

FalseWebs Network Policy Paper: Understanding and Addressing Misinformation in Scotland

Submitted to: Scottish Government

Prepared by: Royal Society of Edinburgh (RSE) FalseWebs Research Network

Date: August 2025

About the [FalseWebs Network](#):

The FalseWebs Network brings together researchers and practitioners working to understand and address the spread of misinformation through research, collaboration, and public engagement. The network includes contributors from a range of academic and professional backgrounds across the UK and beyond.

For queries relating to this report, please contact:

Dr. Marina Wimmer

Edinburgh Napier University

M.Wimmer@Napier.ac.uk

FalseWebs Network Policy Paper: Understanding and Addressing Misinformation in Scotland

This policy paper was developed under the leadership of Dr Marina Wimmer at Edinburgh Napier University. The project was supported by the Royal Society of Edinburgh (RSE).

Contributors: Dr Marina Wimmer, Dr Faye Skelton, Theodore Carlson Webster, Dr Md Zia Ullah, Katarina Alexander, and Emily Spencer (Edinburgh Napier University); Dr Alice Howarth (University of Liverpool); Michael Marshall (Good Thinking Society); Dr Natalia Pasternak (Columbia University, Instituto Questao de Ciencia (Brazil)); Dr David Robertson (University of Strathclyde); Pip Brown and Dr Michaela Gummerum (University of Warwick); Dr Yvonne Skipper (University of Glasgow); Jeremy Wright (Toronto Metropolitan University); Dr Ariana Modirrousta-Galian (University College London); Mansi Pattni, Dr Tina Seabrooke and Professor Philip Higham (University of Southampton); Anna Talley (University of Edinburgh); Dr Chantal den Daas and Professor Marie Johnston (University of Aberdeen); Professor Stephan Lewandowsky (University of Bristol); Dr Charlotte Bücken, Dr Paul Riesthuis, and Professor Henry Otgaar (KU Leuven Research Foundation); Dr Carolina Are (Northumbria University); and Professor John Collomosse (University of Surrey).



Contents

Introduction.....1

Executive Summary.....3

Chapter 1: The Case of Fake News

1.1 What’s the Harm: Conspiracy Beliefs
Continuum.....5

1.2 When Science is Ignored: Global Examples of Misinformation and
Economic
Impact.....8

1.3 Who Falls for Fake News?.....12

1.4 Children’s Trust in the Digital Age: Fostering Digital Literacy in
Adolescents.....15

1.5 Young People and Misinformation: Building Digital Resilience
Through
Education.....18

Chapter 2: A Brief History of Information Disorder

2.1 History of Information Disorder: Production, Distribution, and
Form.....21

Chapter 3: Detecting and Reporting Misinformation: Considerations for Digital Ethics and Education Policies

3.1 Reasoning in Detecting Fake News and Misinformation.....25

3.2 Flagging AI – Generated Images.....29

3.3 Making People Resistant to Fake News.....31

3.4 Remembering Fake News and the Role of Fact-Checking.....34

3.5 Visual Language in Media Literacy Programmes.....37

Chapter 4: Messaging

4.1 Messaging and Behaviour Change.....40

4.2 Misinformation and Peoples Beliefs.....43

**4.3 Flagging, De-Platforming and Appeals: The Moderation of Nuanced
Content on Instagram and TikTok.....46**

Chapter 5: Fake News Algorithms

5.1 Fake News Warning Labels.....49

5.2 Content Provenance in the Fight Against Misinformation.....52

Acknowledgements.....56

References.....57

Introduction

The proliferation of fake news and misinformation poses a growing challenge to democratic societies. In Scotland, this issue intersects with public trust, digital safety, media literacy, and the health of civic discourse. The Scottish Government has already taken steps to address the digital information environment through strategy papers such as the [*National Digital Ethics Public Panel Insight Report \(2021\)*](#) and recommendations from the [*Public Interest Journalism Working Group \(2022\)*](#). These emphasise media literacy, safer online environments, algorithmic transparency, and accessible reporting mechanisms for abuse and harassment.

Recent contributions from the [*International Council of Education Advisers \(2021–2023\)*](#) further stress the importance of integrating digital literacy and news literacy into the Scottish Curriculum for Excellence. In light of technological advances – including AI-generated content – the ability to critically evaluate information is essential for young people and adults alike. Beyond education, there have been wider policy calls for greater “information resilience” across society, including training for political actors and even the appointment of a disinformation commissioner.

This policy paper brings together evidence from researchers across psychology, communication, design, behavioural science, and education to inform the Scottish Government’s ongoing response to misinformation. Its goal is to support a cohesive, evidence-based approach that protects the public from harm, strengthens media and digital literacy, and empowers citizens to navigate the complex information landscape. The content reflects interdisciplinary expertise and international best

practices, aligning with global initiatives such as those from the OECD, the European Commission, the United Nations, and the World Health Organization.

In this paper, we focus on five key areas:

1. **The case of fake news:** understanding harms and the psychology of belief
2. **The history of information disorder:** the production, distribution, and form of fake news
3. **Detecting and Reporting Misinformation:** Considerations for Digital Ethics and Education Policies
4. **Messaging:** Considerations for Digital Ethics and Health & Social Care Innovations in labelling and provenance
5. **Fake News Algorithms:** Tools for combatting fake news and misinformation

Each section concludes with specific, actionable recommendations for Scottish policy development. This paper is intended to inform future parliamentary debate and practice, including through SPICe (Scottish Parliament Information Centre), and to guide collaboration between government, educators, technologists, and civil society.

Executive Summary

This policy paper addresses the growing societal threat of misinformation and disinformation – collectively known as *information disorder* – and explores how Scotland can respond effectively through evidence led interventions, regulation, and education.

Drawing on recent research in psychology, behavioural science, design, AI, and digital governance, the paper outlines the mechanisms by which false information spreads, the harms it causes, and how individuals and institutions can build resilience.

Key Points

- People are not universally susceptible; most can detect misinformation but tend to rely on intuition. Interventions like inductive learning, cognitive reflection, and prebunking can improve accuracy.
- Adolescents and young adults are particularly vulnerable to misleading content, despite high levels of digital engagement.
- AI generated media, particularly images, pose new challenges for detection, trust, and verification. Training and provenance-based technologies offer viable solutions.
- Misinformation not only distorts public understanding but can also implant false memories and alter behaviours, even after correction.

- Moderation practices on platforms like Instagram and TikTok disproportionately affect marginalised communities. Transparent appeals processes and user controls are needed.
- Technological interventions such as content provenance (e.g. C2PA) and independent warning labels offer promising ways to signal trustworthiness and reduce harm.

Recommendations

- Digital Literacy: Expand media education to include visual, AI, and rhetorical literacy, starting early and continuing into adulthood.
- Behavioural Approaches: Apply evidence-based techniques to help individuals better assess, question, and respond to false or misleading content.
- Platform Governance: Improve moderation processes, appeals transparency, and user agency; encourage rehabilitative, not punitive, moderation models.
- Technological Solutions: Support implementation of provenance standards and independent fake news labelling.
- Public Messaging: Tailor government communications using behavioural science insights to support trust and healthy information environments.

This paper provides a roadmap for strengthening Scotland's digital resilience and civic integrity, complementing ongoing government work on AI, education, and public health.

Chapter 1: The Case of Fake News

1.1 What's the Harm: Conspiracy Beliefs Continuum

One significant consequence of exposure to fake news is the increased likelihood of adopting conspiracy beliefs. While these beliefs are sometimes dismissed as trivial or naïve – for example, ideas such as flat earth theory or sovereign citizenship – research shows that belief in one conspiracy theory increases the risk of believing in others, including those with serious health and social consequences.

For example, some individuals believe that a cure for cancer exists but is being deliberately withheld. Such beliefs can lead people to reject proven medical treatments, with life-threatening results. Conspiracy beliefs can undermine trust in institutions, including healthcare, science, and public policy, with implications for both individual and public safety.

People who endorse one conspiracy theory are more likely to adopt others, forming what has been described as a "monological belief system". This progression can lead individuals from relatively benign beliefs to more harmful, socially corrosive, and extreme worldviews.

In many cases, conspiracy belief is motivated by a desire for control, certainty, and community – especially during times of uncertainty. As individuals become more engaged in conspiracy communities, they may begin to withdraw from their existing social networks and align more closely with groups that reinforce their beliefs. Over time, this process can increase vulnerability to radicalisation and, in some cases, violent extremism.

This continuum of harm – from initial distrust to the potential for violence – makes early recognition and intervention critical. Conspiracies that target marginalised groups or promote hatred pose particular risks for social cohesion and public safety.

