & Park, 2024) **Limitations** ● Accuracy nudge effects are inconsistent and weaken with repeated exposure (van der Linden et al., 2025)

		• Both labels and nudges are less effective for people with strong prior beliefs or partisan motivation (van der Linden et al., 2025) • Generic labels risk reducing trust in all media indiscriminately rather than targeting false content specifically
Media & digital literacy education	**What:** Educational programmes building capacity to critically evaluate information, identify manipulation, and navigate digital environments. **How:** Delivered through school curricula, community programmes, online modules, and workplace training. **Types:** • ***Media literacy***: evaluating print and online media messages • ***Digital literacy***: skills required to execute tasks online • ***Social media literacy***: navigating platform affordances, algorithms, and sharing dynamics to reduce engagement with low-quality content • **Lateral reading:** verifying information by leaving the original source to consult external, independent sources. • **Critical ignoring:** strategically choosing what *not* to engage with to preserve attention and reduce susceptibility to manipulation	**Evidence on effectiveness** • **Droog et al. (2025). Systematic review of 80 experimental studies:** Media literacy intervention effectiveness depended more on the outcome variables targeted than on specific intervention characteristics. Most interventions improved misinformation detection, but effects on persuasive outcomes (e.g., attitudes, beliefs) were more inconsistent, suggesting that changing what people believe requires different or additional strategies beyond detection training. • **Boler et al. (2025). Scoping review of adult mis/disinformation literacy interventions:** Found diverse effective formats including course-based, web-based, and game-based programmes, as well as public events and visual resources. Experts recommended teaching about emotion targeting and regulation, algorithmic governance, lateral reading, and visual technology, and using interactive formats. Studies of evaluated interventions outside formal education were scarce. • **Ziapour et al. (2024). Systematic review of social media literacy in infodemic management (11 studies):** Social media literacy has emerged as a significant and effective strategy for managing health-related infodemics. Key vulnerability drivers include users' lack of media knowledge, distrust of government systems, and influence of local peers and influencers. **Design requirements** • Interactive, skills-based programmes anchored to real misinformation examples outperform generic awareness-raising. • Longer interventions produce more durable effects; short programmes of under an hour show limited impact (van der Linden et al., 2025). • Generative AI is creating new challenges for literacy education that existing curricula have not yet fully addressed (Fulsher et al., 2025) • Media and information literacy should be approached as a lifelong learning process, not limited to formal education, as rapid technological change continuously reshapes the skills needed to navigate digital environments (UNESCO, 2013) **Limitations**

		● Evidence for long-term retention of skills is weak. The few studies with extended follow-up generally find effects fading over time (van der Linden et al., 2025) ● Most evidence comes from Western, high-income populations; cross-cultural applicability is unclear ● Literacy programmes are slow to produce population-level change (i.e., years not months) and are therefore insufficient as a rapid response to acute disinformation threats.
Trusted source & authority communication	**What:** Using credible, trusted figures to deliver accurate information and corrections. **How:** Identifying and equipping trusted messengers (health professionals, community leaders, religious figures, peers) to communicate accurate information and challenge misinformation within their networks and communities. **Types:** ● ***Clinical/professional communication*** (patient-facing) ● ***Community champion models*** ● ***Religious and cultural authority figures*** ● ***Influencer-based correction strategies***	**Evidence on effectiveness** ● **Mang et al. (2024). Systematic review (N=91 studies, 64,162 participants):** Source credibility effects in misinformation research are inconsistent. While persuasion research consistently shows that perceived credibility predicts attitude change, research in misinformation correction contexts has yielded widely varying findings. Conceptual factors (how credibility is defined and measured) and methodological factors explain some of the variability. The review provides recommendations for more systematic conceptualisation. ● **Whitehead et al. (2023). Systematic review of vaccine communication interventions (34 studies):** Identified nine intervention approaches. Some strategies appear ineffective and may backfire. Specifically, scare tactics may increase misinformation endorsement, and communicating with certainty (rather than acknowledging uncertainty) about vaccine efficacy or risks was also found to backfire. Promising approaches include communicating weight-of-evidence, scientific consensus, and using humour. ● **Meeran et al. (2025). Narrative review of vaccine hesitancy in India (21 studies, 2013–2024):** Community-based interventions using local trusted figures (e.g., ASHA-led outreach, SMNet mobilization) demonstrated effectiveness through interpersonal communication and localised engagement, in a context where distrust in public health systems was identified as a key driver of hesitancy. Gaps persist in sustainability, frontline risk communication training, and real-time monitoring ● **Ruyang et al. (2025). Systematic review using Social Cognitive Theory (39 studies, 2019–2024):** Influencers and algorithm-driven content amplify unverified health claims, particularly on TikTok and Twitter. Effective countermeasures include media literacy programmes, transparent influencer disclosures, algorithmic reforms, and SCT-informed educational interventions. **Design requirements** ● Trust is highly context-specific; source credibility effects vary substantially depending on how credibility is conceptualised, the topic, and the study population. Conceptual and methodological consistency is needed before practical guidelines can be established (Mang et al., 2024).

		• Community-based interventions using local trusted figures can be effective in contexts where institutional trust is low, but require sustained investment; sustainability gaps and training needs persist (Meeran et al., 2025). • Influencers should be engaged with explicit attention to transparent disclosure; without this, they may function as amplifiers of unverified health claims rather than correctors (Ruyang et al., 2025). • Communication strategies should avoid scare tactics and should not overclaim certainty about vaccine efficacy or safety. Communicating with certainty (rather than acknowledging uncertainty) has been found to backfire (Whitehead et al., 2023). **Limitations** • Source credibility effects in misinformation correction are less settled than commonly assumed; a systematic review of 91 studies found inconsistent findings (Mang et al., 2024). • Identifying, training, and sustaining trusted messengers at scale is resource-intensive • Trusted figures can themselves spread misinformation, and influencers' perceived credibility can amplify unverified claims at scale. Engagement must be structured carefully to avoid this (Ruyang et al., 2025).
System-Level Interventions		

Platform governance & regulation	**What:** Regulatory and policy frameworks governing how platforms moderate, label, amplify, and distribute content. **How:** A spectrum from voluntary self-regulation to co-regulation and statutory frameworks governing platform responsibilities around misinformation. **Types:** ● ***Code of conduct*** (e.g., EU Code of Practice on Disinformation, now operating as a co-regulatory instrument under the Digital Services Act) ● ***Statutory obligations*** (e.g., Digital Services Act, AI Act) ● ***Co-regulatory models with independent audit*** ● ***Mandatory transparency reporting requirements***	**Evidence on effectiveness** ● An analysis of platform baseline reports confirms that major digital platforms (Google, Meta, Microsoft, TikTok, and X) are only partly compliant with the EU Code of Practice. Reporting is characterised by a lack of detail and non-robust quantitative data, highlighting a failure in current self-regulatory transparency (Mündges & Park, 2024). ● Effective counter-disinformation strategies in Europe are not a single-mechanism; they successfully combine regulatory measures (like the DSA), strategic partnerships, and media literacy to safeguard democratic resilience (D'Andrea et al., 2025). ● While government agencies have levers such as transparency and content moderation at their disposal, these must be balanced against significant ethical and legal discussions regarding public security and trust (Stieglitz et al., 2025). ● Evidence suggests that false and misleading information is an urgent sociotechnical problem. Governance must move beyond just technical detection (AI/ML) to address institutional structures and the broader complexity of the information ecosystem (Sanfilippo et al., 2025). **Design requirements** ● Enforcement mechanisms are essential. Voluntary commitments without independent audit and meaningful consequences consistently underdeliver. ● Portfolio approaches combining statutory regulation, public media investment, and independent oversight are better supported than single-mechanism strategies. ● Researcher access to platform data must be a component of any regulatory framework. Without it, the effects of interventions cannot be independently evaluated. **Limitations** ● Regulatory frameworks must carefully balance freedom of expression with protection from harm; however, failing to regulate effectively risks outsourcing these decisions to the most powerful technology companies. Such frameworks adopt a risk-based approach, focusing on harms arising from the design and functioning of platforms, as reflected in instruments such as the Digital Services Act and UNESCO guidelines on the governance of digital platforms. ● Jurisdictional fragmentation limits the reach of national regulations over global platforms ● Climate disinformation demonstrates how financially motivated obstructionist campaigns present distinct regulatory challenges from state-sponsored operations (Gertrudix et al., 2024) ● Disinformation as a tool of foreign interference involves actors and platforms operating outside the reach of domestic platform regulation (García-Estévez et al., 2025)

Algorithmic transparency & platform design	**What:** Interventions targeting the recommendation algorithms and design features of platforms that systematically amplify emotionally charged and misleading content. **How:** Changes to how platforms rank, recommend, and surface content, addressing the structural incentives that favour sensational or misleading material over accurate information. **Types:** ● ***Mandatory disclosure of recommendation logic***: requiring platforms to publicly explain how their algorithms select and prioritise content for users ● ***Demotion of unverified or flagged content:*** reducing the algorithmic reach of content identified as false or misleading before it spreads widely ● ***Friction mechanisms***: introducing delays or accuracy prompts at the platform level before users can share potentially false content ● ***Interoperability requirements:*** enabling users to choose or switch between different algorithmic feeds, reducing dependence on a single platform's recommendation system)	**Evidence on effectiveness** ● A literature review found that algorithmic systems optimised for engagement (likely for monetisation) are a primary driver of misinformation spread, and that efforts to contain misinformation through technology and fact-checking alone are insufficient without addressing algorithmic design (Onifade et al., 2023) ● A systematic review of health misinformation prevalence identified platform algorithms as a consistent factor in amplifying health misinformation across social media platforms (Suarez-Lledo & Álvarez-Gálvez, 2021) ● A literature review found that while AI tools offer potential solutions, the interaction between algorithmic amplification and AI-generated content represents a growing and under-addressed challenge (AbuJarour et al., 2024) ● A cross-disciplinary review identifies algorithmic design as a governance question requiring regulatory attention, not purely a technical one (Sanfilippo et al., 2025) **Design requirements** ● Transparency about how content is ranked and recommended is a precondition for meaningful oversight ● Interventions at the level of platform architecture have the potential to operate at a scale that individual-level approaches cannot reach ● Evidence on the effects of specific design changes is limited due to restricted researcher access to recommendation system data **Limitations** ● Direct evidence on the real-world effects of specific algorithmic interventions is thin (Gauthier et al., 2026). Platforms restrict access to the data needed to study this ● Risk that content demotion is overinclusive and affects legitimate speech alongside false content ● Platforms have commercial incentives to maintain engagement-maximising systems that work against misinformation-reducing design changes

AI-based automated detection	**What:** Machine learning (ML) and Natural Language Processing (NLP) systems that identify false, misleading, or manipulated content at scale across text, images, video, and audio. **How:** Automated systems deployed by platforms, governments, or third parties to detect and flag misinformation before or after it spreads widely. **Types:** ● ***Text-based NLP classifiers:*** algorithms trained to identify false or misleading claims or narratives within written content. ● ***Multimodal detection:*** systems combining text, image, and audio analysis to identify manipulated or synthetic media. ● ***Network-based detection:*** identifying coordinated inauthentic behaviour through patterns of account activity and content distribution. ● ***LLM-based tools:*** large language model systems, including browser extensions, that assess content credibility in real time. ● ***Credibility review architectures:*** transparent, explainable systems that link automated detection to human review processes.	**Evidence on effectiveness** ● A systematic review of AI approaches from 2014 to 2024 found that AI technologies have been extensively applied to detection tasks with growing sophistication, but that significant challenges remain around generalisation to new domains/data and reliability (Saeidnia et al., 2025). ● A scoping review mapping AI and digital tools in public health contexts found promise for monitoring and early warning, but noted that ethical and equity implications are not yet adequately addressed (Cianciulli et al., 2025). ● A systematic review of ML methods for detecting health misinformation found these approaches effective for high-volume content but dependent on high-quality, domain-specific training data (Bariş et al., 2024). ● A scoping review of publicly available LLM-based browser extensions identified a proactive gap: available tools are predominantly reactive, identifying misinformation after it has spread rather than preventing exposure (Kronhardt et al., 2025). ● A benchmarking study found that machine learning models for information disorder detection show significant vulnerabilities to adversarial attacks (Fenza et al., 2024). **Design requirements** ● Effective deployment requires ongoing model retraining, human oversight, and substantial data and research infrastructure. AI tools are not a turnkey solution (Cianciulli et al., 2025). ● Explainability is essential for regulatory and legal use. High-performing models that cannot justify their outputs are not suitable for public-sector deployment (Jareh et al., 2025). ● Training data quality and diversity directly determines detection reliability; biased or narrow training data produces unreliable outputs. ● AI systems are most effective when deployed alongside human expertise: while AI can monitor content at internet scale, human oversight is essential to provide contextual judgement, nuance, and appropriate caveats (“human-in-the-loop” approaches). **Limitations** ● Most models are domain-specific and do not generalise well across topics, languages, or content types (Saeidnia et al., 2025) ● Generative AI has widened the gap between misinformation production capability and detection capability (Kronhardt et al., 2025) ● There is prevailing scepticism about AI reliability in this context, and public trust in AI-based detection tools is not established (Jareh et al., 2025) ● Models show significant vulnerabilities to deliberate adversarial manipulation (Fenza et al., 2024)

Supply-side disruption	**What:** Interventions targeting the production and organised distribution of disinformation, rather than its reception by audiences. **How:** Actions taken against the infrastructure of coordinated disinformation campaigns, disrupting their ability to produce and distribute false content at scale. **Types:** ● ***Takedowns of coordinated inauthentic behaviour accounts*** (removing networks of fake or manipulated accounts engaged in organised influence operations) ● ***Sanctions against state-sponsored foreign interference operations*** (legal and diplomatic measures targeting government-backed disinformation campaigns) ● ***Disruption of disinformation-for-hire networks*** (action against commercial operators producing disinformation on behalf of paying clients) ● ***Defunding mechanisms*** (removing disinformation producers from advertising networks and demonetising their content to reduce financial incentives) ● ***Attribution and public exposure of disinformation campaigns*** (publicly identifying the sources and methods behind organised disinformation to deter future operations)	**Summary of Evidence:** ● A systematic review of astroturfing research from 2004 to 2024 found that coordinated campaigns simulating authentic grassroots movements are pervasive in digital communication and erode democratic deliberation, but that systematic study of the phenomenon remains scarce. While the literature proposes counter-measures including digital media literacy, automated detection tools, and legal regulation, research on their practical implementation and effectiveness remains underdeveloped (García-Estévez et al., 2025) ● A narrative integrative review found that public health disinformation in conflict and disease outbreak contexts is better understood as a systemic governance challenge than a technical one, requiring health diplomacy and political engagement as the primary response, with content moderation serving as a secondary complement (Peters et al., 2025) ● A narrative synthesis of COVID-19 infodemic research identified supply-side measures, including proactive communication to fill information voids and coordinated international responses, as among the recommended countermeasures, though evidence on their effectiveness was limited (Pian et al., 2021) ● A review of social media misinformation research identified supply-side factors including coordinated accounts and malicious actors as key drivers that mitigation strategies must address (Sadiq & Mathew, 2022) ● Expert consultations reinforced that supply-side disruption addresses structural drivers — the attention economy, platform business models, and coordinated production — that individual-level interventions such as literacy or corrections cannot reach. ● Emerging measurement tools, such as the “Cost of Online Social Influence” (COTSI), provide real-time estimates of the financial cost of conducting coordinated manipulation campaigns across platforms and countries, suggesting that increasing the cost of acquiring fake accounts or engagement could be a viable policy lever to disrupt supply-side dynamic (Dek et al., 2025). **Design requirements** ● Addressing state-sponsored foreign interference requires intelligence, diplomatic, and legal capabilities that go beyond platform content moderation ● Filling information voids proactively, before false claims take hold, is more effective than reactive takedowns during crises (Pian et al., 2021) ● International coordination is needed to address cross-border operations (Peters et al., 2025) **Limitations**

