The online information environment

Understanding how the internet shapes people's engagement with scientific information

THE
ROYAL
SOCIETY

The online information environment: Understanding how the internet shapes people's engagement with scientific information
Issued: January 2022 DES7656
ISBN: 978-1-78252-567-7

This report and other project outputs can be viewed at:
royalsociety.org/online-information-environment

Cover image AI generated image using Generative Adversarial Network (GAN) © iStock / Dorin Ionescu.

Contents

Foreword

Professor Frank Kelly CBE FRS.

It is arguably one of the most abundant resources of our age – never before has so much information been available to so many people.

Wherever access to the internet is possible, individuals can access entire libraries-worth of knowledge, decades of news reports, vaults-full of documents and records, speeches, images and videos; and, in the current pandemic, the genome sequence of a novel coronavirus and a torrent of research preprints released before peer review. Once it would take days for news to pass from town to town, but the last century saw a speeding up of information transmission, from the early growth of telephony through to the advent of the World Wide Web in the 1990s and the popularity of social media from the early 2010s. Digital technology has transformed our ability to be informed and to inform others.

But it is not just high-quality information that is being shared.

Inaccurate, misleading and completely false information is shared online in large volumes – both unintentionally by some and maliciously by others. Fictional stories end up being passed around as truth, conspiracies gain weight as they pass through the rumour mill and science becomes mangled beyond recognition.

Misinformation and fake news is not new (see the quote from George Eliot's *Middlemarch*). Alongside this report, we are publishing two literature reviews looking at the spread of misinformation about water fluoridation and vaccination in the 20th century, well before the emergence of the modern information environment. What online technologies have changed, however, is the scale and speed of spread.

The Royal Society's mission since it was established in 1660 has been to promote science for the benefit of humanity, and a major strand of that is to communicate accurately. But false information is interfering with that goal. It is accused of fuelling mistrust in vaccines, confusing discussions about tackling the climate crisis and influencing the debate about genetically modified crops.

Science stands on the edge of error. It is a process of dealing with uncertainties, prodding and testing received wisdom. Science challenges us to continually assess and revise our understanding of the world. What we believed 100 years ago has been replaced with new knowledge. Some people think science is absolute and when it corrects itself it is somehow not to be trusted or believed. We must work to help people recognise that the core ability to correct itself is a strength, not a weakness, of the scientific method. This ability requires the prioritisation of the best data and most trustworthy information, as well as a safe and healthy online information environment which allows robust and open scientific debate. Balancing these necessities is one of the key aims of this report.

Of course, this report can only consider part of a problem as broad and complicated as how to improve the quality of the information environment. Misinformation problems are in part irreducibly political and social in nature. In free and diverse societies we will always have some version of them. In this report we have focused on the part of this where the Royal Society, as the national academy of science, speaks with the greatest authority: on issues pertaining to how science is communicated online, and the technologies underpinning that.

Fact-checking is especially important, and this is an area where the scientific community can help. National academies and learned societies can react to new misinformation threats by quickly providing accurate summaries of what we know. To do this, better access to data is needed for researchers to identify topics of misinformation early in the process of amplification.

This, in itself, will not be enough to counteract the algorithmic amplification of polarising misinformation in an attention economy which incentivises the spread of sensational stories rather than sound understanding. Ultimately, we will need to see legislation which can address the incentives of business models that shape the algorithms determining the spread of content. Scientists will need to work with lawyers and economists to make sure that the particular sensitivities of scientific misinformation are considered when legislation is framed.

The scientific community has its own issues to address in this regard. The incentives for scientific publication and communication need careful consideration, to ensure that novelty isn't overstated simply to grab attention. Open access has been a boon, but in an age of information overload we need tools to identify questionable publishers or platforms. Furthermore, scientists need to be clear and transparent about their own motivations and whether they are seeking to inform or seeking to persuade.

This report represents continuing development of, rather than a final chapter in, the Royal Society's consideration of these issues. Further work going into more detail on some areas covered is planned. An example of this is the Society's ambitious new programme, *Reimagining science*, which seeks to improve the narratives of science in society. Future work will also examine the role of digital technologies, and data access, in scientific research.

Professor Frank Kelly CBE FRS

"But oppositions have the illimitable range of objections at command, which need never stop short at the boundary of knowledge, but can draw forever on the vasts of ignorance. What the opposition in Middlemarch said about the New Hospital and its administration had certainly a great deal of echo in it, for heaven has taken care that everybody shall not be an originator..."

George Eliot,
***Middlemarch* (1872)**

Executive summary

The internet has transformed the way people consume, produce, and disseminate information about the world. In the online information environment, internet users can tailor unlimited content to their own needs and desires. This shift away from limited, gatekept, and pre-scheduled content has democratised access to knowledge and driven societal progress. The COVID-19 pandemic exemplifies this, with global researchers collaborating virtually across borders to mitigate the harms of the disease and vaccinate populations.

The unlimited volume of content, however, means that capturing attention in the online information environment is difficult and highly competitive. This heightened competition for attention presents a challenge for those who wish to communicate trustworthy information to help guide important decisions. The poor navigation or, even, active exploitation of this environment by prominent public figures and political leaders has, on many occasions, led to detrimental advice being disseminated amongst the public. This challenge has caused significant concern with online 'misinformation' content being widely discussed as a factor which impacts democratic elections and incites violence. In recent years, misinformation has also been identified as a challenge in relation to a range of scientific topics, including vaccine safety, climate change, and the rollout of 5G technology.

The Royal Society's mission is to promote excellence in science and support its use for the benefit of humanity. The consumption and production of online scientific information is, therefore, of great interest. This report, *The online information environment*, provides an overview of how the internet has changed, and continues to change, the way society engages with scientific information, and how it may be affecting people's decision-making behaviour – from taking up vaccines to responding to evidence on climate change. It highlights key challenges for creating a healthy online information environment and makes a series of recommendations for policymakers, academics, and online platforms.

These recommendations, when taken together, are intended to help build collective resilience to harmful misinformation content and ensure access to high quality information on both public and private forums.

The report has been guided by a working group of leading experts in this field and informed by a series of activities commissioned by the Royal Society. Firstly, literature reviews were commissioned on historical examples of scientific misinformation; the evidence surrounding echo chambers, filter bubbles, and polarisation; and the effects of information on individuals and groups. Secondly, the Society hosted various workshops and roundtables with prominent academics, fact-checking organisations, and online platforms. Finally, two surveys were commissioned – the first on people's attitudes and behaviours towards online scientific misinformation and the second on people's ability to detect deepfake video content.

The chapters of the report are focused on understanding and explaining essential aspects of the online information environment. They explore a broad range of topics including the ways our minds process information and how this is impacted by accessing information online; how information is generated in a digital context and the role of incentives for content production; and types of synthetic online content and their potential uses, both benign and malicious. However, there are important areas that are not covered in this report, outlined in box 1, which are part of the wider questions around trust in science, in the internet and in institutions. These include the role of traditional science communicators and the wider research community in enabling access to trustworthy information; the issue of online anonymity; and the impact that the online information environment can have on democracy and political events (eg elections).

Within this report, 'scientific misinformation' is defined as information which is presented as factually true but directly counters, or is refuted by, established scientific consensus. This usage includes concepts such as 'disinformation' which relates to the deliberate sharing of misinformation content.

Key findings

- Although misinformation content is prevalent online, the extent of its impact is questionable[1]. For example, the Society's survey of members of the British public[2] found that the vast majority of respondents believe the COVID-19 vaccines are safe, that human activity is responsible for climate change, and that 5G technology is not harmful. The majority believe the internet has improved the public's understanding of science, report that they are likely to fact-check suspicious scientific claims they read online and state that they feel confident to challenge their friends and family on scientific misinformation.

- The existence of echo chambers (where people encounter information that reinforces their own beliefs, online and offline) is less widespread than may be commonly assumed and there is little evidence to support the filter bubble hypothesis (where algorithms cause people to only encounter information that reinforces their own beliefs)[3, 4].

[1] Cabrera Lalinde I. 2021 How misinformation affected the perception of vaccines in the 20th century based on the examples of polio, pertussis and MMR vaccines. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[2] 85% consider the Pfizer-BioNTech vaccine to be safe. AstraZeneca = 80%, Moderna = 74%. 4-5% believe the vaccines are 'not at all safe'. Royal Society / YouGov, July 2021.

[3] Arguedas A, Robertson C, Fletcher R, Nielsen R. 2021 Echo chambers, filter bubbles, and polarisation. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[4] Röttger P, Vedres B. The Information Environment and its Effects on Individuals and Groups. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

- Uncertainty is a core aspect of scientific method, but significant dispute amongst experts can spill over to the wider public[5]. This can be particularly challenging when this uncertainty is prolonged, and the topic has no clear authority. This gap between uncertainty and certainty creates information 'deserts' online with platforms being unable to clearly guide users to trustworthy sources[6]. For example, during the COVID-19 pandemic, organisations such as the World Health Organization and the National Health Service were able to act as authoritative voices online. However, with topics such as 5G telecommunications, it has been more difficult for platforms to quickly identify trustworthy sources of evidence and advice.

- The concept of a single 'anti-vax' movement is misleading and does not represent the range of different reasons for why some people are reluctant to be vaccinated[7]. Those with anti-vaccination sentiments can have distinct concerns including child safety, or act not out of scepticism about the evidence, but out of distrust of governments. In addition, there are various actors involved in creating and spreading anti-vaccination material. These include political actors, particularly when a relevant event (eg a pandemic) is dominating the news cycle[8, 9].

- Technology can play an important though limited role in addressing misinformation content online. In particular, it can be useful in areas such as rapid detection of harmful misinformation content. Provenance enhancing technology, which provides information on the origins of online content and how it may have been altered, shows promise and will become increasingly important as misinformation content grows more sophisticated. Even now, expertly manipulated content appears to be difficult to detect. Survey experiments conducted for this report indicates that most people struggle to identify deepfake video content even when prompted[11].

- Incentives for content production and consumption are the most significant factor to consider when evaluating the online information environment. These incentives can occur on a macro and micro level (affecting both platforms and individual users) and have been described in this report as content which exists for public benefit (eg helping others) or private benefit (eg generating financial profit).

  Understanding how to mitigate the role of these incentives in the spread of misinformation content requires further consideration on the economic and legal aspects of the online information environment.

[5] Royal Society roundtable on Telecommunications and Misinformation, November 2020.

[6] Royal Society roundtable with Major Technology Organisations, March 2021.

[7] Royal Society workshop on Vaccines and Misinformation, July 2020.

[8] Royal Society workshop on Horizon Scanning and Scientific Misinformation, March 2021.

[9] Arguedas A, Robertson C, Fletcher R, Nielsen R. 2021 Echo chambers, filter bubbles, and polarisation. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[10] Royal Society roundtable with Major Technology Organisations, March 2021.

[11] Lewis A, Vu P, Duch R. 2021 Deepfake detection and content warnings: Evidence from two experiments. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

Recommendations

AREA FOR ACTION: PROTECTING ONLINE SAFETY

RECOMMENDATION 1

As part of its online harms strategy, the UK Government must combat misinformation which risk societal harm as well as personalised harm, especially when it comes to a healthy environment for scientific communication.