Understanding how and why conspiracy beliefs develop is essential for designing effective, evidence-based interventions. It is important not only to focus on the content of the beliefs but also to acknowledge the psychological needs they may fulfil. Compassionate responses that seek to understand and address these underlying drivers, while remaining clear-eyed about the potential harms, offer a more sustainable policy approach.

Key Points:

- Belief in conspiracy theories can lead to serious health, social, and security risks.
- Even seemingly harmless conspiracy theories can escalate into more dangerous beliefs.
- Conspiracy thinking often fulfils psychological needs, including a desire for safety, control, and community.
- As individuals become more involved in conspiracy communities, they may withdraw from mainstream social networks and become more vulnerable to radicalisation.
- Understanding and addressing the underlying drivers of conspiracy belief is essential to early intervention and harm reduction. Early recognition and compassionate, evidence-based interventions

addressing underlying motivations are essential to mitigate harms to individuals and society.

Recommendations:

- Continue investing in user-centred research to identify individual traits linked to susceptibility to fake news and misinformation.
- Develop reliable tests that measure the likelihood of falling for and sharing fake news, and use these to raise awareness among low performers.
- Identify and study high performers to understand what psychological characteristics protect them from misinformation.
- Use insights from these findings to develop tailored interventions, such as emotional intelligence or critical thinking training.
- Implement such interventions in schools, ideally around the time young people gain access to social media, to foster early resistance to misinformation.

1.2 When Science is Ignored: Global Examples of Misinformation and Economic Impact

The relationship between government oversight and scientific truth is complex. While regulation plays a necessary role in shaping research ethics, safety, and funding priorities, misinformation and disinformation can also be used by states to manipulate or suppress science in service of political or ideological goals. When complexity is denied in favour of simplistic directives, the consequences can be severe – not only for scientists, but for society as a whole.

Case Study 1: Soviet Union and the banning of genetics

One of the most infamous examples is the suppression of genetics in the Soviet Union under Joseph Stalin. During the 1940s and 1950s, the agronomist Trofim Lysenko convinced Stalin that modern genetics – particularly the work of Darwin, Mendel, and Morgan – was a “bourgeois” science, incompatible with socialist values. In its place, Lysenko promoted pseudoscientific ideas such as training plants to survive the cold through repeated exposure – despite no empirical evidence supporting this claim.

The result was catastrophic: widespread crop failure, famine, and the loss of decades of genetic research. Students were barred from learning about genetics and evolutionary biology; universities were forbidden to teach or study these subjects. Scientists who opposed Lysenkoism were arrested,

exiled, or executed, and generations of Soviet scientists graduated without fundamental knowledge in biology.

Case Study 2: HIV/AIDS denialism in South Africa

More recently, during the AIDS pandemic in the early 2000s, South Africa experienced one of the most damaging cases of state-driven misinformation. President Thabo Mbeki publicly questioned the link between HIV and AIDS, cast doubt on the safety of antiretroviral treatments, and obstructed evidence-based public health initiatives. As a result, South Africa delayed the adoption of life-saving medications, with devastating consequences. More than a decade later, the country still has one of the highest HIV prevalence rates globally, with approximately 18.8% of the adult population living with HIV – nearly 5.5 million people. Denialism led directly to avoidable deaths.

Case Study 3: Organic-only agriculture in Sri Lanka

A third example is Sri Lanka's 2022 attempt to abruptly convert all agriculture to organic farming. Informed by political ideology rather than scientific consensus, the government banned the import and use of synthetic fertilisers and pesticides overnight. The transition, lacking a robust evidence base, severely disrupted food production. A country that was previously self-sufficient in rice was forced to import it, while tea – its major export – suffered reduced yields, contributing to economic instability and worsening food insecurity.

These examples demonstrate that when science is overridden by ideology or misinformation, the effects are not only local – they are global, generational, and deeply damaging.

Key Points:

- Political misuse of scientific authority can have devastating consequences for public health, food security, and scientific progress.
- Historical and contemporary examples show how misinformation or ideologically driven interference in science can lead to unnecessary suffering and economic collapse.
- Safeguards are needed to ensure scientific evidence remains central to decision-making, and that governments resist substituting ideology for expertise.
- Transparent dialogue between policymakers, scientists, and the public helps protect both science and society.

Recommendations:

- Ensure that scientific evidence remains central to policymaking, especially in health, agriculture, and education.
- Strengthen institutional safeguards that prevent political interference in research and the public communication of science.
- Invest in public engagement initiatives to build understanding of the scientific process and counter mistrust.

- Promote open, transparent dialogue between government, scientists, and the public to prevent the misuse of scientific authority.
- Support international knowledge-sharing to resist the spread of pseudoscience and respond more effectively to global challenges.

1.3 Who Falls for Fake News?

The impact of deliberate misinformation on major socio-political events is a growing societal concern. Such misinformation, in the form of fake news, is thought to have had a measurable impact on major global events, from the UK's EU referendum to the 2016 U.S. presidential election and the public response to COVID-19. These effects have been magnified by the rise of social media, where users are not only exposed to misinformation but also play a central role in spreading it, often unknowingly.

Algorithmic fact-checking and warning labels alone are not sufficient. In some cases, such interventions may even backfire, increasing the perceived accuracy of false content that has not been flagged.

This highlights the importance of a user-centred approach: one that investigates why some individuals are more susceptible to fake news than others, and what psychological traits can help protect against its influence.

Research suggests that individual differences play a significant role in fake news susceptibility. People with higher levels of analytical thinking and emotional intelligence are generally better able to spot misleading content and avoid sharing it. These traits help individuals evaluate emotionally charged or exaggerated claims more critically and resist cognitive biases.

To move from insight to action, it is important to develop reliable tools that assess an individual's likelihood of falling for or sharing fake news. These tools can serve dual purposes:

1. Raising awareness among those who are more vulnerable

2. Identifying strengths in those who are more resistant, which can then inform intervention design

High performers on misinformation detection tasks (those least likely to fall for fake news) offer a valuable model for developing targeted interventions. By identifying the traits that protect these individuals (e.g., high emotional intelligence, critical thinking, cognitive reflection), it is possible to design evidence-based training to improve detection accuracy in more vulnerable groups.

To be most effective, such training should be integrated early – ideally in adolescence, when young people first engage with digital media environments. Embedding psychological resilience into education has the potential to create a long-lasting societal defence against misinformation.

Key Points

- The influence of fake news on major political and public health events remains a critical concern, amplified by global social media use.
- While technological approaches such as fact-checking labels are growing in use, they may not always reduce belief in fake news and can even increase its perceived accuracy.
- Psychological traits such as analytical thinking and emotional intelligence are linked to better detection of misinformation.
- A user-centred, individual differences approach can identify who is most and least susceptible to fake news and guide tailored interventions.

- Early intervention – particularly in high schools – can help inoculate young people before they fully engage with social media platforms.

Recommendations:

- Continue investing in user-centred research to identify individual traits linked to susceptibility to fake news and misinformation.
- Develop reliable tests that measure the likelihood of falling for and sharing fake news, and use these to raise awareness among low performers.
- Identify and study high performers to understand what psychological characteristics protect them from misinformation.
- Use insights from these findings to develop tailored interventions, such as emotional intelligence or critical thinking training.
- Implement such interventions in schools, ideally around the time young people gain access to social media, to foster early resistance to misinformation.

1.4 Children's Trust in the Digital Age: Fostering Digital Literacy in Adolescents

Recent evidence underscores the importance of cultivating digital literacy skills among adolescents – particularly in an era where social media has become the primary channel for accessing information. According to [Ofcom \(2023\)](#), adolescents aged 12 to 15 are now more likely to get their news from platforms like TikTok, YouTube, and Instagram than from traditional outlets such as newspapers or television. These platforms are rich with content, but also with “fake news,” conspiracy theories, and misleading claims.

While young people are often considered ‘Digital Natives’ due to their comfort with technology, this fluency does not necessarily translate into critical evaluation skills. Research suggests that adolescents continue to struggle with key aspects of digital literacy, particularly when it comes to evaluating the accuracy and reliability of online information.

This matters for two reasons. First, digital literacy helps adolescents make better-informed decisions for themselves. Second, developing critical evaluation skills at this stage of life may help reduce the societal spread of misinformation and conspiracy theories.

When it comes to detecting erroneous information online, adolescents tend to overestimate their abilities. While they often believe they are better at spotting misinformation than their parents, they frequently fail to detect both typographical (e.g., spelling mistakes) and semantic errors in text. Semantic errors—which involve violations of fact, logic, or prior knowledge – are particularly important for recognising misinformation.

Although older adolescents perform better than younger ones, both groups often base trustworthiness on superficial visual cues, such as how a website appears, rather than on content quality.