		● The evidence base is thin relative to other intervention types. Coordinated operations are covert and difficult to study. ● Takedowns can displace rather than eliminate coordinated networks, and expert consultations noted that even when content is identified and removed, the persuasive impact of content already seen is not eliminated ● Risk of overreach: poorly targeted supply-side measures can suppress legitimate speech ● Attribution of state-sponsored operations requires intelligence capabilities not available in most research or policy settings
Researcher Data Access & Transparency	**What:** Mechanisms enabling (independent) researchers to access platform data to study how false and misleading information spreads, evaluate interventions, and hold platforms accountable. **How:** Through policy and regulatory levers that create conditions for independent research, requiring platforms to share data responsibly, with frameworks modelled on existing medical data governance standards. **Types:** ● ***Mandatory free API access*****:** requiring platforms to provide accredited researchers with structured and direct access to platform data, in a privacy preserving manner. ● ***Ad library and content moderation log transparency*****:** publicly disclosing records of paid advertising and content removal decisions. ● ***Algorithmic audit requirements*****:** independent scrutiny of how platform algorithms select and amplify content.	**Summary of Evidence** ● Major social media platforms have restricted researchers' access to their data since 2022-23, reducing the ability to conduct independent studies of how misinformation spreads and how well platform-level responses work, as well as to carry out cross-platform studies. ● Independent researchers need access to platform data, including information on how content is ranked, recommended, and moderated, to study how false and misleading information spreads and to evaluate whether interventions are working (van der Linden et al., 2025). A cross-disciplinary governance review identifies this as a foundational requirement for evidence-based governance (Sanfilippo et al., 2025). ● A review of how well major platforms met their commitments under the EU Code of Practice on Disinformation found that poor transparency from platforms made it difficult to assess whether they were actually complying, illustrating what happens when data access obligations are weak (Mündges & Park, 2024). ● A review of ethical standards in managing health misinformation found that openness about how content is labelled, moderated, and flagged is a core requirement for public trust in these interventions (Germani et al., 2024). ● Research on system-level interventions, such as changes to platform algorithms or content policies, is constrained by limited access to platform data, making it harder to build the evidence needed to design and improve these approaches (Aghajari et al., 2023). ● Expert consultations identified a broader context: researcher data access is part of a wider regulatory and data access crisis, in which enforcement of existing obligations is weak and access to platform data is reported to be worse since the Digital Services Act than before, further limiting the ability to study the problem and evaluate interventions. ● Early evidence from the EDMO DSA data access pilot highlights significant practical barriers to accessing platform data under the Digital Services Act, including delays, inconsistent processes,

	● ***Responsible data-sharing frameworks***: governed access to sensitive user data, modelled on medical data governance standards to balance research needs with privacy protections).	and limited scope of accessible data, reinforcing concerns that current frameworks are not yet enabling effective independent research (EDMO, 2024) **Design requirements** ● Data-sharing frameworks modelled on those used for sensitive medical data have been proposed as a way to give researchers meaningful access while protecting user privacy (van der Linden et al., 2025) ● Voluntary data sharing by platforms has proved unreliable; statutory requirements with enforcement mechanisms are considered necessary to ensure consistent access. Evidence from the European Digital Media Observatory highlights how voluntary frameworks leave platforms in control of access decisions, creating barriers to research and undermining independence, and has led to calls for intermediary structures to standardise and govern data access (EDMO, 2023). ● Transparency obligations need to be accompanied by independent audit rights, so that platforms cannot solely control how their own compliance is reported. ● An independent intermediary body, separate from both platforms and regulators, has been proposed as a structural solution to facilitate researcher vetting and data access, removing platforms' control over who gains access to their data and helping to standardise review processes across institutions (EDMO, 2023). ● Expert consultations raised the question of digital sovereignty as a foundational constraint: because all major platforms are US-based corporations, they are ultimately subject to US federal law (including the CLOUD Act) which can override domestic regulations such as GDPR. This poses a structural challenge to any national or regional data access framework. **Limitations and Future Directions** ● Restrictions on researcher access to platform data since 2022-23 mean that the evidence base for evaluating platform interventions has become harder to build, not easier. ● Access to data alone is not sufficient. Adequate funding, research capacity, solid research infrastructure, and coordination between institutions are also needed. ● Sharing user interaction data raises legitimate privacy and security concerns that any data access framework must address. ● This intervention is primarily a precondition for generating evidence about other interventions, rather than a direct counter-misinformation tool in its own right.

Source note: *This table synthesises evidence from: (1) a systematic search of 116 review articles tagged 'Countering & Mitigation' identified through Scopus and PsycINFO searches; (2) the APA Consensus Statement on psychological science and health misinformation (van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J., 2025, American Psychologist, https://doi.org/10.1037/amp0001598); and (3) evidence included by experts.*

7 Looking forward: Evidence Gaps, Research Priorities and Trends in Primary Studies from 2025 Onwards

This section looks ahead by identifying key limitations in the current evidence base on false and misleading information, outlining priority areas for future research, and examining how recent primary studies are shaping the field. It highlights where evidence remains incomplete, what types of research are needed to support more effective policy and intervention design, and how emerging trends, particularly in technical and cognitive domains, are influencing the direction of research from 2025 onwards.

7.1. Evidence Gaps and Research Priorities

The preceding analysis highlights a clear imbalance in the current evidence base. Research has advanced significantly in understanding certain aspects of false and misleading information, particularly in health contexts and at the individual level, but remains limited in others, especially where policy action is most urgently required. These gaps reflect not only differences in research attention, but also structural constraints, including limited data access, methodological challenges, and the rapid evolution of the information environment. This section outlines where evidence is currently missing, the types of research required to address these gaps, and the broader strategic directions needed to strengthen the field.

7.1.1 Evidence Gaps: Where Evidence Is Missing or Limited

- A consistent finding across Sections 4 to 6 is that the **evidence on false and misleading information is unevenly distributed**, both in terms of subject matter and methodological approach. The most prominent imbalance is thematic. Research is heavily concentrated in health-related false and misleading information, particularly COVID-19, while other domains, such as climate (and science more broadly), democratic processes, and economic harms, remain comparatively underexplored. This creates a disconnect between the areas most studied and those of growing policy concern, especially as false and misleading information increasingly operates across multiple domains simultaneously.
- A second gap relates to the **measurement of impact**. Much of the existing literature examines how individuals respond to false and misleading information in controlled settings, which is valuable for identifying causal mechanisms but

less suited to understanding how these dynamics unfold in real-world environments. As a result, there is limited evidence linking exposure to false and misleading information to observable behavioural outcomes at scale. This challenge is compounded by the absence of standardised frameworks for measuring harm, making it difficult to compare findings or prioritise responses.

- There is also a **clear imbalance between individual- and system-level evidence**. While Sections 4.3 and 4.4 highlight that the spread of false and misleading information is driven primarily by structural factors, such as algorithmic amplification, platform incentives, and the attention economy, these drivers remain under-measured and under-evaluated. In contrast, individual-level responses to false and misleading information are comparatively well studied. This creates a mismatch between where the problem is understood to originate and where the strongest evidence exists.
- **Data access constraints further limit progress**. Restricted access to platform data has significantly reduced the ability to study how false and misleading information spreads, how it is encountered by users, and how interventions perform in practice. The increased barriers to data access, coupled with the paucity of large-scale longitudinal cross-platform research, have made it more difficult to generate timely and policy-relevant evidence.
- **Emerging AI technologies introduce additional gaps**. The rapid development of generative AI has transformed both the production and distribution of false and misleading information, yet much of the existing evidence base predates these changes. There is therefore limited understanding of how AI-generated false and misleading information operates in practice, particularly in personalised and private information environments where exposure is harder to observe.
- **UK policy responses and platform regulation efforts are lagging behind** those in other jurisdictions. It is now widely acknowledged that more effective whole-of-society policy responses are needed, going beyond content moderation and media literacy and towards business model reform and stronger algorithmic accountability. Approaches from other jurisdictions need to be evaluated promptly for their effectiveness and UK policies improved accordingly.

7.1.2 Research Needed to Address These Gaps

- Addressing these limitations requires a **shift in both the focus and design of research on false and misleading information**. A central priority is the move from controlled experimental settings to large-scale, real-world evaluations embedded

within live systems. This includes studying how false and misleading information circulates on platforms, how users encounter it in practice, and how different interventions perform under real-world conditions. Such approaches are necessary to understand how effects scale, how they vary across populations, and how they interact with platform dynamics.

- There is also a **need for longitudinal research** that examines the **persistence** of both exposure to false and misleading information and the **effects** of interventions over time. While short-term responses are relatively well documented, less is known about whether these effects endure, decay, or require reinforcement. Understanding these dynamics is essential for designing responses that are not only effective but sustainable.
- A further priority is the **measurement of behavioural and decision-relevant outcomes**. Much of the current literature focuses on how false and misleading information affects beliefs and attitudes, yet policy decisions often depend on changes in behaviour, such as sharing content, complying with public guidance, or making financial decisions. Aligning research metrics with these outcomes would significantly improve the policy relevance of the evidence base.
- At the same time, **greater emphasis is needed on system-level research**. This includes evaluating how platform design, algorithmic ranking systems, and engagement-driven incentives shape the distribution and amplification of false and misleading information. It also includes assessing the effectiveness of governance and regulatory approaches that aim to address these structural drivers. Given that these factors are widely understood to be central to the problem, strengthening the evidence base in this area is critical.
- **The rapid evolution of AI** also necessitates a dedicated research focus. This includes understanding how AI-generated false and misleading information is produced, how it is distributed across both public and private channels, and how it interacts with existing amplification systems. Particular attention is needed for risks associated with personalisation, microtargeting, and the potential integration of misleading content into AI training data.
- **Expanding research into underrepresented domains** is equally important. False and misleading information in areas such as climate, economic decision-making, identity-based disinformation, and democratic processes requires more systematic investigation, particularly given its growing prominence. This should be complemented by cross-cultural research that examines how false and

misleading information and responses to it operate in different institutional, linguistic, and socio-economic contexts.

- Another key priority is understanding **how different responses** to false and misleading information **interact when deployed together**. In practice, interventions are rarely used in isolation, yet there is limited evidence on how combinations of approaches reinforce or undermine each other. Research on sequencing, complementarities, and trade-offs would provide valuable guidance for policy design.
- Lastly, the **effectiveness of** the current UK and international **policy responses and platform regulation** approaches needs to be studied continuously, as well as their impact on citizens' exposure to false and misleading information.

7.1.3 Big Picture Research Directions

- Taken together, these priorities point to a broader shift in how research on false and misleading information should be conducted. First, there is a need to **move from a focus on individual pieces of content to a focus on systems**. This involves understanding the structural conditions, particularly algorithmic, economic, and institutional, that enable false and misleading information to spread at scale.
- Second, research must move beyond individual-level explanations toward **a more systemic perspective**. The evidence increasingly suggests that the spread and impact of false and misleading information is not simply a function of individual cognition, but of a broader information environment shaped by platform design, incentives, and trust dynamics.
- Third, the field must adapt to a **rapidly changing and increasingly complex landscape**. The emergence of generative AI, the shift toward personalised information environments, and the convergence of multiple domains into a "polycrisis" context all require research approaches that are flexible and responsive to change.
- Fourth, there is a need for **greater integration across domains**. False and misleading information increasingly operates across health, climate, political, and social issues simultaneously, meaning that siloed research approaches are no longer sufficient. Cross-domain analysis will be essential for understanding how narratives interact and reinforce one another.
- Finally, research must be **more closely aligned with policy and practice**. This includes ensuring that studies are designed with implementation in mind and that findings are directly applicable to real-world decision-making. Strengthening this

connection will be critical for translating insights into effective, scalable responses to false and misleading information.

- In this context, the identified **gaps should not be interpreted as a reason to delay action**. Rather, they highlight the areas where improved evidence can enhance the precision and effectiveness of responses that are already justified by the existing body of research on false and misleading information.

7.2 Trends in Primary Studies from 2025 Onwards

This section provides an overview of the supplementary literature search, which identified 1,239 primary research studies relevant to the three core review questions.

7.2.1 Technical Domain

- The most recently published literature is particularly well aligned with domains concerned with detection, classification and platform dynamics, reflecting the maturation of computational approaches to identifying and tracking misinformation across digital environments.

- A substantial proportion of studies focus on artificial intelligence–driven detection, machine learning classification, and analyses of social media ecosystems, including algorithmic amplification and virality. This cluster of research maps closely onto the “Understanding and Measuring the Problem” review question, especially detection challenges, AI content detection and distribution mechanisms, and demonstrates that recent empirical literature is highly responsive to the technical challenges posed by scale, speed and multimodality in contemporary information systems.

7.2.2 Cognitive and Behavioural Research

- There is a dense and well-developed body of work examining cognitive and behavioural dimensions of misinformation engagement. Studies addressing psychological drivers, belief formation, trust, credibility and narrative persuasion are strongly represented, aligning closely with the “Assessing Impacts” review question.

- This literature frequently intersects with a dominant thematic focus on health misinformation, particularly in relation to COVID-19 and vaccination, which continues to shape the empirical landscape. The prominence of this work reflects

both the urgency of recent public health crises and the relative tractability of studying misinformation in high-salience, data-rich domains.

- In parallel, there is substantial coverage of debunking and fact-checking interventions, including real-time correction mechanisms and platform-based responses, which map effectively onto the “Countering and Mitigation” branch and indicate a mature evidence base around downstream corrective strategies.

7.2.3 Structural, Institutional and Governance Gaps

- When viewed against the full conceptual scope of topics across the evidence map, the most recent literature reveals a skew toward cognitive and technical paradigms, with comparatively limited attention to structural, institutional and governance-oriented dimensions of misinformation. This is in no small part due to the severely limited access to platform data, which is needed to study the latter.

- Work addressing institutional or elite misinformation, including the role of political communication, corporate messaging and strategic ambiguity, is less prominent despite its conceptual importance, as highlighted by discussion with experts.

- Similarly, there is relatively sparse coverage of incentive structures and the political economy of misinformation, including engagement-driven business models, advertising systems and platform monetisation strategies, all of which were highlighted by experts as central to understanding how misinformation is produced and sustained at scale.

- Governance and policy-oriented research is present but uneven, with limited depth in areas such as regulatory design, accountability mechanisms and systemic oversight. In addition, recent literature places relatively little emphasis on information environments as dynamic systems, with concepts such as infodemics, information ecosystems and epistemic conditions appearing less frequently than might be expected given their prominence in policy discourse.

7.3 Implications for the Conceptual Scope of the Evidence Map

- The patterns identified in the most recently published literature suggest that while there is some work supporting understanding of how misinformation spreads, how individuals engage with it and how it can be detected and corrected, there is less coverage to explain why misinformation arises within institutional contexts and how structural conditions shape its persistence.

- The evidence map developed in consultation with experts therefore extends conceptually beyond the current centre of gravity in the evidence base, offering a more integrative perspective that incorporates upstream drivers, systemic dynamics and governance considerations. This divergence highlights critical gaps where further conceptual and empirical development is needed.

8 Conclusion

This report has synthesised a rapidly expanding and evolving evidence-base on false and misleading information, alongside insights from expert consultation. Taken together, the findings point to a fundamental shift in both the nature of the problem and the requirements for effective policy responses. False and misleading information is no longer best understood as a series of discrete incidents, but as a persistent, systemic feature of the modern information environment, shaped by technological change, platform incentives, and evolving actor ecosystems. Across the three core themes examined in this report: understanding and measuring the problem, assessing impacts, and countering and mitigation, a consistent conclusion emerges: effective policy must move beyond reactive, fragmented responses toward coordinated, system-level approaches grounded in evidence, proportionality, and respect for fundamental rights, and informed by an improved understanding of the psychological mechanisms that shape exposure, belief, and sharing of misinformation

8.1 Understanding and Measuring the Problem

The evidence highlights that the challenge is not simply definitional or content-based, but structural. False and misleading information is sustained by persistent, multi-domain dynamics, amplified by platform architectures and increasingly shaped by artificial intelligence, as well as by psychological factors such as repetition, emotional salience, and source credibility, which influence susceptibility and spread.

8.1.1 Key Insights

- **Adopt an impact-led approach to policy design:** Prioritise the harms caused by false and misleading information rather than over-reliance on distinctions based on intent. Definitions should support intervention design, not constrain it.
- **Shift from reactive to systemic monitoring frameworks:** Develop the necessary research infrastructure and accompanying capabilities to monitor persistent, concurrent, cross-domain, and cross-platform narratives, recognising the transition to a “polycrisis” information environment rather than episodic events.
- **Prioritise AI as a cross-cutting risk multiplier:** Treat generative AI not as a standalone category but as an enabling infrastructure that increases the scale, speed, and sophistication of false and misleading information across domains.

- **Strengthen transparency and data access for research:** Establish enforceable mechanisms to enable secure, independent access to platform data, recognising this as a foundational requirement for measurement, evaluation, and accountability, including understanding how algorithmic systems shape exposure at the individual and community level. Complement this with investment into an open-source research infrastructure and computational tools for large-scale, longitudinal, cross-modal and cross-platform analysis of false and misleading information.
- **Focus regulatory attention on amplification mechanisms:** Direct policy toward the systems that drive visibility and reach (e.g. recommender systems, engagement-based ranking), rather than content alone.

8.2 Assessing Impacts

The impacts of false and misleading information are multi-level, interconnected, and unevenly evidenced. While some harms are well established, others remain difficult to measure, creating challenges for prioritisation and policy design, particularly given that effects on beliefs are typically stronger than on attitudes and real-world behaviours.