When considering the potential damage caused by unchecked scientific misinformation online, the framing of 'harm', adopted by the UK Government, has focused primarily on harm caused to individuals rather than society as a whole[12]. For example, this limitation risks excluding misinformation about climate change. While the commissioned YouGov survey suggests that levels of climate change denialism in the UK are very low[13], there is evidence to suggest that misinformation encouraging climate 'inactivism' is on the rise[14, 15].

The consequences of societally harmful misinformation, including its influence on decision-makers and public support for necessary policy changes, could feasibly contribute to physical or psychological harm to individuals in future (eg through failure to mitigate climate catastrophe).

This view is complemented by our YouGov survey which suggests that the public are more likely to consider misinformation about climate change to be harmful[16] than misinformation about 5G technology (a subject which has been significantly cited within discussions on online harms[17, 18, 19]).

There needs to be a recognition that misinformation which affects group societal interests can cause individual harm, especially to infants and future generations who do not have a voice[20]. We recommend that the impact of societal harms on current and future generations, such as misinformation about climate change, is given serious consideration within the UK Government's strategy to combat online harms.

[12] HM Government. 2020 Online Harms White Paper. See: https://www.gov.uk/government/consultations/online-harms-white-paper (accessed 4 November 2021).

[13] 5% do not believe human activity is responsible for climate change. Royal Society / YouGov, July 2021.

[14] Coan T, Boussalis C, Cook J, Nanko M. 2021 Computer-assisted detections and classification of misinformation about climate change. SocArXiv (doi:10.31235/osf.io/crxfm)

[15] Avaaz. 2021 Facebook's Climate of Deception: How Viral Misinformation Fuels the Climate Emergency. See https://secure.avaaz.org/campaign/en/facebook_climate_misinformation/ (accessed 4 November 2021)

[16] 83% consider misinformation about climate change to be harmful, 67% consider misinformation about 5G technology to be harmful. Royal Society / YouGov, July 2021.

[17] HM Government. 2021 Minister launches new strategy to fight online disinformation. See https://www.gov.uk/government/news/minister-launches-new-strategy-to-fight-online-disinformation (accessed 4 November 2021)

[18] HM Government. 2020 Online Harms White Paper. See: https://www.gov.uk/government/consultations/online-harms-white-paper (accessed 4 November 2021).

[19] Hansard. Debate on Online Harms. See https://hansard.parliament.uk/commons/2020-11-19/debates/29AA4774-FDE3-4AB9-BBAB-F072DE3E8074/OnlineHarms (accessed 4 November 2021).

[20] Robinson K. 2020 The Ministry for the Future. London, UK: Orbit Books.

RECOMMENDATION 2

Governments and social media platforms should not rely on content removal as a solution to online scientific misinformation.

Society benefits from honest and open discussion on the veracity of scientific claims[21]. These discussions are an important part of the scientific process and should be protected. When these discussions risk causing harm to individuals or wider society, it is right to seek measures which can mitigate against this. This has often led to calls for online platforms to remove content and ban accounts[22, 23, 24]. However, whilst this approach may be effective and essential for illegal content (eg hate speech, terrorist content, child sexual abuse material) there is little evidence to support the effectiveness of this approach for scientific misinformation, and approaches to addressing the amplification of misinformation may be more effective.

In addition, demonstrating a causal link between online misinformation and offline harm is difficult to achieve[25, 26], and there is a risk that content removal may cause more harm than good by driving misinformation content (and people who may act upon it) towards harder-to-address corners of the internet[27].

Deciding what is and is not scientific misinformation is highly resource intensive[28] and not always immediately possible to achieve as some scientific topics lack consensus[29] or a trusted authority for platforms to seek advice from[30]. What may be feasible and affordable for established social media platforms may be impractical or prohibitively expensive for emerging platforms which experience similar levels of engagement (eg views, uploads, users)[31].

[21] Smith L, Stern N. 2011 Uncertainty in science and its role in climate policy. Phil. Trans. R. Soc. A. 369: 4818-4841. (doi.org/10.1098/rsta.2011.0149)

[22] UK Labour Party. Labour calls for emergency legislation to "stamp out dangerous anti vax content". See https://labour.org.uk/press/labour-calls-for-emergency-legislation-to-stamp-out-dangerous-anti-vax-content/ (accessed 4 November 2021)

[23] Covid vaccine: Social media urged to remove 'disinfo dozen'. BBC News. 26 March 2021. See https://www.bbc.co.uk/news/technology-56536390 (accessed 4 November 2021).

[24] Priti Patel urges social media to remove antivax posts. The Times. 11 February 2021. See https://www.thetimes.co.uk/article/priti-patel-tells-social-media-to-remove-antivax-posts-77ggm5tjn (accessed 4 November 2021).

[25] Miró-Llinares F, Aguerri J. 2021 Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'. European Journal of Criminology. (doi.org/10.1177/1477370821994059)

[26] Greene C, Murphy G. 2021 Quantifying the effects of fake news on behaviour: Evidence from a study on COVID-19 misinformation. Journal of Experimental Psychology. Applied. (doi.org/10.1037/xap0000371)

[27] Royal Society roundtable with Major Technology Organisations, March 2021.

[28] The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People. Motherboard. 23 August 2018. See https://www.vice.com/en/article/xwk9zd/how-facebook-content-moderation-works (accessed 4 November 2021)

[29] Facebook lifts ban on posts claiming Covid-19 was man-made. The Guardian.27 May 2021. See https://www.theguardian.com/technology/2021/may/27/facebook-lifts-ban-on-posts-claiming-covid-19-was-man-made (accessed 4 November 2021)

[30] Royal Society roundtable with Major Technology Organisations, March 2021.

[31] *Ibid*.

Furthermore, removing content may exacerbate feelings of distrust and be exploited by others to promote misinformation content[32, 33, 34].

Finally, misinformation sometimes comes from domestic political actors, civil society groups, or individual citizens who may, in good faith, believe in the content they are spreading, even if it may be harmful to others. It is clear that they may well regard direct action against their expression as outright censorship[35, 36].

Allowing content to remain on platforms with mitigations to manage its impact may be a more effective approach to prioritise. Examples of mitigations include demonetising content (eg by disabling ads on misinformation content); focusing on reducing amplification of those messages by preventing viral spread[37] or regulating the use of algorithmic recommender systems[38]; and annotating content with fact-check labels (see Recommendation 3).

These mechanisms would allow for open and informed discussions on scientific topics whilst acknowledging or addressing any controversies associated with the content.

As this report highlights, the online information environment has provided major benefits for collective scientific understanding by enabling the free exchange of knowledge amongst industry, academia, and members of the wider population. The Royal Society has long believed that the scientific community has a duty to communicate with the public in order to help people make informed decisions about their lives[39]. Removing content and driving users away from platforms which engage with scientific authorities risks making this harder, not easier, to achieve. A more nuanced, sustainable, and focused approach towards misinformation is needed.

[32] BrandNewTube. Ask The Experts (Covid 19 Vaccine) – Now Banned on YouTube and Facebook. See https://brandnewtube.com/watch/ask-the-experts-covid-19-vaccine-now-banned-on-youtube-and-facebook_qlsNohSleSgfz2J.html (accessed 4 November 2021).

[33] Banned.Video – the most banned videos on the internet. See https://www.banned.video/ (accessed 4 November 2021).

[34] Jansen S, Martin B. 2015 The Streisand Effect and Censorship Backfire. International Journal of Communication 9, 16.

[35] Trump says he will sue social media giants over 'censorship'. The Guardian. 7 July 2021. See https://www.theguardian.com/us-news/2021/jul/07/donald-trump-facebook-twitter-google-lawsuit (accessed 4 November 2021).

[36] Censorship concerns as talkRadio removed from YouTube. Society of Editors. 5 January 2021. See https://www.societyofeditors.org/soe_news/censorship-concerns-as-talkradio-removed-from-youtube/

[37] See Chapter 3 – tools and approaches for countering misinformation.

[38] Cobbe J, Singh J. 2019 Regulating recommending: Motivations, considerations, and principles. European Journal of Law and Technology 10, 3.

[39] The Royal Society. 1985 The Public Understanding of Science. See https://royalsociety.org/topics-policy/publications/1985/public-understanding-science/ (accessed 4 November 2021).

RECOMMENDATION 3

To support the UK's nascent fact-checking sector, programmes which foster independence and financial sustainability are necessary. To help address complex scientific misinformation content and 'information deserts', fact checkers could highlight areas of growing scepticism or dispute, for deeper consideration by organisations with strong records in carrying out evidence reviews, such as the UK's national academies and learned societies.

In response to the challenge of misinformation, a number of major online platforms have partnered with independent fact-checkers, certified by the International Fact-Checking Network (IFCN)[40], to help them identify and address misleading content[41]. Google and Facebook have themselves invested in independent fact-checking[42, 43]. As such, fact-checkers – and wider misinformation organisations who also partner with major platforms – have become a vital part of the infrastructure which ensures a healthy online information environment. Although fact-checkers have traditionally been affiliated with traditional media companies, this association is attenuating with a number of independent, dedicated fact-checking organisations being set up[44]. There are now estimated to be 290 fact-checking organisations across the world, with 40% of them based in Europe and North America[45].

A key challenge facing organisations working in the fact-checking sector is sustainable funding[46]. Many are SMEs or NGOs[47]. According to a 2016 Reuters Institute survey of European fact-checking organisations, more than half reported an annual expenditure of less than $50,000 and just over one quarter reported an annual expenditure of more than $100,000[48].

[40] IFCN-certified fact-checkers sign up to a Code of Principles (eg a commitment to transparency).

[41] Royal Society roundtable with Major Technology Organisations, March 2021.

[42] COVID-19: $6.5million to help fight coronavirus misinformation. Google News Initiative. 2 April 2020. See https://www.blog.google/outreach-initiatives/google-news-initiative/covid-19-65-million-help-fight-coronavirus-misinformation/ (accessed 4 November 2021).

[43] Facebook's investments in fact-checking and media literacy Facebook Journalism Project.15 June 2021. See https://www.facebook.com/journalismproject/programs/third-party-fact-checking/industry-investments (accessed 4 November 2021).

[44] The Fact-Check Industry. Columbia Journalism Review. 2019. See https://www.cjr.org/special_report/fact-check-industry-twitter.php (accessed 4 November 2021).

[45] Annual census finds nearly 300 fact-checking projects around the world. Duke Reporters' Lab. 22 June 2020. See https://reporterslab.org/annual-census-finds-nearly-300-fact-checking-projects-around-the-world/ (accessed 4 November 2021).

[46] Royal Society roundtable with Safety Technology Organisations, March 2021.

[47] *Ibid*.

[48] Reuters Institute for the Study of Journalism. 2016 The Rise of Fact-Checking Sites in Europe. See https://reutersinstitute.politics.ox.ac.uk/our-research/rise-fact-checking-sites-europe (accessed 4 November 2021).

A 2020 survey by the International Fact-Checking Network (IFCN) found that 43% of respondents said their main source of income was Facebook's Third Party Fact-Checking Program[49]. 42% reported their income comes from donations, memberships, or grants[50].