One of the most promising avenues for improving digital literacy in adolescence is by fostering **meta-cognition** – the ability to reflect on and evaluate one’s own knowledge and thinking processes. This skill has long been associated with improved reading and critical thinking in offline contexts and is equally important online.

Meta-cognitive training can help adolescents ask the right questions:

- **Planning:** What sources have I chosen to read or follow?
- **Monitoring:** Why is certain content appearing at the top of my feed?
- **Evaluating:** How am I spending my time across different platforms and sources?

By helping adolescents engage more consciously with their media environments, we can empower them to make more informed decisions and become more resilient to the spread of digital misinformation.

Key Points

- Adolescents increasingly rely on social media platforms for news, making digital literacy a crucial skill for informed decision-making and civic resilience.
- Despite being ‘Digital Natives’, adolescents often struggle with evaluating online information, particularly semantic errors that require factual and logical reasoning.

- Meta-cognitive skills such as planning, monitoring, and evaluating are key to developing critical digital literacy.
- Fostering these skills can help adolescents detect misinformation and make better decisions both online and offline.
- Supporting digital literacy in adolescence also contributes to reducing the wider societal spread of misinformation and conspiracy thinking.

Recommendations

- Introduce educational programmes that explicitly teach meta-cognitive skills as part of digital and media literacy curricula in schools.
- Encourage reflection on the types of sources adolescents encounter online and how algorithms shape what they see.
- Support age-appropriate activities that promote planning, monitoring, and evaluating online content.
- Recognise that critical evaluation skills must be actively taught, even to those fluent in digital platforms, and prioritise this in policy related to education and media regulation.
- Frame digital literacy as both a personal and societal good: improving individual decision-making and reducing the broader harms of misinformation.

1.5 Young People and Misinformation: Building Digital Resilience Through Education

Young people in the UK are growing up in an increasingly complex and high-risk digital environment. Misinformation, online harms, and cyber threats pose significant challenges to their safety, wellbeing, and their ability to engage critically with digital content. While digital literacy is widely acknowledged as a vital 21st-century skill, current provision across schools remains inconsistent. Many educators report lacking the time, confidence, or training to address misinformation and digital resilience in a sustained and meaningful way.

Too often, digital literacy and online safety are addressed through one-off events or awareness campaigns that fail to build lasting capabilities. What is needed is a systemic, embedded approach that integrates research-informed digital resilience education throughout the curriculum and equips educators to deliver it effectively.

[Project Real](#) is an example of such an approach. Co-created with young people, teachers, psychologists, influencers, and Police Scotland, the project provides an evidence-informed programme designed to build digital resilience. It includes three integrated strands:

1. **Tackling misinformation**, through structured lesson plans and media literacy activities.
2. **Addressing online safety and digital footprints**, delivered in partnership with policing professionals.
3. **Promoting cybersecurity awareness**, using comics and classroom resources developed with Education Scotland.

Early evaluations of Project Real indicate that young participants increased their confidence in evaluating online content, showed greater awareness of scams and online risks, and improved their cybersecurity knowledge. Teachers similarly reported greater confidence in delivering this material and appreciated that resources were curriculum-aligned and ready to use.

To build a digitally resilient generation, it is essential that such evidence-informed, co-created educational resources are embedded across the school curriculum. Sustained investment in both materials and teacher training will help ensure that young people – regardless of their background – are equipped to safely and critically engage with online content throughout their lives.

Key Points

- Young people are exposed to a complex digital landscape in which misinformation and online harms threaten their wellbeing and trust in information.
- Teachers often lack the resources, training, and confidence to deliver effective digital literacy and online safety education.
- One-off awareness sessions are insufficient; long-term, embedded approaches are needed to develop lasting digital resilience.
- Evidence from the Project Real programme shows that co-created, curriculum-aligned education can improve both student and teacher confidence in managing online risks.

- A whole-school, research-informed approach is necessary to ensure all young people can safely and critically navigate digital environments.

Recommendations

- Embed co-created, research-informed digital resilience education across all levels of the curriculum, ensuring it is sustained and cumulative rather than delivered as one-off events.
- Invest in teacher training and confidence-building to ensure educators are equipped to address misinformation, online safety, and cyber awareness with clarity and confidence.
- Support cross-sector partnerships (e.g. between educators, policing professionals, and digital literacy experts) to ensure real-world relevance and trust in content delivery.
- Ensure equal access to digital resilience education for all pupils, closing gaps in opportunity and protecting those most vulnerable to online harm.

Chapter 2: A Brief History of Information Disorder

2.1 History of Information Disorder: Production, Distribution, and Form

Information disorder is not a new phenomenon. Its historical roots stretch back to the nineteenth century and the emergence of sensationalist journalism. In both form and function, sensational journalism can be seen as a direct precursor to the misinformation we see today. Its subject matter, use of imagery and typographic style, and the economic drivers behind its production offer valuable insights into the continuities between past and present.

Sensationalist publications of the late nineteenth century were highly innovative, seizing on new printing technologies that enabled mass production and experimentation with bold, emotive visuals. These visuals, which often dominated front pages, were central to drawing in readers by appealing directly to emotion. This was further reinforced by new modes of circulation, with papers available through newsstands, home delivery, and via newsboys hawking on street corners.

Competitive market conditions drove prices down, and in order to remain financially viable, newspapers became increasingly dependent on advertising revenue. Thus, the economic model that sustained sensational journalism relied heavily on attention, spectacle, and emotional engagement – principles that continue to underpin information disorder in the digital age.

Today, digital technologies have accelerated and expanded these dynamics. Just as nineteenth-century printing allowed sensational visuals to proliferate, contemporary web templates enable anyone to rapidly create online content. These templates often follow minimalist, modernist design principles long associated with objectivity and trustworthiness. However, unlike historical newspaper layouts, modern templates are usually applied to content after the fact and are not created in relation to the content itself. This results in a visual aesthetic that is disconnected from truth claims, enabling post-factual websites to adopt the appearance of credibility. The visual manipulation of information, therefore, continues – but is now more subtle, capitalising on users' conditioned beliefs about what trustworthy information "should" look like.

The dominance of these web design standards across post-factual websites has given rise to a highly standardised and easily replicable aesthetic. While the content may lack factual grounding, the design lends it an air of legitimacy. Just as newsboys once drove newspaper distribution, social media now plays a key role in the dissemination of post-factual content. The presence of social media icons on these sites is a visual reminder of the new logic of circulation: content is spread not by centralised publishers, but by users themselves. This [prosumer model](#) – where consumers also act as distributors – has been identified as a key mechanism in the spread of misinformation.

While advertising remains a central financial incentive, its structure has shifted. In the digital landscape, profit is no longer generated

primarily through subscriptions, but through clicks and engagement driven by the volume of advertising content seen by users. This platform-based commercial logic incentivises sensationalism and misinformation, particularly when combined with data-driven practices such as micro-targeting and algorithmic personalisation. As [Andrejevic \(2020\)](#) argues, this creates a media environment that prioritises emotional engagement and ideological division over factual accuracy.

Recognising the historical, technological, and economic contexts of information disorder is crucial for effective policymaking. These issues are not simply linguistic but are deeply visual and infrastructural. Design histories, histories of media and journalism, political aesthetics, and political science all offer valuable frameworks for understanding the many dimensions through which misinformation evolves and circulates. In turn, these diverse perspectives can help identify where and how targeted, effective policy interventions can be made.

Key Points

- Information disorder has historical roots, emerging as early as the nineteenth century with the rise of sensationalist journalism, which used emotive imagery, eye-catching typography, and low-cost circulation models to drive profit and influence.
- Advances in technology have historically enabled the spread of misinformation; just as printing innovations boosted sensationalist newspapers in the 1800s, modern digital infrastructures now

facilitate the creation and rapid dissemination of post-factual news online.

- Today's fake news websites often adopt minimalist, modernist design templates that exploit visual cues associated with credibility and objectivity, making false content appear trustworthy.
- The economic logic underpinning misinformation has also remained consistent: both past and present models prioritise advertising revenue and audience reach over editorial accuracy.
- Social media now plays the role that newsboys once did, acting as the primary channel for distribution and amplification of misinformation, especially through 'prosumer' models where users are both consumers and sharers.
- Understanding these technological, economic, and design-driven continuities helps to reveal how information disorder operates today – and where strategic policy interventions might be made.

Recommendations:

- Recognise the structural drivers of misinformation, including digital infrastructure and advertising-based revenue models, to identify points for regulatory or systemic intervention.
- Address both visual and textual manipulation in policies focused on misinformation, acknowledging the role of web design aesthetics and platform interfaces.
- Involve a range of disciplinary experts – including design historians, media scholars, political scientists, and technologists—in shaping future responses to information disorder.