8.2.1 Key Insights

- **Adopt a harms-based framework for prioritisation:** Focus policy attention on harms that are severe, measurable, and actionable, while recognising interdependencies across individual, societal, and organisational impacts.
- **Expand the evidence base beyond health misinformation:** Invest in research on understudied domains, including climate and science more broadly, democratic processes, economic harms, and impacts on marginalised groups.
- **Improve measurement of real-world outcomes:** Move beyond attitudes and beliefs to prioritise behavioural, economic, and societal indicators of harm, aligned with policy objectives.
- **Address data and methodological constraints:** Support new approaches to causal inference, including natural experiments, mixed-methods research, large-scale longitudinal studies, and cross-sector data sharing, to strengthen the evidence base.
- **Rebuild and maintain institutional trust as a policy objective:** Recognise trust as a central mechanism shaping impact, and embed transparent, consistent, and

responsive communication strategies across government, including the use of trusted messengers and community leader

- **Strengthen crisis communication capabilities:** Proactively fill information voids during crises to reduce vulnerability to misinformation, rather than relying solely on reactive correction.

8.3 Countering and Mitigation

The evidence is clear that no single intervention is sufficient. Effective responses require layered, adaptive strategies that combine individual- and system-level approaches, with a rebalancing toward structural interventions, and that reflect the complementary roles and limitations of different intervention types.

8.3.1 Key Insights

- **Adopt a portfolio approach to counter and mitigate:** Combine prebunking, debunking, media literacy, platform governance, and supply-side disruption, recognising their complementary roles.
- **Rebalance policy toward system-level interventions:** Prioritise action on platform design, algorithmic amplification, and business models, which are primary drivers of the problem at scale.
- **Strengthen platform governance and accountability:** Move beyond voluntary self-regulation toward enforceable frameworks, including independent audit, transparency obligations, and meaningful sanctions.
- **Address emerging challenges from AI and private information environments:** Develop evidence-based policy responses for personalised, opaque, and AI-mediated information systems, where traditional interventions are less effective.
- **Mandate integration of quality signals in algorithmic systems:** Require platforms to incorporate and transparently operationalise independent indicators of information quality in ranking and recommendation systems, as well as to label AI-generated and AI-manipulated content.
- **Sustain investment in prebunking and public resilience:** Scale evidence-based prebunking approaches with ongoing reinforcement, while maintaining debunking as a necessary backstop, and including repeated exposure or “booster” interventions to maintain effectiveness over time.

- **Embed context-sensitive intervention design:** Tailor strategies to audience, domain, and trust context, recognising that effectiveness varies across populations and environments.
- **Strengthen international coordination and digital sovereignty considerations:** Recognise the limits of national regulation over global platforms and pursue cross-border cooperation and strategic policy alignment.

9 Methods

The study was conducted between December 2025 and March 2026 and employed a mixed-methods design, incorporating a rapid evidence review and expert consultation.

9.1 Rapid Evidence Review

9.1.1 Bibliographic Literature Searches

Given the timescale for this project, a rapid scoping review was performed. A transparent and reproducible search of two academic databases (Scopus and PsycInfo) was conducted to identify recent evidence syntheses on false and misleading information, published between 2020 and 2026. The search prioritised review-level evidence (i.e., systematic, scoping, and narrative reviews) for two reasons: (i) the review was conducted within a six-month timeframe, necessitating a focused and efficient search strategy, and (ii) evidence syntheses provide a more overarching summary of the literature, aggregating findings across multiple studies and contexts rather than relying on individual empirical results. Predefined search terms were selected and agreed to reflect the varying terminology used across the field of false and misleading information (i.e., *misinformation OR disinformation OR malinformation OR infodemic* OR "information disorder" OR "information manipulation"*). To balance the retrieval of relevant material against the timeframe available for the review, search terms were required to appear in article titles (only).

9.1.2 Supplementary Searches

To ensure the review reflected the rapidly evolving nature of research in the field, a supplementary search for primary empirical studies published from 2025 onwards was also conducted, using the same databases and search terms but without the review-level filter. This reflected two considerations: (i) that evidence syntheses take considerable time to complete and publish, meaning recent primary research may not yet be captured at the review level, and (ii) in areas where syntheses are absent, primary studies represent the best available evidence and can highlight where future work is needed.

Grey literature was identified through internet searches (e.g., Google Scholar and AI-assisted search tools) and through consultation with experts and colleagues across relevant professional and academic networks (e.g., BR-UK, GO-Science) to capture additional or unpublished material. The review also incorporated a review of the

European Digital Media Observatory (EDMO) fact-checking database (2021–2025), which collates fact-checking reports produced by its network of member organisations across Europe. This database helped identify key themes in real-world instances of false or misleading information appearing in the media, mapping of these instances across policy domains, and assessment of how topics and frequency have changed over time. Experts that agreed to take part in this project were also asked to suggest further relevant references to complement the evidence retrieved (see Section 9.2).

Table 9.1.2 Summary of Evidence Identified Across Search Methods

Search	Source	Time period	No. Identified
Reviews	Scopus, PsycInfo	2020- 2026	228
Primary Studies	Scopus, PsycInfo	2025-present	1239
Request for Relevant Sources from Experts	Experts recommendation	2024-2026	19
Grey Literature	Google Scholar, AI-assisted search, Expert recommendation	2020-2025	22
Fact-Checking Database	EDMO Database	2021-2025	N/A

9.1.3 Study screening & Data charting

Articles identified as potentially relevant from the literature searches were imported into a shared Zotero library. Screening of all citations for eligibility was performed by one researcher (LU) and all references were checked by a second reviewer (HZ or HB). Analysis involved: 1. Importing references in Google Gemini to generate themes from the findings of the included reviews; 2. Inductive thematic analysis to ascribe the themes generated to each of the three review questions and to propose an initial thematic structure for the evidence map; and 3. Manual (human) review of the 228 references. Relevant keywords were deductively allocated to the references using the evidence map of the three review questions. Deductive thematic analysis was used to

ascribe relevant keywords from the evidence map to each reference in the Zotero reference management library.

9.1.4 Synthesis of results

Descriptive synthesis of the references was performed to produce narrative summaries for this report. As a scoping review the aim was to produce a high-level overview that comprehensively summarises the evidence base on the topic of false and misleading information, using expert consultation to sense check the completeness of scoping and to provide further suggestions of key themes and relevant references. As such, data regarding the quality or certainty of evidence from the literature retrieved were not extracted. This scoping review did not extract quantitative data regarding the effectiveness of interventions, initiatives or approaches. In order to assess the quality, efficacy and certainty of individual approaches to false or misleading information, further research involving focused, rigorous systematic review approaches would be required, involving comprehensive literature searches, clear eligibility criteria on the interventions of interest, pre-specified methods for best practice data analysis and objective measures for assessing the quality/certainty of evidence.

9.2 Expert Consultation

To complement and strengthen the coverage and suitability of the evidence, 11 experts working in the field of false and misleading information were recruited for a consensus-based exercise.

9.2.1 Expert Identification Strategies

A multi-pronged approach was used to identify potential experts: (i) Leaders of prominent research centres and networks, (ii) Keynote speakers at relevant and recent international conferences, (iii) Principal Investigators of relevant UKRI and ERC projects, (iv) AI-assisted tools (ChatGPT, Gemini) for identifying additional relevant experts, and (v) Networks of BR-UK and GO-Science.

9.2.2 Expert Selection

Selection Criteria

- Inclusion: Recognised academic or sector experts whose work addresses misinformation, disinformation, or influence operations.

- Exclusion: Individuals whose primary research or professional activity does not directly concern misinformation or its adjacent fields.

Selection Process Workflow: The selection followed a rigorous filtering process to ensure high-quality input:

- Initial Pool: HZ and KB generated a longlist of 135 experts using combined manual and AI-supported searches.
- Screening: HZ removed duplicates and reviewed individual biographies, publication lists, and online information to assess each candidate against the inclusion/exclusion criteria. The final longlist was refined to 93 experts across UK and international institutions.
- Invitation: The subject matter expert (KB) shortlisted 20 experts to be invited to the consultation.

Final Cohort: Of the 20 experts invited, 11 accepted and contributed to expert workshops and/or feedback on the draft report. Participants were drawn from academia (N = 6), media (N = 1), NGOs (N = 2), and industry (N = 2).

9.2.3 Workshops and Synthesis

Online Expert Workshops: Consultations were held on 9th and 13th March 2026. Experts were tasked with the following:

- Assessing the Evidence Map’s coverage of three core thematic areas.
- Identifying additional topics, evidence, or emerging sources.
- Evaluating the relevance and specific UK applicability of findings.
- Highlighting prioritisation and critical gaps for policy and research.

Synthesis and Reporting: Expert insights were processed through a formalised review cycle:

- Data Analysis: One researcher (HZ) pseudonymised and analysed the workshop data, synthesising key expert insights across the three thematic areas for the first draft of the report.

- Peer Review: 11 Experts were invited to review and comment on the draft report via a live Google Doc between 25th and 30th March 2026.
- Finalisation: Two researchers (HZ and HB) incorporated all expert comments into the final version of the report.

9.3 Ethics and Governance

The rapid review involved secondary analysis of published material. Expert consultation received ethical approval from the Research Ethics Committee in the School of Psychology at the University of Sheffield (Ref: 072370). Experts were remunerated for their time. The study was overseen by a Project Steering Group including representatives from BR-UK and GO Science. All conflicts of interest were declared in accordance with BR-UK's principles and governance (https://usher.ed.ac.uk/behavioural-research-uk/about-us/governance).

9.4 Use of AI

Due to the volume of material and the rapid nature of this review, AI tools (including NotebookLM and Claude) were used to support aspects of the evidence synthesis. However, human oversight was maintained throughout. All records were manually screened and coded at the title and abstract level by the delivery team, using predefined keywords aligned with the thematic and domain categories of the evidence map. Not all full-text articles were reviewed in their entirety due to time constraints. Instead, key full texts and/or abstracts were identified and uploaded and AI tools were used to assist in summarising and extracting key findings. These summaries were reviewed by the research team and experts to ensure accuracy, relevance, and alignment with the review objectives.

9.5 Delivery Team

The project was delivered by a multidisciplinary team from the University of Sheffield. The delivery team included Dr Harriet Baird (Project Lead and Lecturer in Psychology) with expertise in systematic reviews and meta-analyses; Professor Kalina Bontcheva, an expert in text analytics and misinformation; Dr Lesley Uttley, a specialist in research integrity and evidence synthesis; Professor Thomas Webb, an experienced psychologist with expertise in behaviour and evidence synthesis; and Dr Hui Zhang, an interdisciplinary researcher with experience in systematic reviews. Collectively, the team

brought expertise spanning psychology, management, sustainability, and computer science, alongside capabilities in behavioural research methods and evidence synthesis. This combination of methodological and domain expertise ensured the work remained strategically focused, rigorous, and objective, while minimising bias. The Sheffield team was supported by BR-UK Co-Director Professor Linda Bauld from the University of Edinburgh. The project also benefited from additional support provided by expert contributors.

9.6 Funding

This work was funded through a tender commissioned by the Government Office for Science and conducted as part of Behavioural Research UK (BR-UK). BR-UK is supported by the Economic and Social Research Council [grant number ES/Y001044/1] and includes funding to conduct responsive mode research designed to provide timely, evidence-based insights on emerging societal and economic issues to inform policy, practice, and future research priorities.

9.7 Licence

10 References

10.1 Reference List of 228 Review Studies

Aabeyir, R. (2024). Geoinformation or misinformation? A review of the geographic description of study areas in published academic articles. *African Geographical Review, 43*(5), 665–684. https://doi.org/10.1080/19376812.2023.2230199

Abbas, M. A., Parisapogu, S. A. B., Akhtar, M. Z., & Vandanapu, A. (2025). Exposing misinformation online: A review. *Institute of Electrical and Electronics Engineers*. https://doi.org/10.1109/RAIT65068.2025.11088931

Abuhaloob, L., Purnat, T. D., Tabche, C., Atwan, Z., Dubois, E., & Rawaf, S. (2024). Management of infodemics in outbreaks or health crises: A systematic review. *Frontiers in Public Health, 12*, Article 1343902. https://doi.org/10.3389/fpubh.2024.1343902

AbuJarour, S., Qarariah, A., Saadeh, N., & Salem, M. (2024). AI, misinformation, and fake news: A literature review of ethical and technical approaches. In *Springer Nature*. https://doi.org/10.1007/978-3-031-67547-8_55

Adler Berg, F. S., Lundtofte, T. E., Heiselberg, L., & Frischlich, L. (2025). Children and digital misinformation: A scoping review. *Global Studies of Childhood*. Advance online publication. https://doi.org/10.1177/20436106251398608

Aghajari, Z., Baumer, E. P. S., & DiFranzo, D. (2023). Reviewing interventions to address misinformation: The need to expand our vision beyond an individualistic focus. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), Article 95. https://doi.org/10.1145/3579520

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1), Article 30. https://doi.org/10.1007/s13278-023-01028-5

Al Arfaj, L., Lee, J. S., Shelton, J. A., Ertem, Z., Tran, T., Chen, Y., Blowers, M., & Wysocki, B. T. (2025). Navigating the infodemic: A comprehensive review of research on misinformation's impact on minority youth. *SPIE*. https://doi.org/10.1117/12.3053713

Ali, S. A., & Accessnow.org. (2022). Combatting against Covid-19 & misinformation: A systematic review. *Human Arenas, 5*(2), 337–352. https://doi.org/10.1007/s42087-020-00139-1

Alkhair, K. H., Yusof, M. H., Itam, M. F., Fisal, Z. A. M., Yatim, M. H. M., & Abdul Manaf, R. A. (2023). Analysing public health impact of misinformation during COVID-19 pandemic using the socio-ecological model: A systematic review. *Malaysian Journal of Medicine and Health Sciences, 19*(1), 242–253. https://doi.org/10.47836/mjmhs.19.1.32

Alsararatee, H. H., & Yunusa, N. M. (2025). The impact of nutrition misinformation on public health and practice: A review. *British Journal of Nursing, 34*(19), S18–S26. https://doi.org/10.12968/bjon.2025.0300

Álvarez-Gálvez, J., Carretero-Bravo, J., Lagares-Franco, C., Ramos-Fiol, B., & Ortega-Martin, E. (2025). Development of a conceptual framework of health misinformation during the COVID-19 pandemic: Systematic review of reviews. *JMIR Public Health and Surveillance, 11*, Article e62693. https://doi.org/10.2196/62693

Álvarez-Gálvez, J., Suarez-Lledo, V., & Rojas-García, A. (2021). Determinants of infodemics during disease outbreaks: A systematic review. *Frontiers in Public Health, 9*, Article 603603. https://doi.org/10.3389/fpubh.2021.603603

Angulo, A. J., & Hadley, M. (2025). *Peer review in the misinformation age*. Edward Elgar Publishing.