Providing users with tools to safely navigate the online information environment will be essential to combat harmful scientific misinformation. A survey conducted by YouGov for this report suggests there is already an appetite to fact-check information with the majority of respondents reporting that they would fact-check a suspicious or surprising scientific claim they read online[51]. The important role of fact-checkers is also recognised in the UK Government's Online Media Literacy Strategy[52]. These organisations generally provide a simple mechanism for users to verify the validity of claims made online and play an important role in informing content-moderation decisions. They provide a public benefit, form a core part of anti-misinformation initiatives, and should be supported.

Should the financial viability of organisations in this nascent sector collapse, it could have detrimental effects for the health of the online information environment. As impartiality and financial independence is critical to trust in these organisations, their options for funding are limited[53]. Philanthropic foundations and other grant funders are likely to continue to be necessary in the short to medium term. Platforms, funders and government need to consider sustainable models for long-term funding in this sector.

Furthermore, organisations with expertise in evidence synthesis (such as the UK's national academies) have a role to play and should be engaging with fact-checking organisations to help provide clarity on complex scientific misinformation content where feasible. This could involve fact-checkers highlighting areas of growing scepticism or dispute as being in need of deeper consideration, in order to address the challenges associated with information deserts where there is no clearly recognised scientific authority.

[49] International Fact-Checking Network. 2020 State of Fact Checking 2020. See https://www.poynter.org/wp-content/uploads/2020/06/IFCN_2020_state-of-fact-checking_ok.pdf (accessed 4 November 2021). This survey was 80 organisations that are either current verified signatories of the IFCN Code of Principles or undergoing the renewal process.

[50] *Ibid*.

[51] 68% said they would be likely to fact-check a suspicious scientific claim they saw online. Royal Society / YouGov, July 2021.

[52] HM Government. 2021 Online Media Literacy Strategy. See https://www.gov.uk/government/publications/online-media-literacy-strategy (accessed 4 November 2021).

[53] Royal Society roundtable with Safety Technology Organisations, March 2021.

RECOMMENDATION 4

Ofcom must consider interventions for countering misinformation beyond high-risk, high-reach social media platforms.

Under plans set out in the UK Government's Draft Online Safety Bill, regulations will apply depending on the number of users and/or the type of functionalities which exist on an online platform[54]. Category 1 services, described as 'high-risk, high-reach services' will be expected to take action on content deemed to be legal but harmful[55], which misinformation is likely to fall under[56]. These services will likely include mainstream social media platforms such as Facebook, YouTube, Twitter, and TikTok.

Given the size of these platforms, it is right for them to take appropriate action against harmful misinformation. However, many of these platforms are already taking steps to mitigate the effects of misinformation[57] and it is not clear that focusing on these high-reach services alone is enough to reduce the effects of harmful misinformation. This focus risks excluding small platforms, with significantly lower reach, from higher scrutiny. Some of these smaller platforms host harmful content banned elsewhere, garnering hundreds of thousands of views[58].

It is also unclear whether others, such as online retailers, will be expected to take action on harmful but legal content, despite there being examples of scientific misinformation content being promoted on their platforms[59].

Only a minority of internet users believe in the most prominent examples of scientific misinformation[60]. It may well be the case that this minority of users consume harmful misinformation content on fringe online platforms. However, by prioritising mainstream social media platforms, there is a risk that Ofcom will lack the necessary authority and capacity to address misinformation which exists elsewhere in the online information environment. We recommend a careful consideration of which platforms to focus interventions on and advise Ofcom to include fringe online platforms within their focus.

[54] HM Government. 2021 Draft Online Safety Bill. See https://www.gov.uk/government/publications/draft-online-safety-bill (accessed 4 November 2021).

[55] HM Government. 2020 Online Harms White Paper: Full government response to the consultation. See https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response (accessed 4 November 2021).

[56] UK Parliament. 2020 Misinformation in the COVID-19 Infodemic: Government Response to the Committee's Second Report. See: https://publications.parliament.uk/pa/cm5801/cmselect/cmcumeds/894/89402.htm (accessed 4 November 2021).

[57] See Chapter 2: 'Policies adopted by major online platforms'.

[58] The controversial COVID-19 'Ask the Experts' which discourages the use of vaccines is available on BrandNewTube and has 376,000 views, BrandNewTube, accessed September 2021.

[59] COVID-19: Waterstones and Amazon urged to add warning tags as anti-vaccination book sales surge. Sky News. 5 March 2021. See https://news.sky.com/story/waterstones-and-amazon-urged-to-add-warning-tags-as-anti-vaccination-book-sales-surge-12234972 (accessed 4 November 2021).

[60] 4-5% do not believe the COVID-19 vaccines are safe and 5% do not believe humans activity is responsible for climate change. 5% believe 5G technology is very harmful to human health, however a further 10% believe it is fairly harmful to human health. Royal Society / YouGov, July 2021.

RECOMMENDATION 5

Online platforms and scientific authorities should consider designing interventions for countering misinformation on private messaging platforms.

As users shift away from conversations on open, public platforms in favour of closed, private forums[61, 62], it is likely to become more difficult to analyse the online information environment and design interventions to counter misinformation. This shift will require a re-analysis of society's collective understanding of how information spreads online as lessons learned from public social media platforms are difficult to translate to private forums[63].

Designing interventions which preserve end-to-end encryption is essential for ensuring the security and privacy of people's conversations[64]. It is therefore necessary to design interventions which do not require prior knowledge of a message's content. Current examples of these include mechanisms to understand how messages spread[65] or to limit the number of times they can be shared[66], an option to forward a message to a fact-checker[67], and the creation of official accounts for scientific authorities[68].

[61] Royal Society roundtable with Major Technology Organisations, March 2021.

[62] Reuters Institute for the Study of Journalism. 2018 Digital News Report. See https://www.digitalnewsreport.org/survey/2018/ (accessed 4 November 2021).

[63] Funke D. 2017 Here's why fighting fake news is harder on WhatsApp than on Facebook. Poynter. See https://www.poynter.org/fact-checking/2017/here%C2%92s-why-fighting-fake-news-is-harder-on-whatsapp-than-on-facebook/ (accessed 4 November 2021).

[64] The Royal Society. 2016 Progress and research in cybersecurity: Supporting a resilient and trustworthy system for the UK. See https://royalsociety.org/-/media/policy/projects/cybersecurity-research/cybersecurity-research-report.pdf (accessed 4 November 2021).

[65] Bronstein M, Bruna J, LeCun Y, Szlam A, Vandergheynst P. 2017 Geometric Deep Learning: Going beyond Euclidean data. IEEE Signal Processing Magazine. 34, 18-42. (https://doi.org/10.1109/MSP.2017.2693418).

[66] WhatsApp. About forwarding limits. See https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en (accessed 4 November 2021).

[67] How Line is fighting disinformation without sacrificing privacy. Rest of World, 7 March 2021. See https://restofworld.org/2021/how-line-is-fighting-disinformation-without-sacrificing-privacy/ (accessed 4 November 2021).

[68] WHO Health Alert brings COVID-19 facts to billions via WhatsApp. World Health Organization. 20 March 2021. See https://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp (accessed 4 November 2021).

RECOMMENDATION 5 (continued)

Provenance enhancing technologies also present a potential solution here (see Chapter 3)[69]. These technologies would work by providing users with information about the origins (provenance) of a piece of online content as well as details of any alterations made to it[70]. This could provide a tool to help users verify the validity of any text, images, or videos they receive on a private or public communications channel.

Assuming trends towards private messaging continue[71], misinformation content is likely to become less visible to researchers, regulators, and the platforms themselves. This will therefore become an increasingly important area for those interested in fostering a healthy online information environment. Online platforms and scientific authorities need to consider this behaviour shift in information consumption and design interventions which can promote good quality information and mitigate any harmful effects from misinformation.

[69] McAuley D, Koene A, Chen J. 2020 Response to the Royal Society Call for Evidence: Technologies for Spreading and Detecting Misinformation. (https://doi.org/10.17639/wvk8-0v11).

[70] Content Authenticity Initiative. How it works. See https://contentauthenticity.org/how-it-works (accessed 4 November 2021).

[71] Reuters Institute for the Study of Journalism. 2018 Digital News Report. See https://www.digitalnewsreport.org/survey/2018/ (accessed 4 November 2021).

AREA FOR ACTION: ENABLING GREATER UNDERSTANDING OF THE ONLINE INFORMATION ENVIRONMENT

RECOMMENDATION 6

Social media platforms should establish ways to allow independent researchers access to data in a privacy compliant and secure manner.

Understanding the nature of information production and consumption is critical to ensuring society is prepared for future challenges which arise from the online information environment[72]. Analysis of the rich datasets held by social media platforms can help decision-makers understand the extent of harmful online content, how influential it is, and who is producing it. It should also help enable transparent, independent assessments of the effectiveness of counter-misinformation interventions.

The open nature of some platforms (eg Twitter) makes independent research easier to undertake whilst the more restricted nature of other platforms (eg Facebook, YouTube, TikTok) makes this more difficult[73].

Designing a solution to this and ensuring access to useful data for researchers is highly complex with significant challenges related to privacy, usability, and computing power[74]. Attempts to do so, such as Social Science One[75], have faced criticism from funders[76] and academics[77] for delays and insufficient access.

Developing a safe and privacy preserving means for independent and impartial analysis, such as a trusted research environment[78], is an important challenge for Research Councils, Legal Deposit Libraries, and social media platforms to overcome. Social media platforms have ultimate control of this data and should commence, or continue, efforts to find ways to provide access for independent researchers in a secure and privacy compliant manner.

The Royal Society has an ongoing programme of work related to privacy-preserving data analysis and the role of technology in protecting data subjects and is exploring past attempts, existing barriers, and viable solutions to enable privacy-preserving analysis of data[79].

[72] Omand D, Bartlett J, Miller C. 2012 Introducing social media intelligence (SOCMINT). Intelligence and National Security. 27, 801-823. (https://doi.org/10.1080/02684527.2012.716965).

[73] Arguedas A, Robertson C, Fletcher R, Nielsen R. 2021 Echo chambers, filter bubbles, and polarisation. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[74] The Alan Turing Institute. Data safe havens in the cloud. See https://www.turing.ac.uk/research/research-projects/data-safe-havens-cloud (accessed 4 November 2021).

[75] Harvard University. Social Science One: Building Industry Academic Partnerships. See https://socialscience.one/ (accessed 4 November 2021).

[76] Statement from Social Science Research Council President Alondra Nelson on the Social Media and Democracy Research Grants Program. Social Science Research Council. 27 August 2019. See https://www.ssrc.org/programs/social-data-initiative/social-media-and-democracy-research-grants/statement-from-social-science-research-council-president-alondra-nelson-on-the-social-media-and-democracy-research-grants-program/ (accessed 4 November 2021).

[77] Facebook Said It Would Give Detailed Data To Academics. They're Still Waiting. BuzzFeed News. 22 August 2019. See https://www.buzzfeednews.com/article/craigsilverman/slow-facebook (accessed 4 November 2021).

[78] Health Data Research UK. Trusted Research Environments. See https://www.hdruk.ac.uk/access-to-health-data/trusted-research-environments/ (accessed 4 November 2021).