Chapter 3: Detecting and Reporting Misinformation: Considerations for Digital Ethics and Education Policies

3.1 Reasoning in Detecting Fake News and Misinformation

The ability to detect misinformation is critical – but equally important is the capacity to accept true information. Public health consequences, such as [outbreaks of vaccine-preventable diseases](#), can be driven not only by belief in misinformation (e.g. that vaccines contain harmful substances) but also by a [failure to accept scientifically valid information](#) (e.g. that vaccines are safe and effective). Understanding how individuals make veracity judgements about both true and false news enables researchers and policymakers to measure two distinct but important [factors](#): *discernment* (the ability to correctly distinguish between true and false news), and *response bias* (a general tendency to accept or reject all news as either true or false). These distinctions are crucial for designing and evaluating interventions aimed at improving public resilience to misinformation.

[Recent](#) research has shown that, on average, people are relatively good at identifying true and false news, even without specific training. This evidence challenges common assumptions that the public lacks the capacity to recognise truth. However, while performance is generally above chance, it is far from perfect – meaning that there remains a need for interventions that can help people become even more accurate in their assessments.

One particularly [relevant finding](#) is that individuals are around three times more likely to report using their intuition or gut feeling – rather than applying prior knowledge or specific reasoning strategies – when making judgements about news content. This has implications for the design of interventions. Rather than focusing exclusively on rational-analytic models of decision-making, there may be value in strengthening intuitive judgements through repeated exposure and structured practice.

To explore this, [researchers](#) tested the effectiveness of *inductive learning* – a process by which individuals improve their ability to classify content (e.g. distinguishing between true and false news) through repeated exposure to examples. Findings indicate that this form of practice can significantly improve veracity discernment. Such results point to the potential of low-cost, scalable training methods that build on basic learning principles and support the public in making more accurate, confident news judgements.

Key Points

- The ability to accept true information is just as important as the ability to reject false information, especially in public health and civic contexts
- People are generally above chance at identifying true versus false news without training, but improvements are still needed
- Intuitive reasoning plays a significant role in how individuals make news veracity judgements
- Inductive learning, which involves classifying examples of true and false content, can effectively improve discernment skills

- Measuring both discernment and response bias allows for a clearer evaluation of misinformation interventions

Recommendations

- To evaluate interventions by examining their effects on both belief in true news and rejection of false news
- To apply data analysis methods that distinguish discernment from response bias, such as receiver operating characteristic (ROC) analysis
- To develop interventions that build on basic learning mechanisms, including inductive learning approaches that enhance intuitive reasoning over time

3.2 Flagging AI – Generated Images

Recent advances in artificial intelligence have dramatically increased the prevalence and realism of synthetic media, including [deepfakes](#) – AI-generated images, audio, video, and text. AI-generated imagery has drawn particular attention due to its high visual quality and increasing use in misleading or harmful contexts. For example, during the 2024 United States presidential election, AI-generated images were widely circulated on the [social media platform X](#) (formerly Twitter) to promote disinformation. In parallel, concerns have grown over the use of AI in producing [non-consensual intimate images](#), which pose serious risks to individuals’ privacy, safety, and wellbeing.

One of the most concerning developments is the creation of AI-generated synthetic faces. These images do not depict real individuals, but instead show highly photorealistic people who do not exist. [Recent research](#) has shown that these synthetic faces are often more likely to be perceived as “real” than genuine human faces and are rated as more [trustworthy](#) by observers. Their availability online and ease of access – no technical skill is needed – makes them particularly useful for malicious actors. Because the individuals depicted are fictitious, there are no direct defamation or impersonation risks, making such images appealing for use in [harassment](#), fraud, or [espionage](#).

When people are asked to distinguish between real and AI-generated faces, they often rely on [perceptual cues](#) such as facial symmetry or image quality. Interestingly, these are the same features used successfully by machine learning algorithms to differentiate real from synthetic faces. However, humans typically apply these cues incorrectly.

For example, people tend to assume that highly proportional faces are real – when in fact, this trait is more common in AI-generated faces.

To address this issue, researchers have explored behavioural interventions designed to improve people’s ability to correctly interpret these cues. In particular, a training approach combining verbal instruction with [inductive learning](#) has shown promise. Participants were introduced to key image features – such as proportionality and pixel quality – and asked to judge whether a given face was real or AI-generated, receiving feedback on each trial. This combination of structured exposure and feedback significantly improved participants’ accuracy, even when tested again 20 days later.

These findings suggest that humans can perceive the same features used by AI to identify synthetic images, but they require explicit training to apply them effectively. There is strong potential for scalable behavioural interventions that enhance public resilience to AI-generated misinformation, particularly if integrated into broader media literacy or online safety strategies.

Key Points

- AI-generated faces are often perceived as more real and trustworthy than actual human faces, contributing to the spread of disinformation and online harms
- People use identifiable features—such as image quality and facial symmetry – to judge image authenticity, but tend to apply these features incorrectly

- Machine learning models using these same features achieve high classification accuracy, suggesting humans have the perceptual tools but lack the correct interpretive strategies
- Behavioural interventions that combine verbal instruction with inductive learning can significantly improve image discrimination, even weeks after training
- Such interventions are promising tools to support public media literacy in the age of synthetic content

Recommendations

- Support the development of behavioural training interventions that improve people's ability to distinguish between real and AI-generated images
- Use inductive learning principles to enhance intuitive judgement skills, especially around image features that are commonly misinterpreted
- Incorporate synthetic media literacy into digital education and public awareness campaigns, particularly in contexts vulnerable to visual misinformation

3.3 Making People Resistant to Fake News

The challenge of tackling misinformation lies in improving public resilience without restricting free speech. One promising evidence-based approach to this problem is known as “[inoculation](#)” – or more recently, “prebunking.” Inoculation techniques are designed to build people’s cognitive resistance to misinformation by training them to recognise [common features](#) of manipulative or misleading content.

Inoculation works on the principle that misinformation tends to follow certain predictable patterns. These include logical fallacies such as cherry-picking data, scapegoating particular groups, incoherent argumentation, or falsely presenting debate where there is scientific consensus. These tactics can often be identified through critical thinking and rhetorical awareness – skills long recognised, dating back [to classical traditions of logic and reasoning](#).

Modern inoculation approaches have applied this principle through scalable, low-cost interventions, most notably in the form of short, engaging videos. Research conducted in partnership with Google and academic collaborators has demonstrated the efficacy of such methods. [Studies](#) involving millions of social media users have shown that watching brief videos that explain manipulation techniques significantly [improves viewers’ ability to recognise](#) misinformation in real-world contexts.

This growing body of evidence suggests that prebunking is a viable public education tool. While inoculation is not a silver bullet, it offers a practical and scalable way to bolster information resilience – especially when integrated with wider digital literacy initiatives. The approach is especially

promising for early intervention strategies targeting young people, first-time voters, or those newly entering digital media spaces.

Key Points

- Inoculation, or prebunking, involves teaching people to identify rhetorical strategies used in misinformation, such as cherry-picking or scapegoating
- Short educational interventions, such as online videos, have been shown to significantly improve the public's ability to recognise manipulative content
- These interventions are scalable, low-cost, and compatible with existing digital platforms and public education strategies
- Inoculation enhances critical thinking without limiting freedom of expression, making it a valuable approach for democratic societies

Recommendations

- Support the development and dissemination of inoculation-style public information campaigns using short, evidence-based videos and interactive content
- Incorporate inoculation techniques into school curricula and youth engagement programmes to build early resistance to misinformation
- Use prebunking strategies in advance of high-risk events (e.g., elections, public health campaigns) where misinformation is likely to spread

- Continue cross-sector collaboration between researchers, government, media, and platforms to refine and implement inoculation at scale

3.4 Remembering Fake News and the Role of Fact-Checking

One of the key psychological risks of fake news is its impact on memory. Research consistently shows that exposure to fabricated information can lead people to develop false memories of news stories that never happened. In experimental studies, around 30% of people reported remembering fake news headlines they had never previously encountered – ranging from 15% to 45% depending on the story’s familiarity or alignment with participants’ political beliefs. In one study, 34% of participants claimed to remember the fabricated headline: *“Hundreds of Scottish restaurants permanently shuttered by government for not following Test-and-Protect protocols”*. Despite the story being entirely fictitious, the memory felt real to a significant proportion of participants.

Not all fake news stories are equally memorable. Implausible stories – such as “the earth was discovered to be a cube shape” – rarely generate false memories. However, most fake news is designed to be plausible and emotionally resonant. Familiarity with a subject does not offer protection; in fact, it may increase confidence in one’s false memory. This is particularly pronounced in political contexts, where self-perceived expertise and political alignment can heighten susceptibility. People are significantly more likely to remember and believe fake stories that support their existing views.