Arcos, R., Gertrudix, M., Arribas, C. M., & Cardarilli, M. (2022). Responses to digital disinformation as part of hybrid threats: A systematic review on the effects of disinformation and the effectiveness of fact-checking/debunking. *Open Research Europe, 2*, Article 8. https://doi.org/10.12688/openreseurope.14088.1

Asaad, C., Khaouja, I., Ghogho, M., & Baïna, K. (2025). When infodemic meets epidemic: Systematic literature review. *JMIR Public Health and Surveillance, 11*, Article e55642. https://doi.org/10.2196/55642

Aslani, N., Behmanesh, A., Davoodi, F., Garavand, A., & Shams, R. (2022). Infodemic challenges during COVID-19 pandemic and the strategies to deal with them: A review article. *Archives of Clinical Infectious Diseases, 17*(1), Article e127022. https://doi.org/10.5812/archcid-127022

Balakrishnan, V., Ng, W. Z., Soo, M. C., Han, G. J., & Lee, C. J. (2022). Infodemic and fake news – A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction, 78*, Article 103144. https://doi.org/10.1016/j.ijdrr.2022.103144

Bam, N. E. (2022). Strategies to address conspiracy beliefs and misinformation on COVID-19 in South Africa: A narrative literature review. *Health SA Gesondheid, 27*, Article a1851. https://doi.org/10.4102/hsag.v27i0.1851

Bang, C., Carroll, K., Mistry, N., Presseau, J., Hudek, N., Yanikomeroglu, S., & Brehaut, J. C. (2025). Use of implementation science concepts in the study of misinformation: A scoping review. *Health Education & Behavior, 52*(3), 340–353. https://doi.org/10.1177/10901981241303871

Bariş, I. B., Fernandez, E., Chulvi, B., & Rosso, P. (2024). Automatic detection of health misinformation: A systematic review. *Journal of Ambient Intelligence and Humanized Computing, 15*(3), 2009–2021. https://doi.org/10.1007/s12652-023-04619-4

Bernet, W. (2021). Parental alienation and misinformation proliferation. *Family Court Review, 59*(2), 207. https://doi.org/10.1111/fcre.12564

Bhandari, B., Zafra-Tanaka, J. H., Mahapatra, P., Njelekela, M., Infante-Garcia, M. M., Ramalingam, S., & Gonzalez-Rivas, J. P. (2025). Misinformation on cardiovascular disease spreads through social networks: A scoping review protocol. *BMJ Open, 15*(7), Article e094167. https://doi.org/10.1136/bmjopen-2024-094167

Bhattacharya, S., & Singh, A. (2025). Unravelling the infodemic: A systematic review of misinformation dynamics during the COVID-19 pandemic. *Frontiers in Communication, 10*, Article 1560936. https://doi.org/10.3389/fcomm.2025.1560936

Bianchi, F. P., & Tafuri, S. (2023). Spreading of misinformation on mass media and digital platforms regarding vaccines. A systematic scoping review on stakeholders, policymakers, and sentiments/behavior of Italian consumers. *Human Vaccines & Immunotherapeutics, 19*(2), Article 2259398. https://doi.org/10.1080/21645515.2023.2259398

Birkun, A. A. (2024). Misinformation on resuscitation and first aid as an uncontrolled problem that demands close attention: A brief scoping review. *Public Health, 228*, 147–149. https://doi.org/10.1016/j.puhe.2024.01.005

Bodaghi, A., Schmitt, K. A., Watine, P., & Fung, B. C. M. (2024). A literature review on detecting, verifying, and mitigating online misinformation. *IEEE Transactions on Computational Social Systems, 11*(4), 5119–5145. https://doi.org/10.1109/TCSS.2023.3289031

Boler, M., Gharib, H., Kweon, Y., Trigiani, A., & Perry, B. (2025). Promoting mis/disinformation literacy among adults: A scoping review of interventions and recommendations. *Communication Research*. Advance online publication. https://doi.org/10.1177/00936502251318630

Borges do Nascimento, I. J. B., Pizarro, A. B., Almeida, J. M., Azzopardi-Muscat, N., Gonçalves, M. A., Björklund, M., & Novillo-Ortiz, D. (2022). Infodemics and health misinformation: A systematic review of reviews. *Bulletin of the World Health Organization, 100*(9), 544–561. https://doi.org/10.2471/BLT.21.287654

Boukouvalas, Z., & Shafer, A. (2024). Role of statistics in detecting misinformation: A review of the state of the art, open issues, and future research directions. *Annual Review of Statistics and Its Application, 11*(1), 27–50. https://doi.org/10.1146/annurev-statistics-040622-033806

Bowes, S. M., & Fazio, L. K. (2024). Intellectual humility and misinformation receptivity: A meta-analytic review. *Advances.in/Psychology, 2024*(1), Article e00026. https://doi.org/10.56296/aip00026

Bran, R., Tîru, L., Grosseck, G., Holotescu, C., & Malița, L. (2021). Learning from each other—A bibliometric review of research on information disorders. *Sustainability, 13*(18), Article 10094. https://doi.org/10.3390/su131810094

Brassil, M., O'Mahony, C., Greene, C. M. A., & Alexander, J. R. M. (2024). Do cognitive abilities reduce eyewitness susceptibility to the misinformation effect? A systematic review. *Psychonomic Bulletin & Review, 31*(6), 2410–2436. https://doi.org/10.3758/s13423-024-02512-5

Broda, E., & Strömbäck, J. (2024). Misinformation, disinformation, and fake news: Lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association, 48*(2), 139–166. https://doi.org/10.1080/23808985.2024.2323736

Caballero, A., Chapi-Nitcheu, C., Vallan, L., Flahault, A., & Hasselgard-Rowe, J. (2025). Relationships between misinformation variables and nutritional health strategies: A scoping review. *International Journal of Environmental Research and Public Health, 22*(6), Article 891. https://doi.org/10.3390/ijerph22060891

Chaudhuri, N., Gupta, G., Bagherzadeh, M., Daim, T., & Yalçın, H. (2024). Misinformation on social platforms: A review and research agenda. *Technology in Society, 78*, Article 102654. https://doi.org/10.1016/j.techsoc.2024.102654

Chaufan, C., Hemsing, N., Heredia, C., & McDonald, J. (2024). Trust us—We are the (COVID-19 misinformation) experts: A critical scoping review of expert meanings of "misinformation" in the Covid era. *COVID, 4*(9), 1413–1439. https://doi.org/10.3390/covid4090101

Chernyaeva, O., Lee, O. D., & Hong, T. (2025). From fake to AI-generated: Leveraging information manipulation theory and explainable AI for robust detection of manipulated reviews. *Decision Support Systems, 197*, Article 114522. https://doi.org/10.1016/j.dss.2025.114522

Chikodzi, D., & Nhamo, G. (2023). COVID-19 infodemic and misinformation: A global review and implications for Zimbabwe. In *Springer International Publishing*. https://doi.org/10.1007/978-3-031-21472-1_18

Choukou, M., Sanchez-Ramirez, D. C., Pol, M., Uddin, M., Monnin, C., & Shabbir, S. (2022). COVID-19 infodemic and digital health literacy in vulnerable populations: A scoping review. *Digital Health, 8*, Article 20552076221076927. https://doi.org/10.1177/20552076221076927

Chowdhury, A., Kabir, K. H., Abdulai, A., & Alam, M. F. (2023). Systematic review of misinformation in social and online media for the development of an analytical framework for agri-food sector. *Sustainability, 15*(6), Article 4753. https://doi.org/10.3390/su15064753

Chowdhury, N., Khalid, A., Turin, T. C. A., & Abu-rish, E. Y. (2023). Understanding misinformation infodemic during public health emergencies due to large-scale disease outbreaks: A rapid review. *Journal of Public Health, 31*(4), 553–573. https://doi.org/10.1007/s10389-021-01565-3

Cianciulli, A., Santoro, E., Manente, R., Pacifico, A., Quagliarella, S., Bruno, N., Schettino, V., & Boccia, G. (2025). Artificial intelligence and digital technologies against health misinformation: A scoping review of public health responses. *Healthcare, 13*(20), Article 2623. https://doi.org/10.3390/healthcare13202623

Czerniak, K., Pillai, R., Parmar, A., Ramnath, K., Krocker, J., & Myneni, S. (2023). A scoping review of digital health interventions for combating COVID-19 misinformation and disinformation. *Journal of the American Medical Informatics Association, 30*(4), 752–760. https://doi.org/10.1093/jamia/ocad005

D'Andrea, A., Fusacchia, G., & D'Ulizia, A. (2025). Policy review: Countering disinformation in the digital age - Policies and initiatives to safeguard democracy in Europe. *Information Polity, 30*(1), 82–91. https://doi.org/10.1177/15701255251318900

Dalton, M. E., Duffy, R., Quinn, E., Larsen, K., Peters, C., Brenner, D., Yang, L., & Rainham, D. (2024). A qualitative review of social media sharing and the 2022 monkeypox outbreak: Did early labelling help to curb misinformation or fuel the fire? *Sexual Health, 21*(1), Article SH23158. https://doi.org/10.1071/SH23158

Damerchiloo, M., & Baghalha, F. (2023). Management of the COVID-19 infodemic in Asian countries: What should we know? (Systematic review). *Libri, 73*(3), 187–198. https://doi.org/10.1515/libri-2022-0064

Darnoto, B. R. P., Hasanah, M., Haq, T. I., Mafazy, M. M., Agustinus, J. T., Siahaan, D., & Purwitasari, D. (2024). Deep learning and ensemble approaches to misinformation detection in digital news: A systematic review. *Institute of Electrical and Electronics Engineers*. https://doi.org/10.1109/ITIS64716.2024.10845709

Das, B., & Tsb, S. (2023). Multi-contextual learning in disinformation research: A review of challenges, approaches, and opportunities. *Online Social Networks and Media, 34*, Article 100247. https://doi.org/10.1016/j.osnem.2023.100247

Dastani, M., & Atarodi, A. (2022). A systematic review of infodemic effects on mental health in the COVID-19 crisis. *Health Technology Assessment in Action, 6*(4). https://doi.org/10.18502/htaa.v6i4.12818

Delgado, C. E., Silva, E. A., Castro, E. A. B., da Costa Carbogim, F. D. C., Püschel, V. A. D. A., & Cavalcante, R. B. (2021). COVID-19 infodemic and adult and elderly mental

health: A scoping review. *Revista da Escola de Enfermagem da USP, 55*, Article e20210170. https://doi.org/10.1590/1980-220X-REEUSP-2021-0170

Denaux, R., Gómez-Pérez, J. M., Pan, J. Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., & Kagal, L. (2020). Linked credibility reviews for explainable misinformation detection. In *Springer Science and Business Media Deutschland GmbH*. https://doi.org/10.1007/978-3-030-62419-4_9

Denaux, R., Mensio, M., Gómez-Pérez, J. M., Alani, H., & Zhou, Z. (2021). Weaving a semantic web of credibility reviews for explainable misinformation detection (extended abstract). In *International Joint Conferences on Artificial Intelligence*. https://doi.org/10.24963/ijcai.2021/646

Dickinson, R., Makowski, D., van Marwijk, H., Ford, E. A., & Agley, J. (2025). Interventions for combating COVID-19 misinformation: A systematic realist review. *PLoS ONE, 20*(4), Article e0321818. https://doi.org/10.1371/journal.pone.0321818

Dos Santos Silva, E. M., & Arroio, A. (2025). Teaching acid-base theories in the era of disinformation: A systematic review with proposals for content integration. *Ecletica Quimica, 50*, 1–9. https://doi.org/10.26850/1678-4618.eq.v50.2025.e1530

Droog, E., Vermeulen, I., van Huijstee, D., Harutyunyan, D., Tejedor, S., & Pulido, C. (2025). Combatting the misinformation crisis: A systematic review of the literature on characteristics and effectiveness of media literacy interventions. *Communication Research*. Advance online publication. https://doi.org/10.1177/00936502251363705

Duzen, Z., Riveni, M., & Aktas, M. S. (2022). Misinformation detection in social networks: A systematic literature review. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, & C. Garau (Eds.), *Springer Science and Business Media Deutschland GmbH*. https://doi.org/10.1007/978-3-031-10545-6_5

El Mikati, I. K., Hoteit, R., Harb, T., El Zein, O., Piggott, T., Melki, J., Mustafa, R. A., & Akl, E. A. (2023). Defining misinformation and related terms in health-related literature: Scoping review. *Journal of Medical Internet Research, 25*, Article e45731. https://doi.org/10.2196/45731

Eva, Q., Sakura, O., & Li, G. (2021). Mapping the field of misinformation correction and its effects: A review of four decades of research. *Social Science Information, 60*(4), 522–547. https://doi.org/10.1177/05390184211053759

Fenza, G., Loia, V., Stanzione, C., & Di Gisi, M. (2024). Robustness of models addressing information disorder: A comprehensive review and benchmarking study. *Neurocomputing, 596*, Article 127951. https://doi.org/10.1016/j.neucom.2024.127951

Ferrández-Mas, A., López-López, P. C., Ribeiro, V., Ibáñez, D. B., Castro, L. M., Espinosa, A., Puentes-Rivera, I., & López-López, P. C. (2024). Technologies against disinformation in Ibero-America: Systematic review from the regional to the local level. In *Springer Science and Business Media Deutschland GmbH*. https://doi.org/10.1007/978-981-99-7210-4_27

Finnegan, P., Goh, Y., Murphy, M., & O'Connor, C. (2025). A qualitative review of misinformation on alopecia. *Skin Appendage Disorders, 11*(2), 182–185. https://doi.org/10.1159/000541809

Finnegan, P., Murphy, M., & O'Connor, C. (2023). Corticophobia: A review on online misinformation related to topical steroids. *Clinical and Experimental Dermatology, 48*(2), 112–115. https://doi.org/10.1093/ced/llac019

Finnegan, P., Murphy, M., & O'Connor, C. (2023). Reinventing the wheal: A review of online misinformation and conspiracy theories in urticaria. *Clinical and Experimental Allergy, 53*(1), 118–120. https://doi.org/10.1111/cea.14246

Fletcher, K., Rosas, C., Li, J., Macintosh, H., & Eaton, A. (2025). Love in the age of rage: A scoping review on the impacts of misinformation, conspiracy theories and political polarization on intimate relationships. *Journal of Couple & Relationship Therapy, 24*(4), 411–430. https://doi.org/10.1080/15332691.2025.2546607

Francis, Y., Lantry, F. J., Echeverry, M., & Siddiq, M. (2025). Effective countermeasures to health misinformation and disinformation for global health engagement practitioners: A rapid review. *Military Medicine, 190*(Suppl. 3), 478–482. https://doi.org/10.1093/milmed/usaf237

Freiling, I., Krause, N. M., & Scheufele, D. A. (2023). Science and ethics of “curing” misinformation. *AMA Journal of Ethics, 25*(3), 228–237. https://doi.org/10.1001/amajethics.2023.228

Fulsher, A., Pagkratidou, M., & Kendeou, P. (2025). GenAI and misinformation in education: A systematic scoping review of opportunities and challenges. *AI and Society*. Advance online publication. https://doi.org/10.1007/s00146-025-02536-y

Gabarron, E., Oyeyemi, S. O., & Wynn, R. (2021). Covid-19-related misinformation on social media: A systematic review. *Bulletin of the World Health Organization, 99*(6), 455–463. https://doi.org/10.2471/BLT.20.276782

García-Estévez, N., Ballesteros-Aguayo, L., & Colussi, J. (2025). Disinformation and manipulation of public opinion: A systematic review of astroturfing (2004-2024). *Revista de Comunicacion, 24*(2), 159–181. https://doi.org/10.26441/RC24.2-2025-3988

García-Marín, D. (2021). Research on disinformation: Topics, methodology and impact. A systematic literature review (2016-2020). *Doxa Comunicacion, 33*, 321–346. https://doi.org/10.31921/doxacom.n33a854

Gentili, A., Villani, L., Osti, T., Corona, V. F., Gris, A. V., Zaino, A., Bonacquisti, M., De Maio, L., Solimene, V., Gualano, M. R., Favaretti, C., Ricciardi, W., & Cascini, F. A. (2024). Strategies and bottlenecks to tackle infodemic in public health: A scoping review. *Frontiers in Public Health, 12*, Article 1438981. https://doi.org/10.3389/fpubh.2024.1438981

George, J. F., & Mannina, S. (2025). Detecting disinformation about health in social media: A review of four inductive studies. *Foundations and Trends in Information Systems, 9*(3), 160–246. https://doi.org/10.1561/2900000042

Germani, F., Spitale, G., Machiri, S. V., Ho, C. W. L., Ballalai, I., Biller-Andorno, N., & Reis, A. A. (2024). Ethical considerations in infodemic management: Systematic scoping review. *JMIR Infodemiology, 4*, Article e56307. https://doi.org/10.2196/56307

Gertrudix, M., Carbonell-Alcocer, A., Arcos, R., Arribas, C. M., Codesido-Linares, V., & Benítez-Aranda, N. (2024). Disinformation as an obstructionist strategy in climate change mitigation: A review of the scientific literature for a systemic understanding of the phenomenon. *Open Research Europe, 4*, Article 18180.2. https://doi.org/10.12688/openreseurope.18180.2

Goh, Y., O'Connor, C., & Murphy, M. (2025). What lies beneath: A qualitative review of misinformation on vulval lichen sclerosus. *JEADV Clinical Practice, 4*(1), 352–355. https://doi.org/10.1002/jvc2.561

Grahn, H., Kalsnes, B., Isaksson, E., Mayerhöffer, E., Ólafsson, J. G., Falkheimer, J., Henriksen, F. M., Kristensen, J. B., & Saari, D. (2025). Mapping research on

disinformation and misinformation across the Nordic countries: An integrative review. *Nordicom Review, 46*(Special Issue), 175–220. https://doi.org/10.2478/nor-2025-0015

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., Ecker, U. K. H., & Allen, M. (2023). Best practices for ethical conduct of misinformation research: A scoping review and critical commentary. *Zeitschrift für Psychologie, 28*(3), 139–150. https://doi.org/10.1027/1016-9040/a000491

Grover, H., Nour, R., Zary, N., & Powell, L. (2025). Online interventions addressing health misinformation: Scoping review. *Journal of Medical Internet Research, 27*(1), Article e69618. https://doi.org/10.2196/69618

Gruber, A., Ghiringhelli, M., Edri, O., Abboud, Y., Shiti, A., Shaheen, N., Ballan, N., Neuberger, A., & Caspi, O. (2021). Literature review and knowledge distribution during an outbreak: A methodology for managing infodemics. *Academic Medicine, 96*(7), 1005–1009. https://doi.org/10.1097/ACM.0000000000004073

Guallar, J., Codina, L., Freixa, P., & Pérez-Montoro, M. (2022). Disinformation, hoaxes, curation and verification: Review of studies in Ibero-America 2017-2020. *Online Media and Global Communication, 1*(3), 648–668. https://doi.org/10.1515/omgc-2022-0055

Guo, H., Huang, T., Huang, H., Fan, M., & Friedland, G. (2022). A systematic review of multimodal approaches to online misinformation detection. In *Institute of Electrical and Electronics Engineers*. https://doi.org/10.1109/MIPR54900.2022.00062

Gwiazdzinski, P., Gundersen, A. B., Piksa, M., Krysinska, I., Kunst, J. R., Noworyta, K., Olejniuk, A., Morzy, M., Rygula, R., Wojtowicz, T., & Piasecki, J. (2023). Psychological interventions countering misinformation in social media: A scoping review. *Frontiers in Psychiatry, 13*, Article 974782. https://doi.org/10.3389/fpsyt.2022.974782