[79] The Royal Society. Privacy Enhancing Technologies. See https://royalsociety.org/topics-policy/projects/privacy-enhancing-technologies/ (accessed 4 November 2021).

RECOMMENDATION 7

Focusing solely on the needs of current online platforms risks a repetition of existing problems, as new, underprepared, platforms emerge and gain popularity. To promote standards and guide start-ups, interested parties need to collaborate to develop examples of best practice for countering misinformation as well as datasets, tools, software libraries, and standardised benchmarks.

It is important to consider the health of the online information environment beyond the currently dominant online platforms. New platforms which grow quickly face a challenge of having to address large amounts of misinformation content without the benefit of years of experience[80]. Focusing solely on the needs of current online platforms risks a repetition of the same problems as new, underprepared, platforms emerge and gain popularity.

A particular challenge is the lack of data new platforms will have access to, in order to train automated detection systems for misinformation content[81]. There are already some encouraging examples of attempts to create datasets[82] and machine learning models[83] to assist with this problem. Researchers, policymakers, and platforms must work together to develop further similar initiatives. These should be developed and implemented in a secure, privacy-compliant manner, and published under open licenses, allowing reuse. To ensure high quality data input for machine learning models, the development of data assurance practices should be encouraged[84].

Knowledge for how best to ensure a healthy online information environment exists within various fields of expertise, including computational sociology[85], open-source intelligence[86], library and information science[87], and media literacy[88]. As such, calls for collaboration should encompass all interested parties who can usefully contribute to the development of best practice tools and guidance for future online platforms.

[80] Royal Society roundtable with Major Technology Organisations, March 2021.

[81] *Ibid*.

[82] SFU Discourse Lab. MisInfoText. See https://github.com/sfu-discourse-lab/MisInfoText (accessed 4 November 2021).

[83] Khan J, Khondaker M, Afroz S, Uddin G, Iqbal A. 2021 A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications. 4. (https://doi.org/10.1016/j.mlwa.2021.100032).

[84] Assurance, trust, confidence – what does it all mean for data? Open Data Institute.18 June 2021. See https://theodi.org/article/assurance-trust-confidence-what-does-it-all-mean-for-data/ (accessed 4 November 2021).

[85] Ciampaglia G. 2017 Fighting fake news: A role for computational social science in the fight against digital misinformation. Journal of Computational Social Science. 1, 147-153. (https://doi.org/10.1007/s42001-017-0005-6).

[86] Bellingcat. Bellingcat's Online Investigation Toolkit. See https://docs.google.com/spreadsheets/d/18rtqh8EG2q1xBo2cLNyhIDuK9jrPGwYr9DI2UncoqJQ/edit#gid=930747607 (accessed 4 November 2021).

[87] Revez J, Corujo L. 2021 Librarians against fake news: A systematic literature review of library practices (Jan 2018 – September 2020). The Journal of Academic Librarianship. 47. (https://doi.org/10.1016/j.acalib.2020.102304).

[88] Guess *et al.* 2020 A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proceedings of the National Academy of Sciences July 2020. 117, 15536-15545. (https://doi.org/10.1073/pnas.1920498117).

AREA FOR ACTION: CREATING A HEALTHY AND TRUSTWORTHY ONLINE INFORMATION ENVIRONMENT

RECOMMENDATION 8

Governments and online platforms should implement policies that support healthy and sustainable media plurality.

Many news outlets are a key source of good quality[89] and trusted[90] information. The online information environment has provided, and continues to provide, an ecosystem which allows for increased media plurality with few barriers to entry[91, 92]. It is a feature which exposes users to a wide range of viewpoints and prevents a concentration of influence over public opinion[93]. Reporting about science has also benefited from this plurality with new science and technology media outlets gaining significant online followings[94].

Moves to elevate or prioritise content from 'trustworthy' news outlets[95] in social media feeds presents a risk to online media plurality, is likely to favour established, traditional media outlets over new media outlets[96], and would not necessarily reduce exposure to misinformation content[97]. Although strong arguments have been put forward for online platforms to determine the quality of news content[98], efforts to compare and rate the trustworthiness of different media outlets (eg with nutrition labels) have proven to be complex with some attempts attracting controversy[99, 100].

[89] Ofcom. 2020 News Consumption in the UK. See https://www.ofcom.org.uk/__data/assets/pdf_file/0013/201316/news-consumption-2020-report.pdf (accessed 4 November 2021).

[90] Reuters Institute for the Study of Journalism. 2021 Digital News Report. See https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021 (accessed 4 November 2021).

[91] Reuters Institute for the Study of Journalism. 2012 News Plurality in a Digital World. See https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2017-11/News%20Plurality%20in%20a%20Digital%20World_0.pdf (accessed 4 November 2021).

[92] Open Society Foundations. 2014 Digital journalism: Making news, breaking news. See https://www.opensocietyfoundations.org/uploads/02fc2de9-f4a5-4c07-8131-4fe033398336/mapping-digital-media-overviews-20140828.pdf (accessed 4 November 2021).

[93] Ofcom. 2021 The Future of Media Plurality in the UK. See https://www.ofcom.org.uk/__data/assets/pdf_file/0012/220710/media-plurality-in-the-uk-condoc.pdf (accessed 4 November 2021).

[94] Examples: IFLScience, Rest of World, UNILAD Tech.

[95] Mosseri A. 2018 Helping ensure news on Facebook is from trusted sources. Facebook. 19 January 2018. See https://about.fb.com/news/2018/01/trusted-sources/ (accessed 4 November 2021).

[96] Facebook is changing news feed (again) to stop fake news. Wired. 4 October 2019. See https://www.wired.com/story/facebook-click-gap-news-feed-changes/ (accessed 4 November 2021).

[97] Tsfati Y, Boomgaarden H, Strömbäck J, Vliegenthart R, Damstra A, Lindgren E. 2019 Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. Annals of the International Communication Association. 44, 157-173. (https://doi.org/10.1080/23808985.2020.1759443).

[98] The Cairncross Review. 2019 A sustainable future for journalism. See https://www.gov.uk/government/publications/the-cairncross-review-a-sustainable-future-for-journalism (accessed 4 November 2021).

[99] 'We were wrong': US news rating tool boosts Mail Online trust ranking after talks with unnamed Daily Mail exec. PressGazette. 31 January 2019. See https://www.pressgazette.co.uk/we-were-wrong-us-news-rating-tool-boosts-mail-online-trust-ranking-after-talks-with-unnamed-daily-mail-exec/ (accessed 4 November 2021).

[100] Wikipedia bans Daily Mail as 'unreliable' source. The Guardian. 8 February 2017. See https://www.theguardian.com/technology/2017/feb/08/wikipedia-bans-daily-mail-as-unreliable-source-for-website (accessed 4 November 2021).

RECOMMENDATION 8 (continued)

Furthermore, unilateral decisions about how algorithms present news content in social media feeds and search engines can negatively impact the reach, traffic, and economic performance of both traditional and new media outlets[101].

Governments and online platforms need to consider the impact of any future policies on media plurality and take action to ensure a sustainable future for public interest journalism[102]. Robust, diverse, independent news media and education (see Recommendation 9) together can make people more resilient in the face of any potentially harmful misinformation they come across.

[101] Bailo F, Meese J, Hurcombe E. 2021 The Institutional Impacts of Algorithmic Distribution: Facebook and the Australian News Media. Social Media + Society. 7. (https://doi.org/10.1177%2F20563051211024963).

[102] The Cairncross Review. 2019 A sustainable future for journalism. See https://www.gov.uk/government/publications/the-cairncross-review-a-sustainable-future-for-journalism (accessed 4 November 2021).

RECOMMENDATION 9

The UK Government should invest in lifelong, nationwide, information literacy initiatives.

Ensuring that current and future populations can safely navigate the online information environment will require significant investment in digital information literacy, ensuring that people can effectively evaluate online content. In practice, this could include education on how to assess URLs[103], how to reverse image search[104], and how to identify a deepfake[105].

This education should not be limited to those in schools, colleges, and universities, but extended to all people of all ages. Older adults face a particular challenge with misinformation as they are more likely to be targeted and more likely to be susceptible than younger adults[106]. These groups could be reached through public information campaigns, in workplaces, or on social media platforms. Current initiatives such as the UK Government's 'Don't Feed the Beast' campaign[107] and the Check Before You Share toolkit[108] should be assessed for their effectiveness and improved where necessary.

As the nature of the online information environment is likely to continue evolving over time with new platforms, technologies, actors, and techniques, it is important to consider information literacy as a life skill, supplemented with lifelong learning. These initiatives should be carefully tailored and designed to support people from a broad range of demographics.

There have been widespread calls[109, 110, 111, 112] for digital information literacy to form a core part of future strategies to ensure people can safely navigate the online information environment. Successful implementation of the UK Government's Online Media Literacy Strategy is an important next step[113].

[103] Polizzi G. 2020 Fake news, Covid-19 and digital literacy: Do what experts do. London School of Economics. 17 June 2020. See https://blogs.lse.ac.uk/medialse/2020/06/17/fake-news-covid-19-and-digital-literacy-do-what-the-experts-do/ (accessed 4 November 2021).

[104] *Ibid.*

[105] Microsoft. Spot the Deepfake. See https://www.spotdeepfakes.org/en-US (accessed 4 November 2021).

[106] Moore R, Hancock J. 2020 Older Adults, Social Technologies, and the Coronavirus Pandemic: Challenges, Strengths, and Strategies for Support. Social Media + Society. (https://doi.org/10.1177%2F2056305120948162).

[107] HM Government. 2020 Government cracks down on spread of false coronavirus information online. See https://www.gov.uk/government/news/government-cracks-down-on-spread-of-false-coronavirus-information-online (accessed 4 November 2021).

[108] HM Government. Check Before You Share Toolkit. See https://dcmsblog.uk/check-before-you-share-toolkit/ (accessed 4 November 2021).

[109] European Commission. 2018 Final report of the High Level Expert Group on Fake News and Online Disinformation. See https://www.ecsite.eu/activities-and-services/resources/final-report-high-level-expert-group-fake-news-and-online (accessed 4 November 2021).

[110] House of Commons Digital, Culture, Media and Sport Committee. 2019 Disinformation and 'fake news': Final Report. See https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf (accessed 4 November 2021).

[111] House of Lords Select Committee on Democracy and Digital Technologies. 2020 Digital Technology and the Resurrection of Trust. See https://committees.parliament.uk/publications/1634/documents/17731/default/ (accessed 4 November 2021).

[112] The Alan Turing Institute. 2021 Understanding vulnerability to online misinformation. See https://www.turing.ac.uk/sites/default/files/2021-02/misinformation_report_final1_0.pdf (accessed 4 November 2021).

[113] HM Government. 2021 Online Media Literacy Strategy. See https://www.gov.uk/government/publications/online-media-literacy-strategy (accessed 4 November 2021).

AREA FOR ACTION: ENABLING ACCESS TO SCIENTIFIC INFORMATION

RECOMMENDATION 10

Academic journals and institutions should continue to work together to enable open access publishing of academic research.