This presents a serious challenge for any democratic society that values evidence-based decision-making. Efforts to warn people that a news item may be false – such as through general content warnings – do not reliably reduce false memory formation.

However, targeted interventions show promise. When readers are provided with specific corrective information after reading fake news, and are then tested on their memory for new stories, their [susceptibility to false memories decreases](#). This indicates that people can be “[dehoaxed](#)” – but it requires effortful, explicit fact-checking. Additionally, cognitive reflection, or the ability to take a “second look” and question gut reactions, appears to be a [protective factor](#). Supporting cognitive reflection may therefore be a valuable tool in reducing the long-term cognitive footprint of misinformation.

Key Points

- Exposure to fake news can lead to the formation of false memories, even when individuals are told in advance that they may be reading fabricated content
- Fake news that aligns with an individual’s political beliefs is particularly likely to be remembered and believed
- General warnings are ineffective; however, specific post-exposure corrections can reduce memory errors and increase resistance to future fake news
- Cognitive reflection – the ability to re-evaluate an initial reaction – is associated with greater accuracy in identifying false content

Recommendations

- Encourage the use of detailed corrective information, rather than vague labels, when addressing misinformation in news and media
- Integrate cognitive reflection training into digital literacy and media education programmes, promoting “second-look” thinking
- Support public education campaigns that raise awareness of how memory can be shaped by misinformation, particularly in politically sensitive contexts
- Develop interventions that include post-exposure feedback and memory testing to reduce long-term susceptibility to fake news

3.5 Visual Language in Media Literacy Programmes

Media literacy education is increasingly recognised as a critical tool for addressing the societal harms caused by mis- and disinformation. Most existing media literacy initiatives tend to focus primarily on textual or linguistic aspects of misinformation. However, in the digital information ecosystem, messages are typically delivered through a combination of text, images, design elements, and multimedia formats. As a result, visual literacy – the ability to critically analyse and interpret visual components of communication – must be integrated more centrally into media literacy programmes.

Visual literacy is not just a technical skill. It supports citizens to assess the credibility, motives, and intent of digital communications. [Scholars](#) have argued that visual literacy in the digital age is not only about interpreting graphs or images, but about recognising how visual forms can persuade, mislead, or manipulate. For example, a webpage's typographic style or use of emotionally charged images may cue trust or familiarity, even when the content itself is misleading. These design decisions can serve as rhetorical tools that shape public belief and behaviour.

There is already a growing body of research exploring how visual language can be taught in educational settings. Serafini (2011) and Spalter and Van Dam (2008) have developed frameworks for visual literacy that focus on young learners. Similarly, Pantaleo (2012) found that pupils as young as 11–12 years old could understand how typography could be used to influence meaning in multimodal texts. However, this

body of work is rarely integrated into broader digital literacy policy or programming.

There is a strong case for policy efforts to engage designers, educators, and communication specialists in the co-creation of visually literate curricula. A visual literacy approach would empower individuals to spot visual cues that may indicate manipulation – such as suspicious layouts, vague or emotional imagery, or aesthetic mimicry of official sources. Given the increasing use of AI-generated imagery and synthetic visuals online, this type of education is particularly timely.

Incorporating visual literacy into media education equips learners not only to spot misleading text but also to decode the broader visual language used to construct misinformation. This is a critical capacity in the 21st-century digital information environment.

Key Points

- Visual language plays a major role in how mis/disinformation is constructed, circulated, and interpreted in online spaces
- Current media literacy programmes focus heavily on text, overlooking the persuasive and manipulative potential of visual design elements
- Research shows that children and adolescents are capable of developing visual literacy, including understanding the role of typography and layout in shaping meaning
- Designers and design theorists have an important role to play in shaping future media literacy interventions by contributing expertise in visual rhetoric and design ethics

- There is an urgent need to include education around AI-generated imagery, aesthetic mimicry, and visual emotional manipulation within school curricula and adult learning programmes

Recommendations

- There is an opportunity for design/visual language to be incorporated into media literacy programmes by focussing on the visual aspect of media
- Designers and design theorists can help develop and advise on visual media literacy programmes that support citizens to critically evaluate communication circulating in the public sphere
- Visual media literacy programmes might incorporate education about specific aesthetics of visual rhetoric and manipulation (such as strategies concerning typographic styling, layout and emotive or vague imagery) and identifying AI-generated imagery

Chapter 4: Messaging

4.1 Messaging and Behaviour Change

In behavioural science, changing behaviour does not begin by targeting the behaviour itself, but by addressing the underlying drivers that influence it. These drivers – commonly referred to as determinants – are drawn from decades of research and formalised in theories of behaviour and behaviour change. A well-known example is the [Theory of Planned Behaviour](#), which holds that people’s intentions are the most immediate determinant of behaviour. Intentions are themselves shaped by attitudes, which depend on people’s beliefs about the consequences of a given action and how much they value those outcomes.

When designing interventions to promote behaviour change, the goal is to shift beliefs in order to influence intentions and ultimately change behaviour. [Behaviour change techniques](#) (BCTs) provide evidence-informed [tools](#) for doing this. Several BCTs rely on the provision of information. This includes communicating the health consequences of a behaviour, its emotional or social implications, and even how visible or memorable those consequences might be. Other techniques offer information about patterns that predict the behaviour or the extent to which others approve or disapprove of it.

Notably, these techniques apply equally whether the information is accurate or misleading. Behavioural science theories do not discriminate between information and misinformation in terms of their potential impact on beliefs or behaviours. While public health professionals and academics are ethically bound to share only accurate, evidence-based

content, purveyors of misinformation are under no such obligation. This imbalance became particularly evident during the COVID-19 pandemic (Caceres et al., 2022), when misinformation often outperformed reliable health advice in terms of reach and influence.

There are several reasons for this. [Persuasive communications](#) are most effective when they are personally relevant, align with people's existing beliefs, are emotionally resonant, and include clear instructions. Misinformation campaigns often craft messages that are more tailored, striking, and memorable because they are not constrained by scientific nuance or ethical considerations. This means that even with good intentions and accurate content, official public health messages may fail to connect with the public if they are not designed in ways that reflect how people actually process and respond to information.

This highlights the need for public health messaging that is not only factually correct but also psychologically informed. To compete with misinformation, such messaging must take into account the timing, delivery, relevance, and emotional salience of the message – and must be built around principles of behavioural science.

Key Points

- Behaviour change is driven by underlying beliefs and attitudes, not just information alone
- Misinformation can be more persuasive than factual information because it is unconstrained by ethical or scientific nuance and can be tailored for emotional or cognitive impact

- Behaviour change techniques (BCTs) provide structured, evidence-based ways of shifting beliefs and intentions by delivering specific types of information
- Public health messaging needs to apply behavioural science principles to remain effective in a crowded and competitive digital information environment
- Official messages may underperform if they are not personally relevant, emotionally engaging, and clearly instructive

Recommendations

- Who: Tailor public health information to be relevant to the targeted populations
- When and How: Think about when and how to provide information to minimise the time between people receiving the message and acting
- What: Provide salient consequences and combine these with direct instruction of how to cope with these health threats

4.2 Misinformation and Peoples Beliefs

Exposure to misinformation can have powerful effects on individuals' beliefs and memories, sometimes resulting in false convictions about events that never occurred. [Research](#) shows that around 30% of participants can form false beliefs or memories after encountering misleading or fabricated information. While "fake news" is a well-known source of misinformation, individuals can also be misled by trusted sources such as [family members](#), leading to false autobiographical memories. These may include believing that one experienced events that never occurred or adopting false beliefs about real-world issues, such as the idea that vaccines cause autism.

False beliefs – whether about oneself or the world – can shape behaviour. For example, someone who believes an [unproven treatment](#) is effective may be [willing](#) to pay for it or [recommend it to others](#). In more serious cases, false autobiographical beliefs have influenced food preferences and even contributed to [wrongful criminal convictions](#). Once a false belief is established, it can guide decision-making and influence actions in harmful ways.

Certain factors can increase a person's vulnerability to misinformation. Beliefs that are already aligned with one's existing worldview are more likely to take [hold](#). For example, individuals with anti-vaccine views are more likely to believe and share anti-vaccine misinformation, reinforcing their views and contributing to [echo chambers](#) online. The [Illusory Truth Effect](#) – the tendency to perceive repeated information as more truthful – can further solidify these [false beliefs](#), even if the information has been corrected.

Crucially, anyone can form a false belief as a result of exposure to misinformation. Correcting these beliefs is challenging. Even when misinformation is explicitly retracted, its influence may persist. In particular, corrections that contradict someone's worldview or identity can backfire or be dismissed altogether.