Hartwig, K., Doell, F., & Reuter, C. (2024). The landscape of user-centered misinformation interventions: A systematic literature review. *ACM Computing Surveys, 56*(11), Article 280. https://doi.org/10.1145/3674724

Heley, K., Chou, W. S., D'Angelo, H., Senft Everson, N., Muro, A., Rohde, J. A., Gaysynsky, A., & Agley, J. (2025). Mitigating health and science misinformation: A scoping review of literature from 2017 to 2022. *Health Communication, 40*(1), 79–89. https://doi.org/10.1080/10410236.2024.2332817

Hilberts, S., Govers, M., Petelos, E., & Evers, S. (2025). The impact of misinformation on social media in the context of natural disasters: Narrative review. *JMIR Infodemiology, 5*, Article e70413. https://doi.org/10.2196/70413

Holzer, H., Diviani, N., & Rubinelli, S. (2025). COVID-19 misinformation and healthcare workers: A scoping review. *Patient Education and Counseling, 141*, Article 109309. https://doi.org/10.1016/j.pec.2025.109309

Hove, C., & Cilliers, L. (2023). A structured literature review of the health infodemic on social media in Africa. *Jamba: Journal of Disaster Risk Studies, 15*(1), Article 1484. https://doi.org/10.4102/JAMBA.V15I1.1484

Indu, V., & Thampi, S. M. (2021). A systematic review on the influence of user personality in rumor and misinformation propagation through social networks. In S. M. Thampi, R. M. Hegde, D. Ciuonzo, T. Hanne, & J. Kannan R. (Eds.), *Springer Science and Business Media Deutschland GmbH*. https://doi.org/10.1007/978-981-16-0425-6_17

Iqhrammullah, M., Gusti, N., Muzaffar, A., Khader, Y., Maulana, S., Rademaker, M., & Abdullah, A. (2025). Narrative review and bibliometric analysis on infodemics and health misinformation: A trending global issue. *Health Policy and Technology, 14*(5), Article 101058. https://doi.org/10.1016/j.hlpt.2025.101058

Irfan, A., Bieniek-Tobasco, A., Golembeski, C. (2021). Pandemic of racism: public health implications of political misinformation. *Harvard Public Health Review*. 2021; 26. http://doi.org//10.54111/0001/Z6

Janmohamed, K., Walter, N., Nyhan, K., Khoshnood, K., Tucker, J. D., Sangngam, N., Altice, F. L., Ding, Q., Wong, A., Schwitzky, Z. M., Bauch, C. T., De Choudhury, M., Papakyriakopoulos, O., Kumar, N. A., & Ades, A. (2021). Interventions to mitigate COVID-19 misinformation: A systematic review and meta-analysis. *Journal of Health Communication, 26*(12), 846–857. https://doi.org/10.1080/10810730.2021.2021460

Jareh, A. (2025). An in-depth assessment of the acceptability and effectiveness of AI tools in identifying misinformation in media: Literature review. *SAGE Open, 15*(4), Article 21582440251381272. https://doi.org/10.1177/21582440251381272

Joshi, G., Rathore, T., & Verma, K. (2025). Emotion-induced memory distortions: Insights from Deese-Roediger-McDermott and misinformation paradigms—A

comprehensive review. *Health Sciences Review, 14*, Article 100216. https://doi.org/10.1016/j.hsr.2025.100216

Kalbhor, S., Goyal, D., & Sankhla, K. (2024). Taming misinformation: Fake review detection on social media platform using hybrid ensemble technique. *International Journal of Electrical and Electronics Research, 12*(1), 28–35. https://doi.org/10.37391/ijeer.12bdf05

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society, 23*(5), 1301–1326. https://doi.org/10.1177/1461444820959296

Karami, A., Zain, A., & Jamal, A. (2025). Unveiling the information mirage: A systematic literature review of health misinformation on social media. *Journal of Public Health*. Advance online publication. https://doi.org/10.1007/s10389-025-02639-2

Kaur, K., & Gupta, S. (2023). Towards dissemination, detection and combating misinformation on social media: A literature review. *Journal of Business & Industrial Marketing, 38*(8), 1656–1674. https://doi.org/10.1108/JBIM-02-2022-0066

Kbaier, D., Kane, A., McJury, M., Kenny, I. A., & Abbasi, K. (2024). Prevalence of health misinformation on social media-Challenges and mitigation before, during, and beyond the COVID-19 pandemic: Scoping literature review. *Journal of Medical Internet Research, 26*, Article e38786. https://doi.org/10.2196/38786

Kearney, N., Leahy, M., & Laing, M. (2025). A qualitative review of online content relating to sunscreen misinformation. *Photodermatology, Photoimmunology & Photomedicine, 41*(4), Article e70028. https://doi.org/10.1111/phpp.70028

Kemei, J., Alaazi, D. A., Tulli-Shah, M., Kennedy, M., Tunde-Byass, M., Bailey, P., Sekyi-Otu, A., Murdoch, S., Mohamud, H., Lehman, J., & Salami, B. (2022). A scoping review of COVID-19 online mis/disinformation in Black communities. *Journal of Global Health, 12*, Article 05026. https://doi.org/10.7189/JOGH.12.05026

Keyes, C. (2024). Misinformation nation: Foreign news and the politics of truth in revolutionary America by Jordan E. Taylor (review). *Eighteenth-Century Studies, 57*(4), 574–576. https://doi.org/10.1353/ecs.2024.a931705

Khare, S., Erridge, S., Chidambaram, S., & Sodergren, M. H. (2025). Misinformation about medical cannabis in YouTube videos: Systematic review. *JMIR Formative Research, 9*, Article e76723. https://doi.org/10.2196/76723

Kiili, K., Siuko, J., & Ninaus, M. (2024). Tackling misinformation with games: A systematic literature review. *Interactive Learning Environments, 32*(10), 7086–7101. https://doi.org/10.1080/10494820.2023.2299999

Kisa, S., Kisa, A., & Agley, J. (2024). A comprehensive analysis of COVID-19 misinformation, public health impacts, and communication strategies: Scoping review. *Journal of Medical Internet Research, 26*, Article e56998. https://doi.org/10.2196/56998

Kolosov, M. I. (2023). Review of existing methods and technologies for detection and mitigation of misinformation and distorted data in social networks. *Scientific and Technical Information Processing, 50*(3), 196–202. https://doi.org/10.3103/S0147688223030073

Konstantinou, L., & Karapanos, E. (2025). Behavior change interventions combating online misinformation: A scoping review. In *Association for Computing Machinery*. https://doi.org/10.1145/3706598.3713127

Kont, J., Elving, W., Broersma, M., & Bozdağ, Ç. (2025). What makes audiences resilient to disinformation? Integrating micro, meso, and macro factors based on a systematic literature review. *Communications, 50*(2), 534–555. https://doi.org/10.1515/commun-2023-0078

Krishna, A., & Thompson, T. L. (2021). Misinformation about health: A review of health communication and misinformation scholarship. *American Behavioral Scientist, 65*(2), 316–332. https://doi.org/10.1177/0002764219878223

Kronhardt, K., Lehnert, M. J., Pascher, M., Gerken, J., Sorce, S., Elagroudy, P., & Khamis, M. (2025). The proactive gap: A scoping review of publicly available LLM-based browser extensions and their potential to mitigate information disorder. In *Association for Computing Machinery*. https://doi.org/10.1145/3771882.3771896

Küçükkocaoğlu, G., & Özden, A. T. (2025). Financial markets in the age of infodemics: A systematic review on the impact of information disorder. *Eurasian Economic Review*. Advance online publication. https://doi.org/10.1007/s40822-025-00338-7

Kumar, M., & Sharma, A. (2024). Disinformation and social media: Review of the film *Afwaah*. *Media Asia*, *51*(3), 519–524. https://doi.org/10.1080/01296612.2023.2244747

Kumar, N., Hampsher, S., Walter, N., Nyhan, K., & de Choudhury, M. (2022). Interventions to mitigate vaping misinformation: Protocol for a scoping review. *Systematic Reviews*, *11*(1), Article 108. https://doi.org/10.1186/s13643-022-02094-0

Kumar, N., Walter, N., Nyhan, K., Khoshnood, K., Tucker, J. D., Bauch, C. T., Ding, Q., Jones-Jang, S. M., de Choudhury, M., Schwartz, J. L., Papakyriakopoulos, O., & Forastiere, L. (2022). Interventions to mitigate COVID-19 misinformation: Protocol for a scoping review. *Systematic Reviews*, *11*(1), Article 124. https://doi.org/10.1186/s13643-022-01917-4

la Bella, E., Allen, C., & Lirussi, F. (2021). Communication vs evidence: What hinders the outreach of science during an infodemic? A narrative review. *Integrative Medicine Research*, *10*(4), Article 100731. https://doi.org/10.1016/j.imr.2021.100731

Lan, S., Mahmoud, S., & Franson, K. L. (2024). A narrative review on the impact of online health misinformation on patients' behavior and communication. *American Journal of Health Behavior*, *48*(2), 564–572. https://doi.org/10.5993/AJHB.48.2.25

Larki, M., & Manouchehri, E. (2022). Dispelling fake news and infodemic management about COVID-19 vaccination: A literature review. *Journal of Health Literacy*, *7*(3), 91–105. https://doi.org/10.22038/jhl.2022.65215.1289

Lazić, A., & Žeželj, I. (2021). A systematic review of narrative interventions: Lessons for countering anti-vaccination conspiracy theories and misinformation. *Public Understanding of Science*, *30*(6), 644–670. https://doi.org/10.1177/09636625211011881

Li, Y., Marga, J. J., Cheung, C. M. K., Shen, X., & Lee, M. K. O. (2022). Health misinformation on social media: A systematic literature review and future research directions. *AIS Transactions on Human-Computer Interaction*, *14*(2), 116–149. https://doi.org/10.17705/1thci.00164

Liang, Y., & Chou, T. (2025). Developmental pathways for academic knowledge about public health and social media misinformation: A systematic review through main path analysis. *Current Psychology*, *44*(19), 16078–16094. https://doi.org/10.1007/s12144-025-08298-6

Liu, A. K. C., & Kuru, O. (2025). Understanding the effects of visual misinformation: A systematic review of 10 years (2014–2024). *Mass Communication and Society*. Advance online publication. https://doi.org/10.1080/15205436.2025.2549716

Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: A review. *Information Fusion*, *107*, Article 102300. https://doi.org/10.1016/j.inffus.2024.102300

Llamas, N., & Cristòfol, F. J. (2023). Monkeypox, disinformation, and fact-checking: A review of ten Iberoamerican countries in the context of public health emergency. *Information*, *14*(7), Article 390. https://doi.org/10.3390/info14070390

Löffler, P. (2021). Review: Vaccine myth-buster – Cleaning up with prejudices and dangerous misinformation. *Frontiers in Immunology*, *12*, Article 663280. https://doi.org/10.3389/fimmu.2021.663280

Lojo Lendoiro, S., & Moreno-Sánchez, T. (2022). Occupational radiation and pregnancy: Reality or disinformation? A review of the literature and summary of current clinical guidelines. *Radiologia*, *64*(2), 128–135. https://doi.org/10.1016/j.rx.2021.11.004

López-Borrull, A., & Lopezosa, C. (2025). Mapping the impact of generative AI on disinformation: Insights from a scoping review. *Publications*, *13*(3), Article 33. https://doi.org/10.3390/publications13030033

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X. D. (2023). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *25*, Article e49255. https://doi.org/10.2196/49255

Luo, J., Xue, R., & Hu, J. (2020). COVID-19 infodemic on Chinese social media: A 4P framework, selective review and research directions. *Measurement and Control*, *53*(9–10), 2070–2079. https://doi.org/10.1177/0020294020960957

Ma, Z., & Ma, R. (2025). The role of narratives in countering health misinformation: A scoping review of the literature. *Health Communication*, *40*(11), 2353–2364. https://doi.org/10.1080/10410236.2025.2453451

Maity, A., & Jain, J. (2024). Disinformation and media ethics: Review of the web series *The Broken News*. *Media Asia*. Advance online publication. https://doi.org/10.1080/01296612.2024.2392378

Malhotra, A., Evans, N., Gao, J., Du, J. T., & Zheng, C. (2025). The issues caused by misinformation—How workers and organizations deal with it: A systematic literature review. *Journal of the Association for Information Science and Technology*. Advance online publication. https://doi.org/10.1002/asi.25016

Malik, M., Bauer-Maison, N., Guarna, G., & D'Souza, R. D. (2024). Social media misinformation about pregnancy and COVID-19 vaccines: A systematic review. *Medical Principles and Practice*, *33*(3), 232–241. https://doi.org/10.1159/000538346

Mang, V., Fennis, B. M., & Epstude, K. (2024). Source credibility effects in misinformation research: A review and primer. *Advances.in/Psychology*, *2024*(1), Article e00028. https://doi.org/10.56296/aip00028

Marcos-Vílchez, J. M., Muñiz Velázquez, J. A., & Sánchez-Martín, M. (2026). Effectiveness of training actions aimed at improving critical thinking in the face of mis- and disinformation: A systematic review. *Humanities & Social Sciences Communications*, *51*, Article 51.

Marecos, J., Tude Graça, D., Goiana-Da-Silva, F., Ashrafian, H., & Darzi, A. (2024). Source credibility labels and other nudging interventions in the context of online health misinformation: A systematic literature review. *Journalism and Media*, *5*(2), 702–717. https://doi.org/10.3390/journalmedia5020046

Martel, C., Rand, D. G., & Allen, J. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, *54*, Article 101710. https://doi.org/10.1016/j.copsyc.2023.101710

Meel, P., & Raj, C. (2025). A review of web infodemic analysis and detection trends across multi-modalities using deep neural network. *International Journal of Data Science and Analytics*, *20*(5), 4149–4176. https://doi.org/10.1007/s41060-025-00727-w

Meeran, R. A., Panachanathan, S., Daniel, R. A., Periasamy, P., & Choudhary, A. K. (2025). Understanding vaccine hesitancy in India: A narrative review of parental determinants, misinformation drivers and community-based interventions to improve

childhood immunization. *Journal of Experimental Zoology India*, *28*(2), 1107–1120. https://doi.org/10.51470/jez.2025.28.2.1107

Meza-Gómez, P., García-Tejada, J. E., Mamani-Calcina, J., Vera-Vasquez, C. G., Mamani-Berrios, N., Ortiz-Esparza, M. A., Cardona-Reyes, H., & Lara-Alvarez, C. (2023). Disinformation in social networks: A systematic review on fake news in times of pandemic. *CEUR-WS*.

Miri, A., Karimi-Shahanjarini, A., Afshari, M., Bashirian, S., Tapak, L., & Agley, J. (2024). Understanding the features and effectiveness of randomized controlled trials in reducing COVID-19 misinformation: A systematic review. *Health Education Research*, *39*(6), 495–506. https://doi.org/10.1093/her/cyae036

Miro-Llinares, F., Aguerri, J. C., & Alcacer-Guirao, R. (2023). Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'. *European Journal of Criminology*, *20*(1), 356–374. https://doi.org/10.1177/1477370821994059

Mirza, S. A., Sheikh, A. A. E., Barbera, M., Ijaz, Z., Javaid, M. A., Shekhar, R., Pal, S., & Sheikh, A. B. (2022). COVID-19 and the endocrine system: A review of the current information and misinformation. *Infectious Disease Reports*, *14*(2), 184–197. https://doi.org/10.3390/idr14020023

Morais, R., & Piñeiro-Naval, V. (2025). The presence of regional and local aspects about disinformation in scientific production in Spain and Portugal: A review of the state of the art. *Doxa Comunicacion*, *2025*(41), 341–368. https://doi.org/10.31921/doxacom.n41a2905

Mortimer, J. (2024). Turning the tide on misinformation: Fact-checking in the age of a 'million narratives'. *Byline Times*. https://bylinetimes.com/2024/12/29/turning-the-tide-on-misinformation-fact-checking-in-the-age-of-a-million-narratives/

Moss, C., Ross, K., & Proctor, A. (2023). Topical steroid withdrawal is not a myth. Comment on '#corticophobia: A review on online misinformation related to topical steroids'. *Clinical and Experimental Dermatology*, *48*(6), 697. https://doi.org/10.1093/ced/llad062

Muallem, M. Z., & Sayasneh, A. (2025). Debunking myths and misinformation in cervical cancer: A narrative review on navigating complex treatment choices in locally advanced cases and exploring beyond standard protocols. *Diagnostics*, *15*(9), Article 1174. https://doi.org/10.3390/diagnostics15091174

Muhanga, M., Jesse, A., & Ngowi, E. (2024). Community responses to corona virus disease (COVID-19) in Africa in the face of "infodemic": A scoping review. *Parasite Epidemiology and Control*, *25*, Article e00345. https://doi.org/10.1016/j.parepi.2024.e00345

Mündges, S., & Park, K. (2024). But did they really? Platforms' compliance with the Code of Practice on Disinformation in review. *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1786

Murphy, G., de Saint-Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*, *4*(6). https://doi.org/10.37016/mr-2020-130

Murunga, P. A. (2024). *Fake news and misinformation and their impacts on public perceptions of reality in Kenya: A literature review*. Bloomsbury Publishing.