The ability to easily share and find high quality information is one of the greatest benefits of the online information environment and likely explains why the majority of respondents to the Society's survey believe the internet has improved the public's understanding of science[114]. In particular, the internet's role in opening access to academic research, which would otherwise be locked within physical journals, can often be transformative for society's collective understanding of the world.

Ensuring ease of access to academic research online helps promote more accurate verification of results, reduces duplication of work, and improves public trust in science[115]. As strong supporters of open science[116], the Royal Society is currently working towards transitioning its own primary research journals to open access which will help maximise the dissemination and impact of high-quality scientific research[117].

The COVID-19 pandemic has further incentivised the need for open access publishing[118, 119], and has demonstrated its benefits[120] These benefits can and should be realised for a broad range of societal problems, beyond the pandemic. Moves towards open access publishing[121] are to be welcomed, and academic journals and institutions should work together to enable further open access publishing of academic research.

Novel aspects of open access research, such as the growing popularity of preprints[122] or the use of citations as an indicator of quality[123], have been subject to debate in recent years. We note these concerns and encourage institutions to consider lessons learned for the next generation of academic publishing.

[114] 61% believe the internet has made the public's understanding of science better. Royal Society / YouGov, July 2021.

[115] OECD. 2015 Making Open Science a Reality. See https://www.oecd-ilibrary.org/science-and-technology/making-open-science-a-reality_5jrs2f963zs1-en (accessed 4 November 2021).

[116] This includes open publishing and open data, see https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/ (accessed 10 November 2021).

[117] The Royal Society. 2021 The Royal Society sets 75% threshold to 'flip' its research journals to Open Access over the next five years. See https://royalsociety.org/news/2021/05/royal-society-open-access-plans/ (accessed 4 November 2021).

[118] UNESCO. Open access to facilitate research and information on COVID-19. See https://en.unesco.org/covid19/communicationinformationresponse/opensolutions (accessed 4 November 2021).

[119] Kiley R. 2020 Three lessons COVID-19 has taught us about Open Access publishing. London School of Economics. 6 October 2020. See https://blogs.lse.ac.uk/impactofsocialsciences/2020/10/06/39677/ (accessed 4 November 2021.

[120] European Molecular Biology Laboratory. 2020 Open data sharing accelerates COVID-19 research. See https://www.ebi.ac.uk/about/news/announcements/open-data-sharing-accelerates-covid-19-research (accessed 4 November 2021).

[121] UK Research and Innovation. 2021 UKRI announces new Open Access Policy, UK Research and Innovation. See https://www.ukri.org/news/ukri-announces-new-open-access-policy/ (accessed 4 November 2021).

[122] Soderberg C, Errington T, Nosek B. 2020 Credibility of preprints: an interdisciplinary survey of researchers. Royal Society Open Science. 7, 201520. (https://doi.org/10.1098/rsos.201520).

[123] Aksnes D, Langfeldt L, Wouters P. 2019 Citations, citation indicators, and research quality: An overview of basic concepts and theories. SAGE Open. (https://doi.org/10.1177%2F2158244019829575).

RECOMMENDATION 11

The frameworks governing electronic legal deposit should be reviewed and reformed to allow better access to archived digital content.

In 2013, the UK Government introduced new regulations that required digital publications to be systematically preserved as part of something known as legal deposit. Legal deposit has existed in English law since 1662 and obliges publishers to place at least one copy of everything they publish in the UK and Ireland – from books to music and maps – at a designated library.

Since it was extended to include digital media, the six designated legal deposit libraries in the UK have accumulated around 700 terabytes of archived web data as part of the UK Web Archive, growing by around 70TB every year. The libraries automatically collect – or crawl – UK websites at least once a year to gather a snapshot of what they contain, while some important websites such as news sites are collected daily. They also collect ebooks, electronic journals, videos, pdfs and social media posts – almost everything that is available in a digital format.

Access to this material is extremely limited. Due to the current legislative framework, historic pages for only around 19,000 or so websites can be accessed through the Web Archive's online portal. These are sites where their creators have given explicit permission to allow open access to their content, however contacting every UK website in this way is almost impossible. For the rest, even though access is permitted, and the material is held digitally, researchers must travel to one of nine named sites in person. The framework also permits only one researcher to use a piece of material at any one time; an arbitrary limitation when it comes to digital access.

This framework for access is now out-of-date to how people access and use data, and severely limits the value that trustworthy libraries and archives are able to offer[124]. Opening up the Web Archive would allow it to be mined at scale for high quality information using modern text analysis methods or artificial intelligence. It would enable researchers, businesses, journalists and anyone else with an interest to uncover trends or information hidden in web pages from the past. This will become increasingly important as the online information environment matures and vital source material is digitally archived (see Chapter 4).

The frameworks governing electronic legal deposit need to be reviewed and reformed to allow wider access. Such a review would need to consider the data held in these legal deposits that remains commercially valuable, such as newspaper archives. Rather than act as a barrier to access, systems such as micropayments – like those to authors of books borrowed from libraries already – could be applied to such material in order to support broader access.

[124] Gooding P, Terras M, Berube L. 2019 Towards user-centric evaluation of UK non-print legal deposit: A digital library futures White Paper. Digital Library Futures. See http://eprints.gla.ac.uk/186755/ (accessed 4 November 2021).

BOX 1

Wider questions for further research.

The issues that this report touches on are broad and complex, and we acknowledge that many important issues have not been directly addressed here. These include the following questions and challenges for ongoing research.

- The history of science communication and public engagement with science has been through a long-term process of evolution and development.
  How has the online information environment affected the behaviour and outputs of traditional science communicators? (eg public service broadcasters, university press offices, individual researchers.

- Issues such as vaccine hesitancy are complex phenomena, which may include questions of trust in science but also relate to historic relationships between institutions and society. Some mistrust may stem from the way that marginalised communities have been negatively affected by the actions of those institutions. There are deep questions relating to social justice that need to be addressed.
  What role, if any, does the internet have in building, or damaging, trust amongst marginalised communities, in particular between those communities and public institutions?

- The methods for delivering and presenting content to people using black-box algorithmic recommendation systems has been the focus of much attention in public discourse on internet regulation.
  How significant a role do algorithmic recommendation systems play in amplifying harmful scientific misinformation content and how should they be regulated?

- Scientific topics can often attract the attention of political elites (eg politicians, political commentators), particularly when the topic is dominating the news cycle. The politicisation of these topics has been identified as a factor which may contribute to public opinion becoming divided on scientific issues.
  How do the actions of political elites affect the spread of online scientific misinformation and which groups are most vulnerable?

- The ability to be anonymous online can sometimes be highly beneficial for people (eg for victims of domestic abuse, or those living under authoritarian regimes). However, it has also been under scrutiny in public discourse for its possible role in contributing to online abuse and the spread of misinformation content. **How significant a role do anonymous accounts play in the promotion of online scientific misinformation?**
- The ability to generate financial profit (eg through advertising revenue, public donations) plays a significant role in the production and dissemination of highly engaging and emotive online content. **How could the business models of online platforms be adapted to minimise financial incentives for producing and promoting scientific misinformation content?**
- The growing popularity of preprints has had significant benefits for the rapid dissemination of important research findings, especially during the COVID-19 pandemic. However, there have been questions over how the robustness of these findings have been communicated and understood. **How significant an effect, if any, is the rising popularity of preprints having on the spread of scientific misinformation?**
- Data held as part of the UK Web Archive can offer invaluable insights to researchers exploring changing patterns of behaviour towards information production and consumption. A limited amount of this can be accessed through the Web Archive's online portal. Currently, researchers must travel to one of nine physical sites to gain limited access to the rest of this data. **How can researcher access to public national Web Archive data be improved?**
- Although a clear link can be made between the presence of online misinformation and harmful actions taken offline, evidencing a causal link and clearly defining 'harm' remains a challenge. **How best can researchers, regulators, and platforms evidence a causal link between online misinformation and offline harm? How should 'harm' in a regulatory context be defined?**

-19

Chapter one
Why do we need trustworthy information?

Left
© CasPhotography.

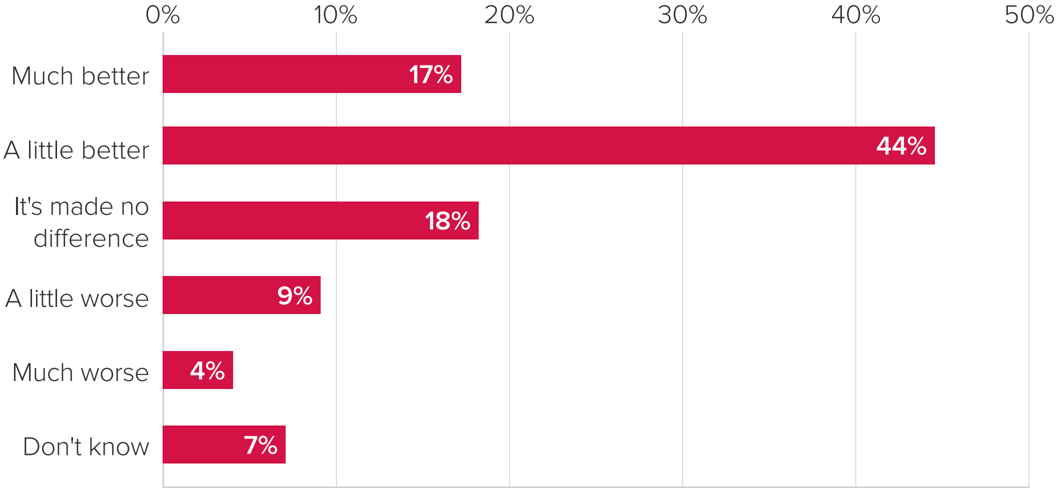
Why do we need trustworthy information?

The open access to knowledge enabled by the online information environment provides major benefits to both individuals and wider society. For scientists, the internet has reduced barriers to the publication of research outputs and allowed for greater scrutiny of results[125]. This has helped accelerate the pace of scientific enquiry and significantly promoted all forms of innovation in society.

For general internet users, who require good information to guide behaviours and support balanced decision making, the online information environment provides access to a broad range of educational material. This includes access to educational resources on healthcare treatments, skills development, and current affairs which, in turn, minimises the need for people to visit a doctor[126], attend a formal educational institution[127], or purchase a physical newspaper[128]. The potential societal benefits from this are reduced pressures on public services, a more skilled population, and a better-informed electorate.

FIGURE 1

Survey results for the question: Overall, how much better, or worse do you think that the internet has made the public's understanding of science, or has it made no difference?



Source: Royal Society / YouGov, July 2021. (n=2,019)

[125] Merrifield R. 2021 How pandemic-driven preprints are driving open scrutiny of research. European Commission. 1 April 2021. See https://ec.europa.eu/research-and-innovation/en/horizon-magazine/how-pandemic-driven-preprints-are-driving-open-scrutiny-research (accessed 4 November 2021).

[126] NHS England. Digital Inclusion in Health and Care. See https://www.england.nhs.uk/ltphimenu/digital-inclusion/digital-inclusion-in-health-and-care/ (accessed 4 November 2021).