There are, however, [protective factors](#). Research indicates that individuals who rely on [intuitive thinking](#) are more likely to believe misinformation, while those with stronger [critical thinking and information literacy skills](#) are more resilient. Teaching people how beliefs and memories are formed – and how they can be manipulated – can strengthen resistance to misinformation, including in autobiographical contexts. These skills are not just important for individuals but have broader societal value in supporting evidence-based public discourse.

Key Points

- Exposure to misinformation can lead to false beliefs or memories, including false autobiographical memories and inaccurate beliefs about the world
- False beliefs can influence behaviour, including willingness to purchase unproven treatments or share misinformation online
- Pre-existing beliefs and worldviews can increase susceptibility to misinformation and contribute to the formation of echo chambers
- The repetition of misinformation, even when followed by retractions, can increase belief in its accuracy (Illusory Truth Effect)
- Anyone can form false beliefs, and efforts to correct them are often ineffective if they contradict personal worldviews

- Critical thinking and information literacy are protective skills that can help individuals resist misinformation and reduce the spread of false beliefs

Recommendations

- Teach critical thinking and information literacy skills across educational levels and age groups
- Encourage the use of these skills before sharing information online to prevent the unintentional spread of misinformation
- Educate the public about how beliefs and memories are formed – and how they can be influenced by misinformation – to build cognitive resilience

4.3 Flagging, De-Platforming and Appeals: The Moderation of Nuanced Content on Instagram and TikTok

Content moderation on social media platforms often prioritises speed, automation, and commercial interests over the lived experiences of users. This approach frequently overlooks the harm caused by abusive reporting, censorship, and platform removal – particularly for marginalised users whose posts relate to bodies, sexuality, sex work, activism, or journalism. For these users, being de-platformed can result in the loss of community, work opportunities, access to education, and significant emotional distress. Research has found that de-platforming can negatively impact mental health and wellbeing.

Content moderation refers to the practice of regulating what users can post and see online. It is a central component of platform governance, intended to enforce platform-specific community guidelines. These guidelines are implemented via a mix of human reviewers and algorithmic decision-making systems. However, automated moderation in particular has disproportionately affected marginalised groups, often focusing more on nudity or sexuality than on abusive or violent content. This reflects broader offline legislative gaps and a desire to protect platform reputation and profit.

Flagging is one of the key tools through which moderation is enforced. It allows users to report content they believe violates platform guidelines. While flagging can support platform accountability, it can also be misused – weaponised by users who seek to silence content they disagree with. As a result, nuanced or marginalised content is often disproportionately removed through de-platforming. When this occurs, creators are meant to

have the opportunity to appeal the decision, but evidence shows that the appeals process is often dysfunctional, slow, or unresponsive.

Research conducted between 2022 and 2023—including a survey of 123 de-platformed creators and qualitative research with 35 additional participants – has documented how this system affects individuals. The findings demonstrate the urgent need for reform in how flagging, de-platforming, and appeals are handled by platforms such as Instagram and TikTok.

Key Points

- Content moderation often prioritises automation and platform interests over the wellbeing of users, with negative consequences for those de-platformed
- Marginalised users – particularly those posting about bodies, sexuality, or activism – are disproportionately affected by overzealous moderation
- Flagging mechanisms can be weaponised, especially against nuanced or politically sensitive content
- Appeals systems are often inadequate, with decisions that lack transparency or avenues for meaningful redress
- De-platforming can result in emotional harm, loss of income, and exclusion from digital spaces of support and opportunity

Recommendations

- Provide clear, detailed, and specific communication to users about moderation decisions, including the rationale for removal or account suspension
- Establish dedicated teams to oversee and respond to appeals processes, ensuring transparency and fairness
- Introduce a rehabilitative approach to policy enforcement that differentiates between the severity of infractions, avoiding blanket penalties
- Empower users to curate their own content experience through granular content controls, rather than relying solely on blanket platform moderation

Chapter 5: Fake News Algorithms

5.1 Fake News Warning Labels

Fake news has significant real-world consequences. For example, the 2024 riots in Southport were traced back to misinformation originating from a Pakistan-based website, Channel3Now. While the creators of fake news may not always intend harm, the effects on public sentiment toward institutions, communities, religions, and society at large are undeniable. The rise of the internet and social media has accelerated the instantaneous spread of false narratives, amplifying this harm.

Current solutions typically place responsibility on social media companies to warn users when links they share might contain misinformation. However, this approach has limitations. Social media platforms have the authority to change how warnings are displayed, which can leave users vulnerable to misinformation. Furthermore, links without warning labels may be mistakenly regarded as trustworthy, despite not having undergone any fact-checking. Additionally, these interventions do not inspect the original news source directly. Users often navigate to websites posing as legitimate news outlets without any external evaluation or overview.

An independent fake news warning system could address these gaps by being applied across all news articles regardless of where they appear online. This system would function similarly to a nutritional label, providing concise information on the content, including facts, opinions, truthfulness, persuasion techniques, and the presence of AI-generated material. Research indicates that warning labels can be effective in reducing the spread of fake news. One promising approach is to develop

a method that identifies check-worthy claims within an article and presents them clearly to readers. Fake news articles often use persuasive language designed to encourage belief and sharing. By presenting facts transparently, readers can better judge whether further research is warranted. Where applicable, independent fact-checked information could also be integrated within the warning label.

Key Points

- Fake news can cause tangible harm to public order and social cohesion.
- Current social media-based warning systems are inconsistent and insufficient to fully protect users.
- Lack of source-level assessment means users can be misled by seemingly legitimate news sites.
- An independent warning label system could provide consistent, transparent information about news veracity.
- Such labels can help readers critically assess persuasive or false claims, potentially reducing misinformation spread.

Recommendations

- Develop and implement an independent, platform-agnostic fake news warning label system accessible across all news sources.
- Design warning labels to include clear, factual, opinion, and AI-generation indicators to aid reader understanding.
- Incorporate check-worthy claim detection to highlight specific statements for reader scrutiny.

- Include links or references to independent fact-checking where available to support informed judgement.
- Promote public awareness of such labels as a tool to empower critical engagement with news content.

5.2 Content Provenance in the Fight Against Misinformation

Misinformation often spreads not only through manipulated media but also through authentic content that is misattributed or taken out of context to support false narratives. While various tools exist to detect AI-generated or manipulated media, their effectiveness is limited because much misinformation involves genuine material presented misleadingly. At the same time, many legitimate news stories use content that has been edited, frequently with the assistance of AI technologies.

Given the rapid evolution of generative AI, detection technologies face constant challenges and risks becoming quickly outdated. In contrast, content provenance technologies provide a more robust solution by tracing the origin and edit history of digital content. This offers users contextual information, empowering them to make better-informed decisions about the trustworthiness of what they see.

An important open standard supporting this approach is the Coalition for Content Provenance and Authenticity (C2PA), which enables digital media to be cryptographically signed with tamper-evident metadata. This metadata records how content was created, edited, and by whom, providing a verifiable trail of authenticity.

Provenance functions best as an opt-in system. Authentic creators — such as journalists, content creators, and public officials — can attach provenance information to their content as a trust signal. Although bad actors are unlikely to participate, widespread adoption among trusted sources creates a credibility layer that benefits the entire information ecosystem.

For instance, a news organisation or public figure can sign their content using C2PA metadata. If that content is later altered or misused to spread misinformation, the original signed version provides cryptographic proof of authenticity, shifting the burden of proof. Rather than attempting to prove a claim false, trusted actors can demonstrate the genuine origin of their material.

The [durability of provenance](#) information relies on ecosystem support. While metadata signed with C2PA is tamper-evident and cryptographically verifiable, many platforms currently strip metadata during upload or compression, weakening trust chains.

To address this, provenance should be reinforced through redundancy. Combining cryptographically signed metadata with complementary techniques – such as watermarking and fingerprinting – helps ensure provenance signals persist even if metadata is removed by platforms or through user actions like screenshots or printing.

Key Points

- Misinformation often involves authentic content that is misattributed or presented out of context rather than overtly manipulated media.
- Detection tools for AI-generated content face ongoing challenges due to rapid technological advancement.
- Content provenance provides contextual trust signals by tracing content origin and editing history.
- The C2PA open standard enables tamper-evident, cryptographically signed provenance metadata.

- Provenance is most effective as an opt-in trust signal used by credible actors, creating a layered system of authenticity.
- Provenance metadata alone is vulnerable if platforms strip metadata; combining it with watermarking and fingerprinting improves durability.

Recommendations

- Encourage individuals and organisations to opt-in to attaching provenance information to their digital content to signal authenticity.
- Urge news and social media platforms to adopt open standards like C2PA and preserve provenance metadata instead of stripping it.
- Promote combining cryptographically signed provenance metadata with watermarking and fingerprinting to ensure signal durability.
- Support initiatives that educate users on the role and benefits of content provenance in assessing information trustworthiness.