Nan, X., Wang, Y., & Thier, K. (2022). Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine*, *314*, Article 115398. https://doi.org/10.1016/j.socscimed.2022.115398

Ng, J. Y., Liu, S., Maini, I., Pereira, W., Cramer, H., & Moher, D. (2023). Complementary, alternative, and integrative medicine-specific COVID-19 misinformation on social media: A scoping review. *Integrative Medicine Research*, *12*(3), Article 100975. https://doi.org/10.1016/j.imr.2023.100975

Ni, Z., Bousquet, C., Vaillant, P., & Jaulent, M. (2023). Rapid review on publicly available datasets for health misinformation detection. In J. Mantas, P. Gallos, E. Zoulias, A. Hasman, M. S. Househ, M. Charalampidou, & A. Magdalinou (Eds.), *Studies in Health Technology and Informatics* (Vol. 305). IOS Press. https://doi.org/10.3233/SHTI230439

Nurani, F., & Suprapto, A. (2025). A systematic literature review on the role of library science in combating disinformation. *IAFOR Journal of Education*, *13*(2), 161–189. https://doi.org/10.22492/ije.13.2.07

Nutter, S., Saunders, J. F., & Alberga, A. S. (2024). Weight stigma and health misinformation: A systematic review of research examining correlates associated with viewing The Biggest Loser. *Stigma and Health*, *9*(3), 337–348. https://doi.org/10.1037/sah0000457

O'Connor, C., & Murphy, M. (2021). Scratching the surface: A review of online misinformation and conspiracy theories in atopic dermatitis. *Clinical and Experimental Dermatology*, *46*(8), 1545–1547. https://doi.org/10.1111/ced.14679

O'Connor, C., O'Grady, C., & Murphy, M. (2022). Spotting fake news: A qualitative review of misinformation and conspiracy theories in acne vulgaris. *Clinical and Experimental Dermatology*, *47*(9), 1707–1711. https://doi.org/10.1111/ced.15222

O'Connor, C., Rafferty, S., & Murphy, M. (2022). A qualitative review of misinformation and conspiracy theories in skin cancer. *Clinical and Experimental Dermatology*, *47*(10), 1848–1852. https://doi.org/10.1111/ced.15249

O'Leary, A., O'Connor, C., Gibson, L., & Murphy, M. (2024). Pouring cold water on fake news: A qualitative review of misinformation related to burns first aid. *Journal of Burn Care & Research*, *45*(3), 753–756. https://doi.org/10.1093/jbcr/irad188

Ohland-Marsoner, B., Nistor, N., Corlatescu, D., Dascalu, M., & Trăuşan-Matu, S. (2023). Misinformation-based dialogical construction of misconceptions on the internet. A literature review based on automated publication analysis. In C. Damsa, M. Borge, E. Koh, & M. Worsley (Eds.), *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning* (pp. 352–359). International Society of the Learning Sciences.

Okuhara, T., Okada, H., Yokota, R., & Kiuchi, T. (2025). Effectiveness and determinants of narrative-based corrections for health misinformation: A systematic review. *Patient Education and Counseling*, *139*, Article 109253. https://doi.org/10.1016/j.pec.2025.109253

Oliveira, T., de Oliveira Cardoso, N., Machado, W., Gonçalves, R. A., Quinan, R., Salvador, E. Z., Almeida, C., & Paes, A. (2024). Confronting misinformation related to

health and the environment: A systematic review. *Journal of Science Communication*, *23*(1), Article 901. https://doi.org/10.22323/2.23010901

Onifade, A. B. (2023). Looking beyond the impressions of algorithms and fact-checking in fighting online misinformation: A literature review. *Education for Information*, *39*(1), 33–49. https://doi.org/10.3233/EFI-211568

Pai, M., Yellapurkar, S., & Shodhan Shetty, A. (2023). Infodemic in public health a reemerging public health threat: A scoping review. *F1000Research*, *12*, Article 687. https://doi.org/10.12688/f1000research.130687.1

Pandey, S., & Ghosh, M. (2023). Bibliometric review of research on misinformation: Reflective analysis on the future of communication. *Journal of Creative Communications*, *18*(2), 149–165. https://doi.org/10.1177/09732586231165577

Pandey, V., Vikrant, V., Vats, S., Yadav, S. P., Purohit, R. V., Yadava, R. L., Kaur, J., Gupta, A., Saini, A. S., Gupta, S., Deo Kumar, B., & Sharma, R. (2024). From misinformation to facts: A detailed review of fake news detection methods. *IEEE*. https://doi.org/10.1109/I3CEET61722.2024.10994139

Park, S., & Nan, X. (2025). Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI and Society*. Advance online publication. https://doi.org/10.1007/s00146-025-02620-3

Pascucci, A., Manna, R., Caterino, C., Masucci, V., Monti, J., Bhatia, A., & Shaikh, S. (2020). Is this hotel review truthful or deceptive? A platform for disinformation detection through computational stylometry. *European Language Resources Association*.

Patel, S. S., Moncayo, O. E., Conroy, K. M., Jordan, D., & Erickson, T. B. (2020). The landscape of disinformation on health crisis communication during the COVID-19 pandemic in Ukraine: Hybrid warfare tactics, fake media news and review of evidence. *Journal of Science Communication*, *19*(5), Article 50202. https://doi.org/10.22323/2.19050202

Patev, A. J., Hood, K. B., & Bartlett, L. A. (2021). Towards a better understanding of abortion misinformation in the USA: A review of the literature. *Culture, Health & Sexuality*, *23*(3), 285–300. https://doi.org/10.1080/13691058.2019.1706001

Peng, W., Lim, S., & Meng, J. (2023). Persuasive strategies in online health misinformation: A systematic review. *Information, Communication & Society*, *26*(11), 2131–2148. https://doi.org/10.1080/1369118X.2022.2085615

Pérez-Escolar, M., Lilleker, D., & Tapia-Frade, A. (2023). A systematic literature review of the phenomenon of disinformation and misinformation. *Media and Communication*, *11*(2), 76–87. https://doi.org/10.17645/mac.v11i2.6453

Peters, L. E. R., Charnley, G. E. C., Roberts, S., & Kelman, I. (2025). Public health disinformation, conflict, and disease outbreaks: A global narrative integrative review to guide new directions for health diplomacy. *Global Health Action*, *18*(1), Article 2562380. https://doi.org/10.1080/16549716.2025.2562380

Pian, W., Chi, J., & Ma, F. (2021). The causes, impacts and countermeasures of COVID-19 “infodemic”: A systematic review using narrative synthesis. *Information Processing & Management*, *58*(6), Article 102713. https://doi.org/10.1016/j.ipm.2021.102713

Plikynas, D., Rizgeliene, I., & Korvel, G. (2025). Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access*, *13*, 17583–17629. https://doi.org/10.1109/ACCESS.2025.3530688

Porter, E., Murphy, M., & O'Connor, C. (2023). Crying wolf: A qualitative review of misinformation and conspiracy theories in lupus erythematosus. *Lupus*, *32*(7), 887–892. https://doi.org/10.1177/09612033231174423

Rani, N., Das, P., & Bhardwaj, A. K. (2022). Rumor, misinformation among web: A contemporary review of rumor detection techniques during different web waves. *Concurrency and Computation: Practice and Experience*, *34*(1), Article e6479. https://doi.org/10.1002/cpe.6479

Rau, M. A., Premo, A. E., & Adams, Z. (2025). Systematic review of educational approaches to misinformation. *Educational Psychology Review*, *37*(2), Article 27. https://doi.org/10.1007/s10648-025-10012-8

Ravichandran, B. D., & Keikhosrokiani, P. (2023). Classification of Covid-19 misinformation on social media based on neuro-fuzzy and neural network: A systematic

review. *Neural Computing and Applications*, *35*(1), 699–717. https://doi.org/10.1007/s00521-022-07797-y

Reis, A. B. P., & de Oliveira Malaquias, F. F. (2025). Infodemic among students: A systematic literature review. *Journal of Technology Management and Innovation*, *20*(2), 92–102. https://doi.org/10.4067/s0718-27242025000200092

Resnick, P., Alfayez, A., Im, J., Gilbert, E. A., & Allen, J. (2023). Searching for or reviewing evidence improves crowd workers' misinformation judgments and reduces partisan bias. *Collective Intelligence*, *2*(2), Article 26339137231173407. https://doi.org/10.1177/26339137231173407

Revez, J., & Corujo, L. (2024). Scientists' behaviour towards information disorder: A systematic review. *Journal of Information Science*. Advance online publication. https://doi.org/10.1177/01655515241244460

Ries, M. (2022). The COVID-19 infodemic: Mechanism, impact, and counter-measures—A review of reviews. *Sustainability*, *14*(5), Article 2605. https://doi.org/10.3390/su14052605

Romy, R. W., Chen, J., & Wang, Y. (2025). Evaluating inclusiveness and diversity in health misinformation correction research: A scoping review. *Health Information and Libraries Journal*. Advance online publication. https://doi.org/10.1111/hir.12584

Ruyang, L., & Hedi, Y. (2025). Wellness misinformation on social media: A systematic review using social cognitive theory. *Health Communication*. Advance online publication. https://doi.org/10.1080/10410236.2025.2555614

Rynne, R., Murphy, M., & O'Connor, C. (2024). White lies? A qualitative review of internet-based vitiligo-related misinformation. *JEADV Clinical Practice*, *3*(2), 735–737. https://doi.org/10.1002/jvc2.339

Sadiq, S., & Mathew, S. K. (2022). The disaster of misinformation: A review of research in social media. *International Journal of Data Science and Analytics*, *13*(4), 271–285. https://doi.org/10.1007/s41060-022-00311-6

Saeidnia, H. R., Hosseini, E., Lund, B., Tehrani, M. A., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: A

systematic review of challenges and approaches. *Knowledge and Information Systems*, *67*(4), 3139–3158. https://doi.org/10.1007/s10115-024-02337-7

Sanaullah, A. R., Das, A., Das, A., Kabir, M. A., & Shu, K. (2022). Applications of machine learning for COVID-19 misinformation: A systematic review. *Social Network Analysis and Mining*, *12*(1), Article 114. https://doi.org/10.1007/s13278-022-00921-9

Sanfilippo, M. R., Zhu, X., & Yang, S. (2025). Sociotechnical governance of misinformation: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, *76*(1), 289–325. https://doi.org/10.1002/asi.24953

Schmid, P., Altay, S., & Scherer, L. D. (2023). The psychological impacts and message features of health misinformation: A systematic review of randomized controlled trials. *Zeitschrift für Psychologie*, *28*(3), 162–172. https://doi.org/10.1027/1016-9040/a000494

Segado-Fernández, S., Jiménez-Gómez, B., Jiménez-Hidalgo, P. J., Lozano Estevan, M. C., & Herrera-Peco, I. (2025). Disinformation about diet and nutrition on social networks: A review of the literature. *Nutricion Hospitalaria*, *42*(2), 366–375. https://doi.org/10.20960/nh.05533

Senteio, C., Fields, S. D., Pritam Singh, R. K., Namata Kamoga, R. M., Andrews, E., Gandsman, D., Halton, C., Rysinova, V., & Snow, S. (2025). Overcoming health misinformation in marginalized groups: A systematic review. *International Journal for Equity in Health*, *24*(1), Article 17. https://doi.org/10.1186/s12939-025-02657-2

Sharma, A., & Kumar, M. (2024). Infodemic, migration and social inequality: Review of the film *Bheed*. *Media Asia*, *51*(3), 512–518. https://doi.org/10.1080/01296612.2023.2210433

Sharma, A., Hasija, Y., Misra, R., Shyamasundar, R. K., & Chaturvedi, A. (2022). Misinformation–A challenge to medical sciences: A systematic review. In *Proceedings of the 2022 International Conference on Computational Intelligence and Data Analytics* (pp. 165–174). Springer. https://doi.org/10.1007/978-3-030-82469-3_14

Sharma, P. R., Wade, K. A., & Jobson, L. (2023). A systematic review of the relationship between emotion and susceptibility to misinformation. *Memory*, *31*(1), 1–21. https://doi.org/10.1080/09658211.2022.2120623

Skafle, I., Nordahl-Hansen, A., Quintana, D. S., Wynn, R., & Gabarron, E. (2022). Misinformation about COVID-19 vaccines on social media: Rapid review. *Journal of Medical Internet Research*, *24*(8), Article e37367. https://doi.org/10.2196/37367

Smith, R., Chen, K., Winner, D., Friedhoff, S., & Wardle, C. (2023). A systematic review of COVID-19 misinformation interventions: Lessons learned. *Health Affairs*, *42*(12), 1738–1746. https://doi.org/10.1377/hlthaff.2023.00717

Sreeraag, S., & Padinjappurathu Gopalan, S. P. (2023). From fake reviews to fake news: A novel pandemic model of misinformation in digital networks. *Journal of Theoretical and Applied Electronic Commerce Research*, *18*(2), 1069–1085. https://doi.org/10.3390/jtaer18020054

Stanciu, A., Ciupercă, E., & Stapleton, L. (2024). Can deepfakes benefit the metaverse in an era of disinformation? Insights from a systematic review. *IFAC-PapersOnLine*, *57*(3), 195–200. https://doi.org/10.1016/j.ifacol.2024.07.125

Starvaggi, I., Dierckman, C., Lorenzo-Luaces, L., & Adams, C. (2024). Mental health misinformation on social media: Review and future directions. *Current Opinion in Psychology*, *56*, Article 101738. https://doi.org/10.1016/j.copsyc.2023.101738

Stieglitz, S., Fromm, J., Kocur, A., Rostalski, F., Duda, M., Evans, A., Rieskamp, J., Sievi, L., Pawelec, M., Loh, W., Heesen, J., Fuchss, C., Eyilmez, K., Beverungen, D., Lehrer, C., & Trier, M. (2025). What measures can government agencies in Germany take against digital disinformation? A systematic literature review and ethical-legal discussion. In *Proceedings of the 2025 Conference on Digital Governance* (pp. 201–215). Springer. https://doi.org/10.1007/978-3-031-80125-9_19

Suarez-Lledo, V., & Álvarez-Gálvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research*, *23*(1), Article e17187. https://doi.org/10.2196/17187

Sun, Z., Yim, W., Uzuner, Ö., Xia, F., & Yetişgen, M. (2025). A scoping review of natural language processing in addressing medically inaccurate information: Errors, misinformation, and hallucination. *Journal of Biomedical Informatics*, *169*, Article 104866. https://doi.org/10.1016/j.jbi.2025.104866

Sundelson, A. E., Grönvall, G. K., Ackerman, G., Limaye, R., Watson, C., & Sell, T. K. (2025). Diplomacy disrupted: A mixed-methods analysis of Russian disinformation at

the Ninth Review Conference of the Biological and Toxin Weapons Convention. *Politics and the Life Sciences*, *44*(1), 28–48. https://doi.org/10.1017/pls.2025.3

Surjatmodjo, D., Unde, A. A., Cangara, H., & Sonni, A. F. (2024). Information pandemic: A critical review of disinformation spread on social media and its implications for state resilience. *Social Sciences*, *13*(8), Article 418. https://doi.org/10.3390/socsci13080418

Tadesse, T. T., Jara, D., & Kifle, D. (2026). Health misinformation in Ethiopia: Myths, media dynamics, public response, and policy implications: A narrative review. *Public Health Challenges*, *5*(1), Article e70181. https://doi.org/10.1002/puh2.70181

Thomas, D. D., Xu, L., Yu, B., Alanis, O., Adamek, P. J., Canton, P. I., Lin, X., Luo, P. Y., & Mullen, P. S. P. (2025). Physical activity misinformation on social media: Systematic review. *JMIR Infodemiology*, *5*, Article e62760. https://doi.org/10.2196/62760

Treadgold, B. M., Coulson, N. S., Campbell, J. L., Lambert, J., Pitchforth, E., & Adams, S. A. (2025). Quality and misinformation about health conditions in online peer support groups: Scoping review. *Journal of Medical Internet Research*, *27*, Article e71140. https://doi.org/10.2196/71140

Udry, J., Barber, S. J., & Alter, A. L. (2024). The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology*, *56*, Article 101736. https://doi.org/10.1016/j.copsyc.2023.101736

Utama Chandra, Y. U., & Maydian, N. (2021). Factors influencing disinformation on social media: A systematic literature review. In *Proceedings of the 2021 International Conference on Information Management and Technology* (pp. 512–517). IEEE. https://doi.org/10.1109/ICIMTech53080.2021.9535001