[127] OECD. 2020 The potential of online learning for adults: Early lessons from the COVID-19 crisis. See https://www.oecd.org/coronavirus/policy-responses/the-potential-of-online-learning-for-adults-early-lessons-from-the-covid-19-crisis-ee040002/ (accessed 4 November 2021).

[128] Reuters Institute for the Study of Journalism. 2021 Digital News Report. See https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021 (accessed 4 November 2021).

The low to zero cost of publishing information on the internet, however, has led to a phenomenon known as 'information overload' with an overabundance of content competing for attention[129]. Within this attention-seeking content is information which may be misleading, inaccurate, or dangerous. Prominent examples of this include content which creates undue fear about vaccines, distrust in the integrity of elections, and divisions between communities. The influence of harmful online content has been pointed to as a possible contributing factor to acts of vandalism[130], suicide[131], and genocide[132].

During the COVID-19 pandemic, both the best and worst qualities of the online information environment have been exposed. The ability to easily share data and research freely and quickly online, in addition to the use of remote working tools has enabled international collaboration on an unprecedented scale. A key example is the use of the online open-source repository, GitHub. During the pandemic, GitHub contributors have provided access to datasets, contact tracing apps, tracking tools, forecasting techniques, and diagnostic solutions[133]. Meanwhile, long-held fears that a global pandemic would be made worse by viral misinformation[134] came to fruition with doubts being spread (including by political leaders and celebrities[135, 136, 137, 138]) about the efficacy of official public health advice and false remedies being sold online[139]. At one stage of the pandemic, the number one bestseller on Amazon for books about children's vaccination and immunisation was one authored by a renowned vaccine conspiracy theorist[140].

The ability to easily share data and research freely and quickly online has enabled international collaboration on an unprecedented scale.

[129] Bawden D, Robinson L. 2020 Information Overload: An Overview. In: Oxford Encyclopaedia of Political Decision Making. Oxford, Oxford University Press. See https://core.ac.uk/download/pdf/286715468.pdf (accessed 4 November 2021).

[130] Experts say echo chambers from apps like Parler and Gab contributed to attack on Capitol. ABC News. 12 January 2021. See https://abcnews.go.com/US/experts-echo-chambers-apps-parler-gab-contributed-attack/story?id=75141014 (accessed 4 November 2021).

[131] Carlyle K, Guidry J, Williams K, Tabaac A, Perrin P. 2018 Suicide conversations on Instagram: contagion or caring? Journal of Communication in Healthcare. 11, 12-18. (https://doi.org/10.1080/17538068.2018.1436500).

[132] United Nations Human Rights Council. 2018 Report of the detailed findings of the independent fact-finding mission on Myanmar. See https://digitallibrary.un.org/record/1643079?ln=en (accessed 4 November 2021).

[133] Wang L, Li R, Zhu J, Bai G, Wang H. 2020 When the Open-Source Community Meets COVID-19: Characterising COVID-19 themed GitHub Repositories. See https://arxiv.org/abs/2010.12218 (accessed 4 November 2021).

[134] Larson H. 2018 The biggest pandemic risk? Viral misinformation. Nature. 16 October 2018. See https://www.nature.com/articles/d41586-018-07034-4 (accessed 4 November 2021).

[135] Coronavirus: Outcry after Trump suggests injecting disinfectant as treatment. BBC News. 24 April 2020. See https://www.bbc.co.uk/news/world-us-canada-52407177 (accessed 4 November 2021).

[136] Brazil's Bolsonaro warns virus vaccine can turn people into 'crocodiles'. France 24. 18 December 2020. See https://www.france24.com/en/live-news/20201218-brazil-s-bolsonaro-warns-virus-vaccine-can-turn-people-into-crocodiles (accessed 4 November 2021).

[137] Actress Letitia Wright criticised for sharing vaccine doubter's video. BBC News. 4 December 2020. See https://www.bbc.co.uk/news/entertainment-arts-55185119 (accessed 4 November 2021).

[138] Keeps D. 2021 Eric Clapton's anti-vaccine diatribe blames 'propaganda' for 'disastrous' experience. Rolling Stone. 16 May 2021. See https://www.rollingstone.com/music/music-news/eric-clapton-disastrous-vaccine-propaganda-1170264/ (accessed 4 November 2021).

[139] Hansson *et al.* 2021 COVID-19 information disorder: six types of harmful information during the pandemic in Europe. Journal of Risk Research. 24, 380-393. (https://doi.org/10.1080/13669877.2020.1871058).

[140] COVID-19: Waterstones and Amazon urged to add warning tags as anti-vaccination book sales surge. Sky News. 5 March 2021. See https://news.sky.com/story/waterstones-and-amazon-urged-to-add-warning-tags-as-anti-vaccination-book-sales-surge-12234972 (accessed 4 November 2021).

This competition between those seeking attention for trustworthy information and those seeking attention for untrustworthy information is emblematic of the broad challenges facing the online information environment. In the online 'attention economy'[141] – the systems for filtering, supplying, and promoting content are of paramount importance. They have the potential to both upgrade and degrade the quality of life for individual users and wider society. Research in this area is generally focused on understanding the side-effects of these systems and theorising concepts to explain how they promote positive and negative impacts.

Evidence of impact

The internet has had a mixed impact on the information environment. Assertions about social media platforms creating filter bubbles (a "unique universe of information for each of us")[142] in which users are directed by algorithms towards hyper-personalised news content are not well-evidenced, with academic studies finding little to no support for the hypothesis[143]. Instead, the evidence shows that algorithmic selection generally diversifies the news content an internet user consumes[144].

The similar, but distinct, phenomenon of 'echo chambers' ("a bounded, enclosed media space that has the potential to both magnify the messages delivered within it and insulate them from rebuttal")[145] is also potentially overstated in public commentary[146]. However, whilst studies show that the vast majority of people do not opt into echo chambers, there is some evidence that this may be the case for those with highly partisan views[147].

[141] Simon H. 1971 Designing organizations for an information-rich world. In: Greenberger M (ed.) Computers, communications, and the public interest. Baltimore, The John Hopkins Press.

[142] Pariser E, 2011 Filter Bubble. London, UK: Penguin.

[143] Arguedas A, Robertson C, Fletcher R, Nielsen R. 2021 Echo chambers, filter bubbles, and polarisation. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[144] *Ibid.*

[145] Jamieson K, Cappella J. 2008 Echo Chamber. Oxford, UK: Oxford University Press.

[146] Röttger P, Vedres B. The Information Environment and its Effects on Individuals and Groups. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[147] Arguedas A, Robertson C, Fletcher R, Nielsen R. 2021 Echo chambers, filter bubbles, and polarisation. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

There is limited research and evidence outside of the United States on the prevalence of polarisation induced by social media use. Research from the US, however, suggests that exposure to like-minded political content – especially from political elites – can potentially polarise or strengthen the attitudes of those with existing partisan views. This could also apply to public conversations on science, however there is, as yet, little empirical research on this[148].

Polling commissioned for this report provides an insight into the penetration of harmful scientific misinformation amongst internet users in the UK[149]. The polling found that 5% of respondents do not believe the COVID-19 vaccines to be safe, 5% of respondents do not believe humans are at all responsible for climate change, and 15% believe 5G technology is harmful to human health. This suggests that harmful online scientific conspiracy theories are believed by a minority – albeit a significant minority – of the population. The majority of respondents believe that the internet has improved the public's understanding of science and feel confident to challenge suspicious scientific claims made by friends and family members.

Whilst the evidence shows that negative aspects of the online information environment may be affecting only a few percent of internet users, this still equates to a significant number of people. For example, if it is the case – as per the survey conducted for this report – that 5% of the UK's online adult population[150] do not believe the COVID-19 vaccines are safe, this would be equivalent to approximately 2.4 million people. If even a small fraction of these people vocalised their opinions on social media, it could lead to thousands of posts online which may or may not influence others (including key decision-makers).

Furthermore, there are numerous reports of harmful behaviours which have been linked to misinformation consumed online. For example, in 2020, an online conspiracy theory linking 5G telecommunications towers to the spread of COVID-19 is alleged to have contributed to multiple arson attacks as well as the stabbing and hospitalisation of an engineer[151].

Researchers have also found that increased susceptibility to misinformation negatively affects people's self-reported compliance with COVID-19 public health measures as well as their willingness to be vaccinated[152]. Estimates have found that online pages and accounts promoting anti-vaccination messages have millions of followers and have surged in recent years[153].

Harmful online scientific conspiracy theories are believed by a minority – albeit a significant minority – of the population.

[148] *Ibid.*

[149] Royal Society / YouGov, July 2021.

[150] Office for National Statistics. Internet users, UK: 2020. See https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2020 (accessed 5 November 2021).

[151] 77 cell phone towers have been set on fire so far due to a weird coronavirus 5G conspiracy theory. Business Insider. 6 May 2020. See https://www.businessinsider.com/77-phone-masts-fire-coronavirus-5g-conspiracy-theory-2020-5?r=US&IR=T (accessed 4 November 2021).

[152] Roozenbeek *et al.* 2020 Susceptibility to misinformation about COVID-19 around the world. Royal Society Open Science. 7, 201199. (https://doi.org/10.1098/rsos.201199).

[153] The Royal Society. 2020 COVID-19 vaccine deployment: Behaviour, ethics, misinformation and policy strategies. See https://royalsociety.org/-/media/policy/projects/set-c/set-c-vaccine-deployment.pdf (accessed 4 November 2021).

Misinformation about climate change – which the public are more likely to consider harmful than misinformation about 5G[154] – has also reached millions of online users in recent years with a report by Avaaz suggesting that tactics are switching away from climate denialism to climate inactivism[155]. This involves seeding doubt in climate science, making unfound assertions about climate solutions, and promoting 'doomism' (that it is too late to act)[156].

Globally, 'fake news' is regarded by internet users as a greater concern than other risks such as online abuse and fraud. A recent worldwide poll conducted by Lloyds Register Foundation and Gallup found that 57% of internet users across society view fake news as a major concern, particularly in regions of high economic inequality[157].

Taken together, current evidence suggests that whilst the negative aspects of the online information environment appear to be affecting only a minority of internet users, the harm caused creates significant concern and can have real-world consequences – although these may not be as a result of echo chambers and filter bubbles. Combined with early research which suggests that information overload is affecting people's attentional capacities[158, 159] there is sufficient reason to be concerned about the negative impacts of the online information environment and to consider potential mitigations such as those set out in the Recommendations chapter.

[154] 83% consider misinformation about climate change to be harmful, 67% consider misinformation about 5G technology to be harmful. Royal Society / YouGov, July 2021.

[155] Avaaz. 2021 Facebook's Climate of Deception: How Viral Misinformation Fuels the Climate Emergency. See https://secure.avaaz.org/campaign/en/facebook_climate_misinformation/ (accessed 4 November 2021).

[156] *Ibid*.

[157] The Lloyd's Register Foundation World Risk Poll. See https://wrp.lrfoundation.org.uk/explore-the-poll/fake-news-is-the-number-one-worry-for-internet-users-worldwide/ (accessed 9 November 2021).

[158] Firth *et al.* 2019 The "online brain": how the Internet may be changing our cognition. World Psychiatry. 18, 119-129. (https://doi.org/10.1002/wps.20617).