Acknowledgements

This policy paper was developed with the contributions of leading researchers across multiple institutions, drawing on interdisciplinary expertise in psychology, behavioural science, communication, media literacy, and public policy.

We gratefully acknowledge the authors of each chapter and section:

- **Dr Alice Howarth**, *University of Liverpool* (1.1)
- **Dr Natalia Pasternak**, *Columbia University* (1.2)
- **Dr David Robertson**, *University of Strathclyde* (1.3)
- **Pip Brown and Dr Michaela Gummerum**, *University of Warwick* (1.4)
- **Dr Yvonne Skipper**, *University of Glasgow* (1.5)
- **Anna Talley**, *University of Edinburgh* (2.1, 3.5)
- **Dr Ariana Modirrousta-Galian, Dr Tina Seabrooke and Professor Philip Higham**, *University College London; University of Southampton* (3.1)
- **Mansi Pattni, Dr Tina Seabrooke and Professor Philip Higham**, *University of Southampton* (3.2)
- **Professor Stephan Lewandowsky**, *University of Bristol* (3.3)
- **Dr Faye Skelton, Dr Marina Wimmer and Theodore Carlson Webster**, *Edinburgh Napier University* (3.4)

- **Dr Chantal den Daas and Professor Marie Johnston**, *University of Aberdeen* (4.1)
- **Dr Charlotte Bücken, Dr Paul Riesthuis and Professor Henry Otgaar**, *KU Leuven* (4.2)
- **Dr Carolina Are**, *Northumbria University* (4.3)
- **Dr Md Zia Ullah and Katarina Alexander**, *Edinburgh Napier University* (5.1)
- **Professor John Collomosse**, *University of Surrey* (5.2)
- **Emily Spencer**, *Edinburgh Napier University*
- **Jeremy Wright**, *Toronto Metropolitan University*
- **Michael Marshall**, *Good Thinking Society*

References

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179-211. [https://doi:10.1016/0749-5978\(91\)90020-T](https://doi:10.1016/0749-5978(91)90020-T)
- Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social Media + Society*, 9, <https://doi:20563051221150412>.
- Andrejevic, M. (2020) 'Political Function of Fake News: Disorganized Propaganda in the Era of Automated Media', in M. Zimdars and K. McLeod (eds) *Fake news: Understanding media and misinformation in the digital age*. Cambridge, Massachusetts: The MIT Press (Information policy), pp. 19–28.
- Anson, I. G. (2018). Partisanship, political knowledge, and the Dunning-Kruger effect. *Political Psychology*, 39, 1173-1192. <https://doi.org/10.1111/pops.12490>
- Are, C. (2023). The assemblages of flagging and de-platforming against marginalised content creators. *Convergence*, 30, 922-937. <https://doi.org/10.1177/13548565231218629>
- Are, C. (2024). 'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation. *Information, Communication & Society*, 1–18. <https://doi.org/10.1080/1369118X.2024.2396621>
- Are, C. (2024). Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *New Media & Society*, 0(0). <https://doi.org/10.1177/14614448241228544>

- Are, C., & Briggs, P. (2023). The Emotional and Financial Impact of De-platforming on Creators at the Margins. *Social Media + Society*, 9(1). <https://doi.org/10.1177/20563051231155103>
- Are, C., Collingham, H., Carrothers, A.M. and Fox, E. (2023). Co-designing platform governance policies: Tackling malicious flagging and de-platforming with impacted social media users. *Centre for Digital Citizens*, https://digitalcitizens.uk/blog/platform_governance_inequalities/.
- Bernstein, D. M., & Loftus, E. F. (2009). The consequences of false memories for food preferences and choices. *Perspectives on Psychological Science*, 4(2), 135-139. <https://doi.org/10.1111/j.1745-6924.2009.01113.x>
- Collomosse, J., & Parsons, A. (2024). To Authenticity, and Beyond! Building safe and fair generative AI upon the three pillars of provenance. *IEEE Computer Graphics and Applications*, 82-90. DOI: [10.1109/MCG.2024.3380168](https://doi.org/10.1109/MCG.2024.3380168)
- Rottweiler, B., & Gill, P. (2022). Conspiracy Beliefs and Violent Extremist Intentions: The Contingent Effects of Self-efficacy, Self-control and Law-related Morality. *Terrorism and Political Violence*, 34, 1485-1504, <https://doi.org/10.1080/09546553.2020.1803288>
- Booth, E., Lee, J., Rizoiu, M.-A., & Farid, H. (2024). Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. *Journal of Sociology*, 60, 440-457. <https://doi.org/10.1177/14407833241231756>
- Borinskaya, S. A., Ermolaev, A. I., & Kolchinsky, E. I. (2019). Lysenkoism against genetics: The meeting of the Lenin All-Union Academy of Agricultural Sciences of August 1948, its background, causes, and aftermath. *Genetics*, 212, 1-12. <https://doi.org/10.1534/genetics.118.301413>

- Caceres, M. M. F., Sosa, J. P., Lawrence, J. A., Sestacovschi, C., Tidd-Johnson, A., Rasool, M. H. U., . . . Cuevas-Lou, C. (2022). The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*, 9, 262. doi: [10.3934/publichealth.2022018](https://doi.org/10.3934/publichealth.2022018)
- Calvillo, D. P., Harris, J. D., & Hawkins, W. C. (2023). Partisan bias in false memories for misinformation about the 2021 US Capitol riot. *Memory*, 31, 137-146. <https://doi.org/10.1080/09658211.2022.2127771>
- Carnegie UK Trust. (2021). *National Digital Ethics Public Panel: Insight report*. Carnegie UK Trust.
- Chigwedere, P., Seage III, G. R., Gruskin, S., Lee, T. H., & Essex, M. (2008). Estimating the lost benefits of antiretroviral drug use in South Africa. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 49, 410-415. DOI: 10.1097/QAI.0b013e31818a6cd5
- Davidson, B. M., & Kobayashi, T. (2022). The effect of message modality on memory for political disinformation: Lessons from the 2021 US capitol riots. *Computers in Human Behavior*, 132, 107241. <https://doi.org/10.1016/j.chb.2022.107241>
- Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., ... & Quattrociocchi, W. (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences*, 113, 554-559. <https://doi.org/10.1073/pnas.1517441113>
- Digital Ethics Expert Group. (2022). *Building trust in the digital era: Achieving Scotland's aspirations as an ethical digital nation*. Scottish Government.

- Douglas, K. M., Sutton, R. M., & Cichocka, A. (2017). The psychology of conspiracy theories. *Current Directions in Psychological Science*, 26, 538-542. <https://doi.org/10.1177/0963721417718261>
- Dubé, E., Laberge, C., Guay, M., Bramadat, P., Roy, R., & Bettinger, J. A. (2013). Vaccine hesitancy: An overview. *Human Vaccines & Immunotherapeutics*, 9, 1763-1773. <https://doi.org/10.4161/hv.24657>
- Eldredge, L. K. B., Markham, C. M., Ruitter, R. A., Fernández, M. E., Kok, G., & Parcel, G. S. (2016). *Planning health promotion programs: an intervention mapping approach*: John Wiley & Sons.
- Fishbein, M., & Ajzen, I. (2011). *Predicting and changing behavior: The reasoned action approach*: Psychology Press.
<https://doi.org/10.4324/9780203838020>
- Gershberg, Z. and Illing, S.D. (2022) *The paradox of democracy: free speech, open media, and perilous persuasion*. Chicago: University of Chicago Press.
- Greene, C. M., & Murphy, G. (2023). Debriefing works: Successful retraction of misinformation following a fake news study. *PloS one*, 18(1), e0280295. <https://doi.org/10.1371/journal.pone.0280295>
- Greene, C. M., Nash, R. A., & Murphy, G. (2021). Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory*, 29, 587-604.
<https://doi.org/10.1080/09658211.2021.1923754>
- Greene, C. M., Ryan, K. M., Ballantyne, L., Barrett, E., Cowman, C. S., Dawson, C. A., ... & Murphy, G. (2024). Unringing the bell: Successful debriefing following a rich false memory study. *Memory & Cognition*, 52, 1079-1092. doi: [10.3758/s13421-024-01524-9](https://doi.org/10.3758/s13421-024-01524-9)