Vaidya, Y., Saini, J. R., Rathod, R., & Gaikwad, H. (2023). Multi-modal misinformation detection: An exhaustive review. In *Proceedings of the 2023 International Conference on Computing, Communication, Security and Intelligent Systems* (pp. 120–128). IEEE. https://doi.org/10.1109/ICCUBEA58933.2023.10392005

Valverde-Berrocoso, J., Gonzalez-Fernandez, A., Acevedo-Borrega, J., & Aguaded, I. (2022). Disinformation and multiliteracy: A systematic review of the literature. *Comunicar*, *30*(70), 93–105. https://doi.org/10.3916/C70-2022-07

Vivion, M., Reid, V., Trottier, V., Bergeron, F., Savard, I., Dionne, E., & Tourigny, A. (2025). Interventions to counter health misinformation among older people: Protocol for a scoping review. *JMIR Research Protocols*, *14*, Article e74138. https://doi.org/10.2196/74138

Vivion, M., Trottier, V., Bouhêlier, È., Goupil-Sormany, I., & Diallo, T. (2024). Misinformation about climate change and related environmental events on social media: Protocol for a scoping review. *JMIR Research Protocols*, *13*, Article e59345. https://doi.org/10.2196/59345

Vyas, P., Vyas, G., & Liu, J. (2021). Proliferation of health misinformation on social media platforms: A systematic literature review. *Issues in Information Systems*, *22*(3), 73–85. https://doi.org/10.48009/3_iis_2021_81-95

Wang, S., Su, F., Ye, L., & Jing, Y. (2022). Disinformation: A bibliometric review. *International Journal of Environmental Research and Public Health*, *19*(24), Article 16849. https://doi.org/10.3390/ijerph192416849

Wang, Y., Thier, K., & Nan, X. (2023). HPV vaccine misinformation online: A narrative scoping review. In *Proceedings of the 2023 International Conference on Health Communication* (pp. 45–58). Springer. https://doi.org/10.1007/978-3-031-24490-2_3

Westberry, C., Palmer, X., Potter, L., & Arai, K. (2023). Social media and health misinformation: A literature review. In *Proceedings of the 2023 International Conference on Media and Society* (pp. 210–225). Springer. https://doi.org/10.1007/978-3-031-47457-6_26

Whitehead, H. S., French, C. E., Caldwell, D. M., Letley, L., & Mounier-Jack, S. (2023). A systematic review of communication interventions for countering vaccine misinformation. *Vaccine*, *41*(5), 1018–1034. https://doi.org/10.1016/j.vaccine.2022.12.059

Wise, J. (2025). HPV vaccine safe and reduces risk of cervical cancer, anti-misinformation review finds. *BMJ*, *391*, Article r2479. https://doi.org/10.1136/bmj.r2479

Yan, X., Li, Z., Cao, C., Huang, L., Li, Y., Meng, X., Zhang, B., Yu, M., Huang, T., Chen, J., Li, W., Hao, L., Huang, D., Yi, B., Zhang, M., Zha, S., Yang, H., Yao, J., Qian, P., ... Shui, T. (2024). Characteristics, influence, prevention, and control measures of the

mpox infodemic: Scoping review of infodemiology studies. *Journal of Medical Internet Research*, *26*, Article e54874. https://doi.org/10.2196/54874

Yap, J. M., Barátné Hajdu, A. H., & Kiszl, P. (2024). Civic roles of libraries in combating information disorders in social media: A scoping review. *Education for Information*, *40*(1), 21–44. https://doi.org/10.3233/EFI-220038

Ying, W., Cheng, C., & Brewster, T. (2021). Public emotional and coping responses to the COVID-19 infodemic: A review and recommendations. *Frontiers in Psychiatry*, *12*, Article 755938. https://doi.org/10.3389/fpsyt.2021.755938

Yu, W., Payton, B., Sun, M., Jia, W., & Huang, G. (2023). Toward an integrated framework for misinformation and correction sharing: A systematic review across domains. *New Media & Society*, *25*(8), 2241–2267. https://doi.org/10.1177/14614448221116569

Zalpour, A., Hashemian, M., Geraei, E., & Zare-Farashbandi, F. (2024). Health information disorders models: A scoping review. *Iranian Journal of Nursing and Midwifery Research*, *29*(6), 637–648. https://doi.org/10.4103/ijnmr.ijnmr_414_23

Zhang, S., Zhou, H., & Zhu, Y. (2024). Have we found a solution for health misinformation? A ten-year systematic review of health misinformation literature 2013–2022. *International Journal of Medical Informatics*, *188*, Article 105478. https://doi.org/10.1016/j.ijmedinf.2024.105478

Zhao, S., Hu, S., Zhou, X., Song, S., Wang, Q., Zheng, H., Zhang, Y., & Hou, Z. (2023). The prevalence, features, influencing factors, and solutions for COVID-19 vaccine misinformation: Systematic review. *JMIR Public Health and Surveillance*, *9*, Article e40201. https://doi.org/10.2196/40201

Ziapour, A., Malekzadeh, R., Darabi, F., Yıldırım, M., Montazeri, N., Kianipour, N., & Nejhaddadgar, N. (2024). The role of social media literacy in infodemic management: A systematic review. *Frontiers in Digital Health*, *6*, Article 1277499. https://doi.org/10.3389/fdgth.2024.1277499

Ziemer, C., Rothmund, T., & Altay, S. (2024). Psychological underpinnings of misinformation countermeasures: A systematic scoping review. *Journal of Media Psychology*, *36*(6), 397–409. https://doi.org/10.1027/1864-1105/a000407

10.2 Other References

Altay, S., Berriche, M., Heuer, H., Farkas, J., & Rathje, S. (2023). A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-119

Arney, J. (2024). Data dump: Meta killed CrowdTangle. What does it mean for researchers, reporters? *University of Colorado Boulder, College of communication, media, Design and Information - Online Blog 23 Aug 2024*. https://www.colorado.edu/cmdi/news/2024/08/23/research-info-crowdtangle-disinformation-keegan

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., & West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour,* *6*(10), 1372–1380. https://doi.org/10.1038/s41562-022-01449-3

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Bentzen, N. (2025a). European democracy shield (Report No. PE 775.835). *European Parliamentary Research Service*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775835/EPRS_BRI(2025)775835_EN.pdf

Bentzen, N. (2025b). Information manipulation in the age of generative artificial intelligence (Report No. PE 779.259). *European Parliamentary Research Service*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/779259/EPRS_BRI(2025)779259_EN.pdf

Bhandari, B., Zafra-Tanaka, J. H., Mahapatra, P., Njelekela, M., Ramalingam, S., Pavlovska, I., et al. (2026). Misinformation on cardiovascular disease spreads through social networks: a scoping review. *BMJ Public Health*. https://doi.org/10.1136/bmjph-2025-003225

Bontcheva, K., Papadopoulos, S., Tsalakanidou, F., & Gallotti, R., Dutkiewicz, L., Krack, N., Teyssou, D., Nucci, S. F., Spangenberg, J., Srba, I., Aichroth, P., Cuccovillo, L. &

Verdoliva, L. (2024). Generative AI And Disinformation: Recent Advances, Challenges, And Opportunities (White Paper). *European Digital Media Observatory*. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

Bontcheva, K. & Posetti, J. (Eds). (2020). Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression. *Broadband Commission for Sustainable Development.* https://www.broadbandcommission.org/publication/balancing-act-countering-digital-disinformation/

Brashier, N. M., & Schacter, D. L. (2020). Aging in an Era of Fake News. *Current Directions in Psychological Science*, *29*(3), 316-323. https://doi.org/10.1177/0963721420915872

Brysse, K., Oreskes, N., O'Reilly, J., & Oppenheimer, M. (2013). Climate change prediction: Erring on the side of least drama? *Global Environmental Change, 23*(1), 327–337. https://doi.org/10.1016/j.gloenvcha.2012.10.008

Bursztyn, L., Rao, A., Roth, C., & Yanagizawa-Drott, D. (2023). Opinions as facts. *The Review of Economic Studies, 90*(4), 1832–1864. https://doi.org/10.1093/restud/rdac065

Carrella, F., Simchon, A., Edwards, M., & Lewandowsky, S. (2025). Warning people that they are being microtargeted fails to eliminate persuasive advantage. *Communications Psychology, 3*(1), Article 77. https://doi.org/10.1038/s44271-025-00188-8

Center for Countering Digital Hate. (2024a). Hate pays: How X accounts are exploiting the Israel-Gaza conflict to grow and profit. https://counterhate.com/research/hate-pays/

Center for Countering Digital Hate. (2024b). Rated not helpful: How X's community notes system falls short on misleading election claims. https://counterhate.com/wp-content/uploads/2024/10/CCDH.CommunityNotes.FINAL-30.10.pdf

Chadwick, A. (2017). *The hybrid media system: Politics and power*. Oxford University Press.

Chan, M.-p. S., & Albarracín, D. (2023). A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour, 7*, 1514–1525. https://doi.org/10.1038/s41562-023-01588-5

Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Choi, S., Anderson, A. A., Cagle, S., Long, M., & Kelp, N. (2023). Scientists' deficit perception of the public impedes their behavioral intentions to correct misinformation. *PLOS ONE*, *18*(8), e0287870. https://doi.org/10.1371/journal.pone.0287870

Combrink, H. M. V. E., & Mkungeka, P. (2025). Misfluencers, the Human Agents Behind AI-Driven Infodemics. *Journal of Health Communication*, 30(10–12), 349–355. https://doi.org/10.1080/10810730.2025.2538529

Cook, J. (2024, January 15). *America misled: How the fossil fuel industry deliberately misled Americans about climate change*. Skeptical Science. https://skepticalscience.com/america-misled.html

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., van der Linden, D., Roscoe, J. F., Ayravainen, L. & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour* 7, 2054–2057 (2023). https://doi.org/10.1038/s41562-023-01750-2

Dek, A., Kyrychenko, Y., van der Linden, S., & Roozenbeek, J. (2025). Mapping the online manipulation economy: A market perspective on digital manipulation may help improve online trust and safety. *Science, 390*(6778), 1112–1114. https://doi.org/10.1126/science.adw8154

de Quay, E., & Sethi, P. (2026). *The Reform UK party's approach to climate change and net zero in local councils*. Grantham Research Institute on Climate Change and the Environment, London School of Economics and Political Science. https://www.lse.ac.uk/granthaminstitute/wp-content/uploads/2026/03/Reform-UK-approach-to-climate-change-and-net-zero-in-local-councils.pdf

Drolsbach, c. P., Solovev, K. & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media, *PNAS Nexus*, Volume 3, Issue 7, July 2024, page 217, https://doi.org/10.1093/pnasnexus/pgae217

Ecker, U. K. H., Lewandowsky, S., Cook, J., et al. (2022). The psychological drivers of misinformation and its resistance to correction. *Nature Reviews Psychology, 1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y

Ecker, U. K. H., Tay, L. Q., Roozenbeek, J., van der Linden, S., Cook, J., Oreskes, N., & Lewandowsky, S. (2025). Why misinformation must not be ignored. *American Psychologist, 80*(6), 867–878. https://doi.org/10.1037/amp0001448

Essien, E. O. (2025). Climate Change Disinformation on Social Media: A Meta-Synthesis on Epistemic Welfare in the Post-Truth Era. *Social Sciences,* 14(5), 304. https://doi.org/10.3390/socsci14050304

European Commission. (2026, January 26). *Commission investigates Grok and X's recommender systems under the Digital Services Act* [Press release]. https://digital-strategy.ec.europa.eu/en/news/commission-investigates-grok-and-xs-recommender-systems-under-digital-services-act

European Commission. (2025, February 13). *The Code of Conduct on Disinformation*. https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

EEAS (European Union External Action Service). (2025). 3rd EEAS Report On Foreign Information Manipulation And Interference Threats: Exposing the architecture of FIMI operations https://www.eeas.europa.eu/sites/default/files/documents/2025/EEAS-3nd-ThreatReport-March-2025-05-Digital-HD.pdf

EEAS (European Union External Action Service). (2026). 4th EEAS Report On Foreign Information Manipulation And Interference Threats: Dismantling The FIMI House Of Cards. https://www.eeas.europa.eu/sites/default/files/2026/documents/EEAS%204th%20Threat%20Report_web%20version_1.pdf

Farrelly, M. C., Healton, C. G., Davis, K. C., Messeri, P., Hersey, J. C., & Haviland, M. L. (2002). Getting to the truth: Evaluating national tobacco countermarketing campaigns. *American Journal of Public Health, 92*, 901–907. https://doi.org/10.2105/AJPH.92.6.901

Fazio, L. K., Pillai, R. M. & Patel, D. (2022). The effects of repetition on belief in naturalistic settings, *Journal of Experimental Psychology: General,* 151(10): 2604–2613, https://doi.org/10.1037/xge0001211

Feuerriegel, S., DiResta, R., Goldstein, J. A., Kumar, S., Lorenz-Apreen, P., Tomz, M. & Pröllochs, N. (2023). Research can help to tackle AI-generated disinformation. *Nature Human Behaviour*, *7*(11), 1818–1821. https://doi.org/10.1038/s41562-023-01726-2

Flore, M. (2025). Synthetic Friends: AI Companions and the Future of Disinformation. *SSRN*: http://dx.doi.org/10.2139/ssrn.5920643

Full Fact. (2025). *Full Fact report 2025*. https://fullfact.org/policy/reports/full-fact-report-2025/

Gauthier, G., Hodler, R., Widmer, P., et al. (2026). The political effects of X's feed algorithm. *Nature*. https://doi.org/10.1038/s41586-026-10098-2

Gehrke, M., & Amit-Danhi, E. R. (2025). Gendered disinformation as violence: A new analytical agenda. *Harvard Kennedy School Misinformation Review*, *6*(3). https://doi.org/10.37016/mr-2020-177

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv.* https://doi.org/10.48550/arXiv.2301.04246

Hameleers, M. (2026). Spaces of Unmoderated Hate? The Legitimization of Anti-Immigration Narratives on Right-Wing Alternative Media. *Digital Journalism*, 14(3), 511–530. https://doi.org/10.1080/21670811.2026.2643289

Han, J., Cha, M., & Lee, W. (2020). Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School Misinformation Review, 1*(3). https://doi.org/10.37016/mr-2020-011

Hastuti, H., Maulana, H. F., Lawelai, H., & Suherman, A. (2025). Algorithmic influence and media legitimacy: A systematic review of social media's impact on news production. *Frontiers in Communication*, *10*, Article 1667471. https://doi.org/10.3389/fcomm.2025.1667471

Hiaeshutter-Rice, D., Chinn, S. & Chen, K. (2021). Platform Effects on Alternative Influencer Content: Understanding How Audiences and Channels Shape Misinformation Online. *Frontiers Political Science.* 3:642394. https://doi.org/10.3389/fpos.2021.642394

Huet, N. Sadeghi, M. & Labbe, C. (2025). Russian propaganda campaign targets France with AI-fabricated scandals, drawing 55 million views on social media. *Newsguard.* https://www.newsguardtech.com/special-reports/russian-propaganda-campaign-targets-france-with-ai-fabricated-scandals/

Hitzig, Z. (2026, February 11). OpenAI is making the mistakes Facebook made. I quit. *The New York Times*. https://www.nytimes.com/2026/02/11/opinion/openai-ads-chatgpt.html

Holtgrave, D. R., Wunderink, K. A., Vallone, D. M., & Healton, C. G. (2009). Cost–utility analysis of the national truth campaign to prevent youth smoking. *American Journal of Preventive Medicine, 36*, 385–388. https://doi.org/10.1016/j.amepre.2009.01.020

Hyzen, A., Van den Bulck, H., Puppis, M., Kulig, M. & Paulussen, S. (2026). Epistemic welfare and algorithmic recommender systems: overcoming the epistemic crisis in the digitalized public sphere. *Communication Theory*, Volume 36, Issue 1, Pages 46–57, https://doi.org/10.1093/ct/qtaf018

Jiménez Durán, R., Müller, K., & Schwarz, C. (2022). The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4230296

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., & Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences, 120*, e2210666120. https://doi.org/10.1073/pnas.2210666120

Lamb, W. F., Mattioli, G., Levi, S., Roberts, J. T., Capstick, S., Creutzig, F., Minx, J. C., Müller-Hansen, F., Culhane, T. & Steinberger, J. K. (2020). Discourses of climate delay. *Global Sustainability*, 3, e17. https://doi.org/10.1017/sus.2020.13

Lewandowsky, S. (2022). Fake news and participatory propaganda. In R. Pohl (Ed.), *Cognitive illusions* (pp. 324–340). Routledge. https://doi.org/10.4324/9781003154730-23

Lewandowsky, S. (2024). Truth and democracy in an era of misinformation. *Science, 386*. https://doi.org/10.1126/science.ads5695

Lewandowsky, S. (2025). Free speech, fact checking, and the right to accurate information. *Science, 387*(6734). https://doi.org/10.1126/science.adv4632

Lewandowsky, S., Cook, J., Ecker, U. K., et al. (2020). *The Debunking Handbook 2020*. Databrary. https://doi.org/10.17605/OSF.IO/NFSXW