[159] Information Overload Helps Fake News Spread, and Social Media Knows It. Scientific American. 1 December 2020. See https://www.scientificamerican.com/article/information-overload-helps-fake-news-spread-and-social-media-knows-it/ (accessed 4 November 2021).

CASE STUDY 1

Vaccine misinformation.

During the COVID-19 pandemic, online misinformation about the vaccines has attracted significant attention amongst the media, politicians, and the wider public. Conspiracy theories about the vaccines include accusations that they implant microchips, can alter DNA, and that they were created prior to the onset of the pandemic[160]. Misinformation about vaccines, however, predates the internet and is not a new concern. The history of how misinformation affected the perception of the polio, pertussis, and MMR vaccines is explored in a literature review commissioned for this report[161] and outlined below.

The concept of a single 'anti-vax movement' is a misleading one. A range of different groups are involved in creating and spreading anti-vaccination material and, those holding 'anti-vax' views have different concerns, for example they may be very concerned about child safety and potential side effects of vaccines. They may be in broad opposition to government and public institutions or may adhere to particular political philosophies such as libertarianism. These different groups tend not to interact, coalescing around different interests and in different fora.

A large proportion of those spreading anti-vaccination material do so with genuine concern for the risk vaccines might cause to individuals' health or to society at large. They share material based on the belief that the material is trustworthy and helpful to others within their network[162]. Critically, it is often information which they do not see being shared by mainstream news or medical sources despite its perceived importance.

The introduction of routine vaccinations for pertussis in the 1950s (in the UK) drastically reduced incidences of the illness from an average of 122,000 cases per year in 1956 to 20,000 by the 1970s. However, following the publication of a book *A Shot in the Dark* and a television programme *Vaccine Roulette*, which alleged that the vaccine was giving children severe disabilities, concerns about the vaccine peaked and led to the formation of the Association of Parents of Vaccine Damaged Children[163]. Furthermore, publications from medical professionals, including the Hospital of Sick Children (now known as Great Ormond Street Hospital) and a doctor called Gordon Stewart, questioned the safety of the vaccine and added to public concern. This led to a drastic fall in confidence and to further epidemics[164]. ▶

[160] Islam *et al.* 2021 COVID-19 vaccine rumours and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. PLoS ONE. (https://doi.org/10.1371/journal.pone.0251605).

[161] Cabrera Lalinde I. 2021 How misinformation affected the perception of vaccines in the 20th century based on the examples of polio, pertussis and MMR vaccines. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[162] Moran M, Lucas M, Everhart K, Morgan A, Prickett E. 2016 What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. Journal of Communication in Healthcare. 9, 151-163. (https://doi.org/10.1080/17538068.2016.1235531).

[163] Cabrera Lalinde I. 2021 How misinformation affected the perception of vaccines in the 20th century based on the examples of polio, pertussis and MMR vaccines. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[164] *Ibid*.

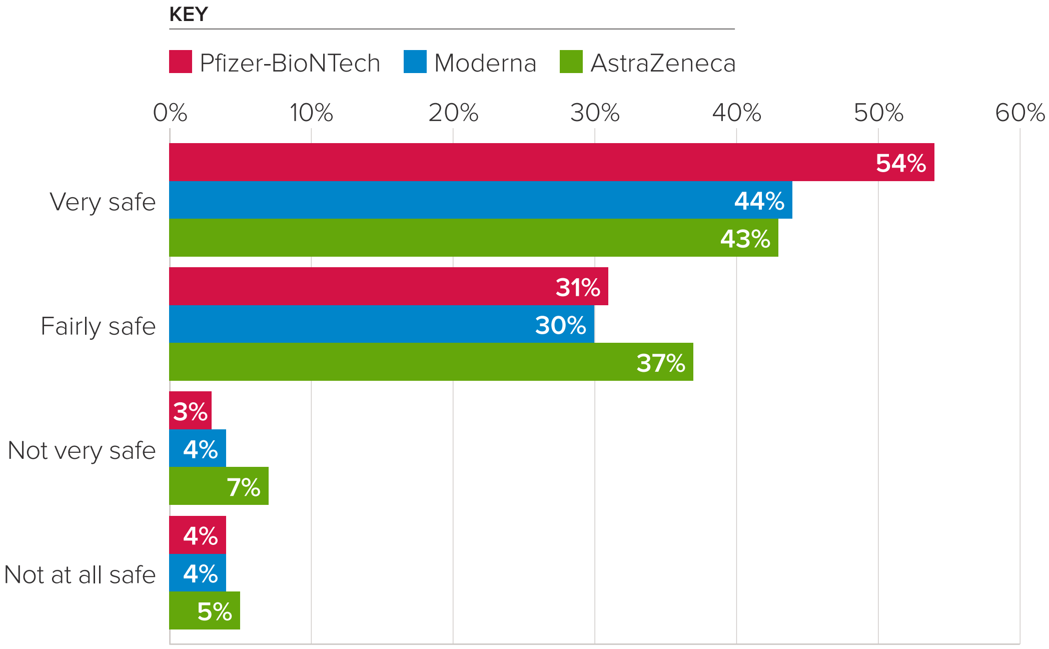
CASE STUDY 1 (continued)

A similar series of events occurred with the MMR vaccine in the 1990s. A heavily discredited[165] paper by former physician, Andrew Wakefield, published in the Lancet, claimed to establish a link between the vaccine and autism in children. The study was later partially retracted by 10 of the 12 authors, who stated that no "causal link had been established between MMR vaccine and autism as the data were insufficient"[166]. As with pertussis, these concerns were amplified by media outlets and perpetuated by organisations providing parental support[167].

Many of these themes have re-emerged during the COVID-19 pandemic with different communities with different concerns expressing hesitancy or reluctance to be vaccinated.

FIGURE 2

Survey results for the question: Overall, how safe, if at all, do you think each of the following COVID-19 vaccines are?



Source: Royal Society / YouGov, July 2021. (n=2,019)

[165] Wakefield's article linking MMR vaccine and autism was fraudulent. The British Medical Journal. 6 January 2011. See https://www.bmj.com/content/342/bmj.c7452 (accessed 4 November 2021).

[166] Murch *et al.* 2004 Retraction of an interpretation. The Lancet. 6 March 2004. See https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(04)15715-2/fulltext (accessed 4 November 2021).

[167] Cabrera Lalinde I. 2021 How misinformation affected the perception of vaccines in the 20th century based on the examples of polio, pertussis and MMR vaccines. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

To counter this, the world's major public health bodies, as well as specialists from academia, non-governmental organisations (NGOs), and international charities, have sought to create extensive pro-vaccine public engagement campaigns. While pro-vaccine sentiment largely falls to these larger organisations, there do exist a number of vocal pro-vaccine individuals whose passionate, information-seeking behaviour mirrors that of sections of the anti-vaccination actors[168].

Pro-vaccine actors' sharing of information is predominantly driven by the intention of attaining or maintaining a high level of vaccine uptake throughout the world. Major players in this area include the WHO, governments acting via public health bodies, and NGOs such as the Bill and Melinda Gates Foundation, acting through a united consensus. Online platforms have also taken steps to counter anti-vaccination content on their platforms with some removing discussion forums[169], banning hashtags[170], and labelling content[171].

These institutions, while being the fundamental drivers for vaccine engagement, may also be exacerbating the views of anti-vaccine groups in a form of psychological reactance (a tendency for people to react negatively when they feel their choices are being taken away)[172]. In these instances, the unified consensus reinforces a narrative that a pro-vaccine agenda is being driven by international elites seeking to impose their interests, and which should be resisted. Individuals who harbour any mistrust in modern medicine, political elites, mainstream media, or who hold a conspiracy theorist's mindset are especially prone to this psychological reactance. However, there are also individuals who may be introduced to this set of anti-establishment conspiracies with anti-vaccine material being their 'gateway drug' – the psychology of conspiracy theories shows that individuals that believe in one type of conspiracy are likely to believe in others.

Also contributing to this environment are people with a financial interest arising from an individual's appetite for information on vaccines. These range from wellness advocates to social media influencers using media platforms such as YouTube and Facebook to share content with their followers[173]. In the latter case, money is generated via advertising revenue, a share of which is kept by the online platforms. ▶

[168] Moxon E, Siegrist C. 2011 The next decade of vaccines: societal and scientific challenges. The Lancet. 9 June 2011. See https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(11)60407-8/fulltext (accessed 4 November 2021)

[169] Reddit bans COVID misinformation forum after 'go dark' protest. The Guardian. 1 September 2021. See https://www.theguardian.com/technology/2021/sep/01/reddit-communities-go-dark-in-protest-over-covid-misinformation (accessed 4 November 2021).

[170] Instagram blocks vaccine hoax hashtags. BBC News. 10 May 2019. See https://www.bbc.co.uk/news/technology-48227377 (accessed 4 November 2021).

[171] Twitter. 2021 Updates to our work on COVID-19 vaccine misinformation. See https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation (accessed 4 November 2021).

[172] Steindl C, Jonas E, Sittenthaler S, Traut-Mattausch E, Greenberg J. 2015 Understanding psychological reactance: New developments and findings. Zeitschrift fur Psychologie. 223, 205-214. (https://doi.org/10.1027/2151-2604/a000222).

[173] How the wellness and influencer crowd serve conspiracies to the masses. The Guardian. 25 February 2021. See https://www.theguardian.com/australia-news/2021/feb/25/how-the-wellness-and-influencer-crowd-served-conspiracies-to-the-masses (accessed 4 November 2021).

CASE STUDY 1 (continued)

There also exists a body of 'trolls' amplifying anti-vaccination material, who do not necessarily share a genuine belief in either side's cause, but rather seek to sow discord and polarisation within wider society. There are reports of some of these trolls being state-sponsored[174]. Adding to this artificial amplification are social media 'bots' sharing material and support for both sides.

What is the role of technology in the sharing of vaccine misinformation?

Online news providers act in a relatively restrained manner with regards to sharing or platforming of misinformation when compared to social media. While not generally actively promoting misinformation themselves they may still contribute to its spread via loosely moderated below-the-line comment functions or having editorial stances which challenge their readers' beliefs in experts, government, 'big pharma' or other elites – in turn providing fertile ground for vaccine misinformation to be spread elsewhere.

Social media platforms allow for the sharing of misinformation at unprecedented speed and scale. The amount of potential material for users to engage with sets a competition for 'eyeballs' resulting in an attention economy of likes, shares and commentary, in which misinformation offers highly shareable content[175]. Social media platforms optimise their services to secure high levels of user engagement, making use of sophisticated algorithms to help drive this engagement.

There have been examples of companies taking steps to reduce misinformation – Reddit removed specific sub-Reddits (forums or threads on specific subjects within the Reddit website) it deemed to be sharing harmful content, and Instagram banned a subset of anti-vaxx hashtags – but those seeking to promote such misinformation have been able to quickly adjust their tactics. Some social media platforms have also begun to use fact-checking services or introduce measures to verify content. However, these efforts are constrained by technological limitations and are often human-resource intensive, meaning they are unable to react to the rate at which content can be produced and cultural context can be subverted. Compounding this challenge, misinformation is increasingly shared via media which is difficult to regulate technologically. Images, GIFs (short motion images) and videos are harder for algorithmic detection, classification and contextualisation.