- International Council of Education Advisers. (2023). *Third formal report 2021–2023*. Scottish Government. <https://www.gov.scot/>
- James, M. (2020). The Role of Ethics Online and Among Social Media Designers. In L. Scherling & A. DeRosa (Ed.). *Ethics in Design and Communication: Critical Perspectives* (pp. 152–162). London: Bloomsbury Academic. Retrieved July 15, 2024, from <http://dx.doi.org/10.5040/9781350077027.0029>
- Johnston, M., Carey, R. N., Connell Bohlen, L. E., Johnston, D. W., Rothman, A. J., de Bruin, M., . . . Michie, S. (2020). Development of an online tool for linking behavior change techniques and mechanisms of action based on triangulation of findings from literature synthesis and expert consensus. *Translational Behavioral Medicine*. <https://doi.org/10.1093/tbm/ibaa050>
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the “enemy of induction”? *Psychological Science*, *19*, 585-592. <https://doi.org/10.1111/j.1467-9280.2008.02127.x>
- Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing—therefore,(climate) science is a hoax: An anatomy of the motivated rejection of science. *Psychological Science*, *24*, 622-633. <https://doi.org/10.1177/0956797612457686>
- 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*, *13*, 106-131. <https://doi.org/10.1177/1529100612451018>
- MacFarlane, D., Tay, L. Q., Hurlstone, M. J., & Ecker, U. K. (2021). Refuting spurious COVID-19 treatment claims reduces demand and

misinformation sharing. *Journal of Applied Research in Memory and Cognition*, 10, 248-258. <https://doi.org/10.1016/j.jarmac.2020.12.005>

Machete, P., & Turpin, M. (2020). The use of critical thinking to identify fake news: A systematic literature review. In *Responsible Design, Implementation and Use of Information and Communication Technology: 19th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2020, Skukuza, South Africa, April 6–8, 2020, Proceedings, Part II 19* (pp. 235-246). Springer International Publishing. https://doi.org/10.1007/978-3-030-45002-1_20

Marchese, C. (2021) *Information design for the common good: human-centric approaches to contemporary design challenges*. London ; New York: Bloomsbury Visual Arts.

Marques, M. M., Wright, A. J., Corker, E., Johnston, M., West, R., Hastings, J., . . . Michie, S. (2023). The behaviour change technique ontology: transforming the behaviour change technique taxonomy v1. *Wellcome Open Research*, 8.

Michael, R. B., & Breaux, B. O. (2021). The relationship between political affiliation and beliefs about sources of “fake news”. *Cognitive Research: Principles and Implications*, 6, 1-15. <https://doi.org/10.1186/s41235-021-00278-1>

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., . . . Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, 46(1), 81-95. <https://doi.org/10.1007/s12160-013-9486-6>

- Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General*, *152*, 2411. <http://dx.doi.org/10.31234/osf.io/4bgkd>
- Modirrousta-Galian, A., Higham, P. A., & Seabrooke, T. (2024). Wordless wisdom: The dominant role of tacit knowledge in true and fake news discrimination. *Journal of Applied Research in Memory and Cognition*, *14*, 231-240. <https://psycnet.apa.org/doi/10.1037/mac0000151>
- Modirrousta-Galian, A., Seabrooke, T., Hanoach, Y., Kelley, N. J., & Higham, P. A. (2024). An Inductive Learning Intervention to Improve News Veracity Discernment. https://doi.org/10.31234/osf.io/6j7fr_v1
- Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological Science*, *30*, 1449-1459. <https://doi.org/10.1177/0956797619864887>
- Musi, E., & Reed, C. (2022). From fallacies to semi-fake news: Improving the identification of misinformation triggers across digital media. *Discourse & Society*, *33*, 349-370. <https://doi.org/10.1177/09579265221076609>
- Newman M. (2018). Is cancer fundraising fuelling quackery? *BMJ*, *362*. <https://doi.org/10.1136/bmj.k3829>
- Nightingale, S., & Farid, H. (2022). Synthetic faces are more trustworthy than real faces. *Journal of Vision*, *22*, 3068-3068. <https://doi.org/10.1167/jov.22.14.3068>

- Oeberst, A., Wachendörfer, M. M., Imhoff, R., & Blank, H. (2021). Rich false memories of autobiographical events can be reversed. *Proceedings of the National Academy of Sciences*, 118, e2026447118. <https://doi.org/10.1073/pnas.2026447118>
- OECD. (2020). *Transparency, communication and trust: The role of public communication in responding to the wave of disinformation about the new coronavirus*. OECD Policy Responses to Coronavirus (COVID-19). <https://www.oecd.org/>
- Pantaleo, S. (2012) 'Middle years students thinking with and about typography in multimodal texts', *Literacy Learning: The Middle Years*, 20, p. 37+. <https://search.informit.org/doi/10.3316/informit.902017433567687>
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66, 4944-4957. <https://doi.org/10.1287/mnsc.2019.3478>
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147, 1865. <https://psycnet.apa.org/doi/10.1037/xge0000465>
- Pfänder, J., & Altay, S. (2025). Spotting false news and doubting true news: a systematic review and meta-analysis of news judgements. *Nature Human Behaviour*, 1-12. <https://doi.org/10.1038/s41562-024-02086-1>
- Preston, S., Anderson, A., Robertson, D. J., Shephard, M. P., & Huhe, N. (2021). Detecting fake news on Facebook: The role of emotional intelligence. *Plos one*, 16(3), e0246757. <https://doi.org/10.1371/journal.pone.0246757>

- Principe, G. F., & Schindewolf, E. (2012). Natural conversations as a source of false memories in children: Implications for the testimony of young witnesses. *Developmental Review, 32*(3), 205-223.
<https://doi.org/10.1016/j.dr.2012.06.003>
- Public Interest Journalism Working Group. (2022). *Scottish Government's response to recommendations of the Public Interest Journalism Working Group*. Scottish Government.
- Richards, A. 2021: <https://conspiracychart.com/>
- Roozenbeek, J., & Van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*, 1-10. <https://doi.org/10.1057/s41599-019-0279-9>
- Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L., Recchia, G., ... & Van Der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science, 7*, 201199. <https://doi.org/10.1098/rsos.201199>
- Roozenbeek, J., Van Der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances, 8*, eabo6254. <https://doi.org/10.1126/sciadv.abo6254>
- Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences, 116*, 7662-7669. <https://doi.org/10.1073/pnas.1805871115>
- Scoboria, A., Wade, K. A., Lindsay, D. S., Azad, T., Strange, D., Ost, J., & Hyman, I. E. (2017). A mega-analysis of memory reports from eight peer-reviewed

false memory implantation studies. *Memory*, 25, 146-163.
<https://doi.org/10.1080/09658211.2016.1260747>

Scoboria, A., Wysman, L., & Otgaar, H. (2012). Credible suggestions affect false autobiographical beliefs. *Memory*, 20, 429-442.
<https://doi.org/10.1080/09658211.2012.677449>

Serafini, F. (2011) 'Expanding Perspectives for Comprehending Visual Images in Multimodal Texts', *Journal of Adolescent & Adult Literacy*, 54, pp. 342–350. Available at: <https://doi.org/10.1598/JAAL.54.5.4>.

Shephard, M. P., Robertson, D. J., Huhe, N., & Anderson, A. (2023). Everyday non-partisan fake news: Sharing behavior, platform specificity, and detection. *Frontiers in Psychology*, 14, 1118407.
<https://doi.org/10.3389/fpsyg.2023.1118407>

Silverman, K.N. and Piedmont, J. (2016) 'A Visual Literacy Curriculum for Today', *Knowledge quest, American Library Association*, 44, pp. 32–37.

Spalter, A.M. and Van Dam, A. (2008) 'Digital Visual Literacy', *Theory Into Practice*, 47, 93–101. <https://doi.org/10.1080/00405840801992256>.

Swami, V., Coles, R., Stieger, S., Pietschnig, J., Furnham, A., Rehim, S., & Voracek, M. (2011). Conspiracist ideation in Britain and Austria: Evidence of a monological belief system and associations between individual psychological differences and real-world and fictitious conspiracy theories. *British Journal of Psychology*, 102, 443-463.
<https://doi.org/10.1111/j.2044-8295.2010.02004.x>

Szpitalak, M., Woltmann, A., Polczyk, R., & Kękuś, M. (2021). Memory training as a method for reducing the misinformation effect. *Current Psychology*, 40, 5410-5419. <https://doi.org/10.1007/s12144-019-00490-9>

- Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13, e0203958.
<https://doi.org/10.1371/journal.pone.0203958>
- Trevors, G., & Duffy, M. C. (2020). Correcting COVID-19 misconceptions requires caution. *Educational Researcher*, 49, 538-542.
<https://doi.org/10.3102/0013189X20953825>
- Varet, F., Fournier, V., & Delouvée, S. (2025). Assessing the role of conspiracy beliefs in oncological treatment decisions: An experimental approach. *Applied Psychology: Health and Well-Being*, 17, e12615.
<https://doi.org/10.1111/aphw.12615>