Lewandowsky, S., Ecker, U. K. H., Cook, J., et al. (2024). Liars know they are lying: Differentiating disinformation from disagreement. *Humanities and Social Sciences Communications, 11*, 986. https://doi.org/10.1057/s41599-024-03503-6

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological science in the public interest : a journal of the American Psychological Society,* 13(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Oreskes, N., Risbey, J. S., Newell, B. R., & Smithson, M. (2015). Seepage: Climate change denial and its effect on the scientific community. *Global Environmental Change, 33*, 1–13. https://doi.org/10.1016/j.gloenvcha.2015.02.013

Littrell, S., Klofstad, C., Diekman, A., Funchion, J., Murthi, M., Premaratne, K., Seelig, M., Verdear, D., Wuchty, S., & Uscinski, J. E. (2023). Who knowingly shares false political information online?. *Harvard Kennedy School, Misinformation Review*. https://doi.org/10.37016/mr-2020-121

Lopez-Borrull, A., & Lopezosa, C. (2025). Mapping the impact of generative AI on disinformation: Insights from a scoping review. *Publications, 13*(3), 33. https://doi.org/10.3390/publications13030033

Lopez-López, E., Abels, C. M., Holford, D., Herzog, S. M., & Lewandowsky, S. (2025). Generative artificial intelligence–mediated confirmation bias in health information seeking. *Annals of the New York Academy of Sciences, 1550*(1), 23–36. https://doi.org/10.1111/nyas.15413

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., & Hertwig, R. (2022). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour, 7*, 74–101. https://doi.org/10.1038/s41562-022-01460-1

Lyons, B., Montgomery, J. M. & Reifler, J. (2024). Partisanship and Older Americans' Engagement with Dubious Political News, *Public Opinion Quarterly*, Volume 88, Issue 3, Fall 2024, Pages 962–990, https://doi.org/10.1093/poq/nfae044

Maertens, R., Roozenbeek, J., Simons, J. S., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., & van der Linden, S. (2025). Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications, 16.* https://doi.org/10.1038/s41467-025-57205-x

Maaß, S., Wortelker, J., & Rott, A. (2024). Evaluating the regulation of social media: An empirical study of the German NetzDG and Facebook. *Telecommunications Policy, 48*, 102719. https://doi.org/10.1016/j.telpol.2024.102719

Manor, I. (2025). AI Companions: The New Frontier of Disinformation. USC Center on Public Diplomacy, *CPD Online Blog 25 Nov 2025.* https://uscpublicdiplomacy.org/blog/ai-companions-new-frontier-disinformation

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5*(1), 47. https://doi.org/10.1186/s41235-020-00252-3

Martins-Rodal, B., & López Bolás, A. (2026). From search to imaginary: The algorithmic construction of the refugee figure and the reproduction of bias in the era of generative AI. The Spanish case study of Google AI overviews and Gemini. *Frontiers in Communication*, *11*, Article 1739665. https://doi.org/10.3389/fcomm.2026.1739665

Metzler, H., & Garcia, D. (2024). Social Drivers and Algorithmic Mechanisms on Digital Media. *Perspectives on Psychological Science*, *19*(5), 735-748. https://doi.org/10.1177/17456916231185057

Meyrowitsch, D. W., Jensen, A. K., Sørensen, J. B., & Varga, T. V. (2023). AI chatbots and (mis)information in public health: impact on vulnerable communities. *Frontiers in public health*, 11, 1226776. https://doi.org/10.3389/fpubh.2023.1226776

Morosoli, S. & Humprecht, E. (2025). Motivations behind misinformation engagement: approving, disapproving, and ignoring. A study on individual characteristics in connection with supporting and renouncing online misinformation. *Journal Of Elections, Public Opinion And Parties.* Vol. 35, No. 3, pp. 360–383. https://doi.org/10.1080/17457289.2025.2514200

Motyl, M., Allen, J., Gurley, S. & Bonilla, S. (2025). Platform Datasets: Challenges, Insights and Examples for Researchers under Article 40 of the Digital Services Act. *EDMO Reports and Analyses, 21 Aug 2025.* https://edmo.eu/publications/platform-datasets-challenges-insights-and-examples-for-researchers-under-article-40-of-the-digital-services-act/

Müller, K., & Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association, 19*(4), 2131–2167. https://doi.org/10.1093/jeea/jvaa045

Neff, T., & Pickard, V. (2023). Building Better Local Media Systems: A Comparative Policy Discourse Analysis of Initiatives to Renew Journalism Around the World. *Journalism Studies*, 24(15), 1877–1897. https://doi.org/10.1080/1461670X.2023.2253928

Ó Fathaigh, R., Helberger, N., & Appelman, N. (2021). The perils of legally defining disinformation. *Internet Policy Review*, 10(4). https://doi.org/10.14763/2021.4.1584

Ofcom. (2026). Adults' Media Use and Attitudes Report. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/media-literacy-research/adults/adults-media-use-and-attitudes-2026/adults-media-use-and-attitudes-2026-report.pdf

Ognyanova, K., Lazer, D., Robertson, R. E. & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. Harvard Kennedy School, Misinformation Review. https://doi.org/10.37016/mr-2020-024

Parr, C. S. (2025). Does social media undermine trust? Institutional trust in civil society and governance institutions. *Journal of Public Policy,* 45(4), 737–760. https://doi.org//10.1017/S0143814X25100834

Pennycook, G., Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, 2021; vol. 25,issue 5, pp. 388-402 https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(21)00051-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS1364661321000516%3Fshowall%3Dtrue

Piccardi, T., Saveski, M., Jia, C., Hancock, J., Tsai, J. L., & Bernstein, M. S. (2025). Reranking partisan animosity in algorithmic social media feeds alters affective polarization. *Science, 390*(6776). https://doi.org/10.1126/science.adu5584

Pillai, R. M., & Fazio, L. K. (2025). Repeated by many versus repeated by one: Examining the role of social consensus in the relationship between repetition and belief. *Journal of Applied Research in Memory and Cognition*, 14(2), 154–166. https://doi.org/10.1037/mac0000166

Plikynas, D., Rizgelienė, I., & Korvel, G. (2025). Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through

machine learning. *IEEE Access*, vol. 13, pp. 17583-17629, 2025, https://doi.org/10.1109/ACCESS.2025.3530688

Posetti, J., Shabbir, N., Maynard, D., Bontcheva, K., & Aboulez, N. (2021). The Chilling: Global trends in online violence against women journalists. UNESCO-ICFJ Research Discussion Paper. https://unesdoc.unesco.org/ark:/48223/pf0000377223

Prama, T. T., Bagchi, C., Kalakonnavar, V., Krauß, P., & Grabowicz, P. A. (2025). *Political biases on X before the 2025 German federal election*. arXiv. https://doi.org/10.48550/arXiv.2503.02888

Purnat, T. D., Wilhelm, E., Scales, D., Wardle, C., Bastien, S., Ganatra, B., Lavelanet, A., Mburu, G., Tamrat, T. & Nihlén, Å. (2025). Impacts of Sexual and Reproductive Health and Rights Misinformation in Digital Spaces on Human Rights Protection and Promotion: Scoping Review. *JMIR Infodemiology* 2025,vol. 5: e83747. https://doi.org/10.2196/83747

Purnat, T. D., Vacca, P., Czerniak, C., Ball, S., Burzo, S., Zecchin, T., Wright, A., Bezbaruah, S., Tanggol, F., Dubé, È., Labbé, F., Dionne, M., Lamichhane, J., Mahajan, A., Briand, S., & Nguyen, T. (2021). Infodemic Signal Detection During the COVID-19 Pandemic: Development of a Methodology for Identifying Potential Information Voids in Online Conversations. *JMIR infodemiology*, *1*(1), e30971. https://doi.org/10.2196/30971

Razuvayevskaya, O., & Bontcheva, K. (2026). Truth with a twist: The rhetoric of persuasion in professional vs. community-authored fact-checks. In *Proceedings of the ACM Web Conference 2026 (WWW '26)* (pp. 13–17). *ACM*. https://doi.org/10.1145/3774904.3792938

RCNi. (2025, October 2). *Ex-nurse 'adversely influenced' daughter who died after refusing chemo*. https://rcni.com/nursing-older-people/newsroom/news/ex-nurse-adversely-influenced-daughter-who-died-after-refusing-chemo-217541

Rheault, L., Rayment, E., & Musulan, A. (2019). Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics,* 6(1). https://journals.sagepub.com/doi/10.1177/2053168018816228

Roeloffs, M. W. (2023). *Musk: X* Users Won't Make Money Off Corrected Tweets. *Forbes*. https://www.forbes.com/sites/maryroeloffs/2023/10/29/musk-x-users-wont-make-money-off-corrected-tweets/

Roozenbeek, J., Culloty, E., & Suiter, J. (2023). Countering misinformation: evidence, knowledge gaps, and implications of current interventions. *European Psychologist,* 28(3), 189–205. https://doi.org/10.1027/1016-9040/a000492

Saner, K. (2026). Introduction to the special issue on gendered disinformation. *Journal of Gender Studies*, vol. 35(3), pp. 603–612. https://doi.org/10.1080/09589236.2026.2629862

Santini, R. M., Salles, D., Mattos, B., Moreira, A., Mello, D., Haddad, J. G., Fernandez, M. E., Dias, B., & Benzecry, L. (2025). Transparency under threat: progress and setbacks between CrowdTangle and the Meta Content Library. *NetLab UFRJ*. netlab.eco.ufrj.br/post/transparencia-sob-ameaca

Sato, Y., & Wiebrecht, F. (2024). Disinformation and Regime Survival. *Political Research Quarterly,* 77(3), 1010-1025. https://doi.org/10.1177/10659129241252811

Sauer, M. A., Truelove, S., Gerste, A. K. & Limaye, R. J. (2021). A Failure to Communicate? How Public Messaging Has Strained the COVID-19 Response in the United States. *Health Secur.* 2021 Jan-Feb;19(1):65-74. https://doi.org/10.1089/hs.2020.0190

Scholtens, M., Pizano, P., Karpawich, M., & Kuckes, G. (2024). The Disinformation Economy. *The Carter Center; McCain Institute for International Leadership.* https://www.cartercenter.org/resources/pdfs/news/peace_publications/democracy/the-disinformation-economy-mccain-may-2024.pdf

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative AI. *PNAS Nexus, 3*(3), Article pgae035. https://doi.org/10.1093/pnasnexus/pgae035

Simchon, A., Sutton, A., Edwards, M., & Lewandowsky, S. (2023). Online reading habits can reveal personality traits: Towards detecting psychological microtargeting. *PNAS Nexus, 2*(7), Article pgad191. https://doi.org/10.1093/pnasnexus/pgad191

Simonov, A., Sacher, S., Dubé, J.-P., & Biswas, S. (2020). *The persuasive effect of Fox News: Non-compliance with social distancing during the Covid-19 pandemic* (NBER Working Paper No. 27237). National Bureau of Economic Research. https://doi.org/10.3386/w27237

Skibinski, M. (2021). Special report: Top brands are sending $2.6 billion to misinformation websites each year. *NewsGuard.* https://www.newsguardtech.com/special-reports/brands-send-billions-to-misinformation-websites-newsguard-comscore-report/

Spampatti, T., Pillai, R., Globig, L. K., Sternisko, A. & Van Bavel, J. J. (on behalf of the Center for Conflict and Cooperation) (2025). Disinformation Is A Systemic Risk To Human Rights: Input To The Study On "The Impact Of Disinformation On The Enjoyment And Realization Of Human Rights" Of The UN Human Rights Council Advisory Committee https://www.ohchr.org/sites/default/files/documents/hrbodies/hrcouncil/advisorycommittee/cfi-disinformation/subm-impact-disinformation-enjoyment-aca-center-conflict-cooperation.pdf

Sultan, M., Tump, A. N., Ehmann, N., Lorenz-Spreen, P., Hertwig, R., Gollwitzer, A., & Kurvers, R. H. J. M. (2024). Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, *121*(47), Article e2409329121. https://doi.org/10.1073/pnas.2409329121

Supran, G., & Oreskes, N. (2017). Assessing ExxonMobil's climate change communications (1977–2014). *Environmental Research Letters, 12*(8), 084019. https://doi.org/10.1088/1748-9326/aa815f

Supran, G., & Oreskes, N. (2021). Rhetoric and frame analysis of ExxonMobil's climate change communications. *One Earth,* *4*(5), 696–719. https://doi.org/10.1016/j.oneear.2021.04.014

Surjatmodjo, D., Unde, A. A., Cangara, H., & Sonni, A. F. (2024). Information Pandemic: A Critical Review of Disinformation Spread on Social Media and Its Implications for State Resilience. *Social Sciences*, *13*(8), 418. https://doi.org/10.3390/socsci13080418

Szakács, J., & Bognár, É. (2021). The impact of disinformation campaigns about migrants and minority groups in the EU (Report No. PE 653.641). European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/653641/EXPO_IDA(2021)653641_EN.pdf

Terzian, G. (2025). The epistemic dangers of journalistic balance. *Episteme*, 22(4), 931–951. https://doi.org/10.1017/epi.2024.45

Tian, L., & Rizoiu, M. (2025, February 24). How we tricked AI chatbots into creating misinformation despite safety measures. *The Conversation*. https://theconversation.com/how-we-tricked-ai-chatbots-into-creating-misinformation-despite-safety-measures-264184

Tiller, N. B., Marcon, A. R., Zenone, M., Kidd, K. E., Jeukendrup, A. E., Master, Z. & Caulfield, T. (2026). Generative artificial intelligence-driven chatbots and medical misinformation: an accuracy, referencing and readability audit. *BMJ Open* 2026;16:e112695. https://doi.org/10.1136/bmjopen-2025-112695

Torre, L., Ramos, G., Noronha, M., & Jerónimo, P. (2024). Sourcing Local Information in News Deserts. *Journalism and Media*, 5(3), 1228-1243. https://doi.org/10.3390/journalmedia5030078

UK Parliament, Foreign Affairs Committee. (2026, March). *Foreign disinformation: "The new warfare and open liberal democracies are sitting ducks"*. https://committees.parliament.uk/committee/78/foreign-affairs-committee/news/212849/foreign-disinformation-the-new-warfare-and-open-liberal-democracies-are-sitting-ducks

United Nations. (2025). *UN global risk report*. https://digitallibrary.un.org/record/4090928/files/1441689-en.pdf

van der Linden, S., Albarracín, D., Fazio, L., Freelon, D., Roozenbeek, J., Swire-Thompson, B., & Van Bavel, J. (2025). Using Psychological Science to Understand and Fight Health Misinformation: An APA Consensus Statement. *American Psychologist. Advance online publication*. https://doi.org/10.1037/amp0001598

van der Linden, S., Louison-Lavoy, D., Blazer, N., Noble, N., & Roozenbeek, J. (2026). Prebunking misinformation techniques in social media feeds: Results from an Instagram field study. *Harvard Kennedy School (HKS) Misinformation Review*. https://doi.org/10.37016/mr-2020-193

Vincent, E., & Crisan, D. (2026). *What our second measurement says about misinformation on major platforms in Europe*. Science Feedback. https://science.feedback.org/second-measurement-mis-disinformation-major-platforms-europe/

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science* vol. 359, issue 6380, pp. 1146-1151(2018). https://doi.org/10.1126/science.aap9559

Waldman, S. & E&E News. (2025). Elon Musk's Grok Chatbot Has Started Reciting Climate Denial Talking Points. *Scientific American*. https://www.scientificamerican.com/article/elon-musks-ai-chatbot-grok-is-reciting-climate-denial-talking-points/

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (Report No. DGI(2017)09). Council of Europe. https://www.firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf

Weikmann, T., Wouters, F., Tulin, M., Hameleers, M., de Vreese, C., Zarouali, B., & Opgenhaffen, M. (2026). On the same page? Experts are mostly, but not always aligned about disinformation in times of generative AI. *Harvard Kennedy School Misinformation Review,* 7(2). https://doi.org/10.37016/mr-2020-196

Williams, K. (2024). California legislative session roundup: Which key privacy and AI bills were enacted and which were vetoed? *Electronic Privacy Information Center.* https://epic.org/california-legislative-session-roundup-which-key-privacy-and-ai-bills-were-enacted-and-which-were-vetoed/

Wire. (2025). What the CLOUD Act really means for EU data sovereignty. https://wire.com/en/blog/cloud-act-eu-data-sovereignty

World Economic Forum. (2026). *The global risks report 2026* (21st ed.). https://www.weforum.org/publications/global-risks-report-2026/

Wouters, F., & Opgenhaffen, M. (2024). Regional facts matter: A comparative perspective of sub-state fact-checking initiatives in Europe. *Media and Communication*, *12*, Article 8758. https://doi.org/10.17645/mac.8758

Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H. and Lee, Y. (2025). The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). *Association for Computing Machinery, New York, NY, USA*, Article 13, 1–17. https://doi.org/10.1145/3706598.3713429