[174] GCHQ in cyberwar on anti-vaccine propaganda. The Times. 9 November 2020. See https://www.thetimes.co.uk/article/gchq-in-cyberwar-on-anti-vaccine-propaganda-mcjgjhmb2 (accessed 4 November 2021).

[175] Ryan C, Schaul A, Butner B, Swarthout J. 2020 Monetizing disinformation in the attention economy: The case of genetically modified organisms (GMOs). European Management Journal. 38, 7-18. (https://doi.org/10.1016/j.emj.2019.11.002).

In this context, it largely falls to individual users to apply their own judgement about which sources of information can be trusted, which leads to the challenge of how to empower citizens to have greater critical analysis skills and be more resilient to misinformation.

Policy and technological interventions

Regulating misinformation is not an insurmountable challenge. In the case of anti-vax YouTube channels identified by the Centre for Countering Digital Hate, for example, 409 active accounts have existed since 2018 or earlier, while measures to moderate or take down content or channels which frequently share misinformation within hours or days are both feasible and already put into practice in the case of examples like hate speech[176]. The technological barriers to keeping up with the speed at which culture, language and context evolve are complex, but becoming increasingly blurred. TikTok has demonstrated that by actively involving its content producers and consumers in its detection of successful trends, it is allowing content-providers to optimise their output, while simultaneously optimising its own algorithm and understanding of what draws users in.

Most countries already have in place legislation that regulates the advertising of products with unsubstantiated health benefits. There may be ways of applying this legislation against content which generates revenue from anti-vaccination information, in particular the examples where alternative medicine products are simultaneously marketed.

Digital platforms could be more effectively utilised by or collaborate with governments and public health agencies to help implement a wider variety of responses and share a diversity of information. The drier style of text-heavy, officious public communication by government and public health bodies does not currently compete with other more friendly, image-based and emotionally-led approaches adopted by those spreading misinformation. In addition, digital engagement with patients should be carried out, where possible, as public dialogue and provide space for discussion of any legitimate concerns on vaccines rather than countering all scepticism with ‘authoritative’ information. The challenge, therefore lies in presenting cogent, honest narratives tailored to the interests of individuals motivations, avoiding the risk of ’blowback’ or psychological reactance.

[176] Center for Countering Digital Hate. 2020 The Anti-Vaxx Industry See https://www.counterhate.com/anti-vaxx-industry (accessed 4 November 2021).

19 of 19
AFL
10cm
SP L10.4
SL 5.0
FoV 240*240
256*512s
Sag>Tra(5.7)>Cor(-3.9)

Chapter two
How the information environment is shaped and navigated

Left
© Rixipix.

How the information environment is shaped and navigated

How do our minds process information?

An understanding of how the online information environment may be affecting the way people produce and consume content can be gained from research undertaken in the emerging field of computational social and communication sciences, as well as literature from the cognitive and behavioural sciences.

An interdisciplinary literature review commissioned for this report[177] draws on this research and highlights the following three foundational theories of cognition which, when combined, form a framework to help explain how people engage with information:

1. Cognitive heuristics and biases
   Heuristics (mental shortcuts) are essential for fast decision-making in human judgement, but they can be systematically wrong and induce biases which deviate from rational behaviour[178]. Examples of common heuristics[179] include availability (judgements on probability based on events people can recall); anchoring and adjustment (judgements based around irrelevant information); and affect (judgements based on emotions and gut responses)[180].

2. Dual process theory
   The dual process theory assigns heuristics and other cognitive processes to two types of thinking, one intuitive and the other reflective[181]. Intuitive thinking requires minimal mental effort, relies on heuristics, and is activated in response to stimuli[182]. Reflective thinking is slower, more analytical, and applies hypothetical thinking[183]. Reflective thinking does not necessarily lead to better judgements as both processes can suffer from biases and yield incorrect conclusions.

3. Motivated reasoning
   The theory of motivated reasoning argues that desired conclusions play a role in determining which type of cognitive process (intuitive or reflective) is applied on a given occasion[184]. These motivations can be defined under the following categories:

- The defence motive, whereby individuals defend their attitudes, beliefs, or behaviours by avoiding, or engaging in a biased manner with, information likely to challenge them and favouring information likely to support them[185].

[177] Röttger P, Vedres B. The Information Environment and its Effects on Individuals and Groups. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[178] Tversky A, Kahneman D. 1974 Judgment under uncertainty: Heuristics and biases. Science. 185, 1124-1131. (https://doi.org/10.1126/science.185.4157.1124).

[179] Kahnemann D. 2011 Thinking Fast and Slow. London, UK: Penguin.

[180] Finucane M, Alhakami A, Slovic P, Johnson S. 2000 The Affect Heuristic in Judgements of Risks and Benefits. Journal of Behavioral Decision Making. 13, 1-17. (https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1%3C1::AID-BDM333%3E3.0.CO;2-S).

[181] Evans J, Stanovich K. 2013 Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science. 8, 223-241. (https://doi.org/10.1177%2F1745691612460685).

[182] Röttger P, Vedres B. The Information Environment and its Effects on Individuals and Groups. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[183] *Ibid.*

[184] Kunda Z. 1990 The case for motivated reasoning. Psychological Bulletin. 108, 480-498. (https://doi.org/10.1037/0033-2909.108.3.480).

[185] *Ibid.*

- The accuracy motive, whereby individuals engage with information in an objective, open-minded fashion to reach a normatively correct conclusion[186].
- The impression motive, whereby individuals engage with information to satisfy social goals[187].

Building upon this framework, a well-established phenomenon, termed confirmation bias, plays an influential role in cognitive outcomes. It represents the most relevant bias for understanding individual behaviour when faced with new information and refers to the tendency for individuals to be influenced by prior beliefs and expectations[188]. In general, individuals exhibit larger confirmation bias in settings which accentuate the defence motive over the accuracy motive[189].

Credibility presents an additional dimension for understanding how individuals engage with information. People tend to favour information sources they find more intuitively believable. Factors that shape perceptions of information credibility are evident through all components of information transmission channels, and include the following:

- **Author characteristics**
  Individuals are more likely to trust sources which they consider to be qualified, dynamic, and independent. In addition, positive testimonials can increase credibility as readers adopt a 'bandwagon' heuristic, where the number of prior endorsements positively relates to perceived source credibility[190]. New forms of trust simulations have emerged to accentuate this, such as reviews and kitemarks.

- **Familiarity and semantic quality in message content**
  Technical quality[191], grammatical correctness[192] and repeated exposure positively relate to credibility. Indicators of credibility are widespread in terminology online too, eg 'Ask the expert'.

In general, individuals exhibit larger confirmation bias in settings which accentuate the defence motive over the accuracy motive.

[186] Kruglanski A. 1989 The psychology of being "right": The problem of accuracy in social perception and cognition. Psychological Bulletin. 106, 395-409. (https://psycnet.apa.org/doi/10.1037/0033-2909.106.3.395).

[187] Chaiken S, Giner-Sorolla R, Chen S. 1996 Beyond accuracy: Defense and impression motives in heuristic and systematic information processing. In: P Gollwitzer & J Bargh (eds), The psychology of action: Linking cognition and motivation to behavior. The Guildford Press.

[188] Nickerson R. 1998 Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology. 2, 175-220. (https://doi.org/10.1037%2F1089-2680.2.2.175).

[189] Röttger P, Vedres B. The Information Environment and its Effects on Individuals and Groups. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[190] *Ibid*.

[191] Sundar S. 1999 Exploring receivers' criteria for perception of print and online news. Journalism & Mass Communication Quarterly. 76, 373-386. (https://doi.org/10.1177%2F107769909907600213).

[192] Maier S. 2005 Accuracy matters: A cross-market assessment of newspaper error and credibility. Journalism & Mass Communication Quarterly. 82, 533-551. (https://doi.org/10.1177%2F107769900508200304).

- **Platform characteristics and media attitudes**
  As the media landscape evolves, different channels are perceived as more or less trustworthy with levels of trust in online, print, and broadcast media fluctuating over time[193].

- **Audience attributes, beliefs, and attitudes**
  The individual attributes of the information consumer, particularly their ideological congeniality, can instil trust. Information which does not challenge an individual's attitudes and beliefs is considered to be more credible[194]. These existing beliefs can also lead to confirmation bias when consuming information.

Existing research on the cognitive processes involved in human judgement, reasoning and decision-making can help inform our understanding of the online information environment. For example, the format of online content may lend itself more to intuitive or reflective thinking processes; the public nature of social media platforms may accentuate the defence motive over the accuracy motive; and novel indicators of credibility may be emerging and influencing public attitudes. In addition, emerging research suggests that the internet may be limiting people's attentional capacities[195] which may in turn lead us to favour more intuitive thinking processes and place greater reliance on heuristics.

However, whilst useful for understanding how cognitive processes may affect behaviour, the evolving nature of the online information environment, heterogenous global media consumption patterns, and greater access to data may require these frameworks to be revised and updated in future.

Types of misinformation actors

The actors involved in producing and disseminating misinformation content can be broadly categorised as intentional or unintentional actors, and further differentiated by motivation. These actors can exist across all sections of society and often include those in positions of power and influence (eg political leaders, public figures, and media outlets). We identify four types of misinformation actors:

Good Samaritans

These users unknowingly produce and share misinformation content. Their motivation is to help others by sharing useful information which they believe to be true. Examples of this could include unknowingly sharing an ineffective health treatment[196] or an inaccurate election schedule[197].

[193] Röttger P, Vedres B. The Information Environment and its Effects on Individuals and Groups. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[194] *Ibid*.

[195] Firth *et al.* 2019 The "online brain": how the Internet may be changing our cognition. World Psychiatry. 18, 119-129. (https://doi.org/10.1002/wps.20617).

[196] Will lemons and hot water cure or prevent COVID-19? Snopes. 26 March 2020. See https://www.snopes.com/fact-check/lemons-coronavirus/ (accessed 4 November 2021).

[197] Fact check: Fake West Bengal election schedule circulates on social media. India Today. 17 February 2021. See https://www.indiatoday.in/fact-check/story/fact-check-fake-west-bengal-election-schedule-circulates-on-social-media-1770180-2021-02-17 (accessed 4 November 2021).

Profiteers

These users either knowingly share misinformation content or are ambivalent about the content's veracity. The consumption of their content generates profit for them[198] with greater engagement resulting in higher profit. Examples include writers for explicitly false news outlets being paid directly to a Google Ads account[199], companies selling fraudulent health treatments[200], and video content creators profiting from advertising revenue[201]. Profit, in this context, is not restricted to monetary value and can include other forms of personal gain (eg more votes or greater reach).

Coordinated influence operators

These users knowingly produce and share misinformation content. Their motivation is to sway public opinion in a manner that will benefit the agenda of their organisation, industry, or government. The aim is to either convince consumers of an alternate story or to undermine faith in trusted institutions. Examples include successfully publishing political opinion pieces by a fabricated expert in reputable online news outlets[202] and using automated social media accounts (bots) to promote climate change denialism[203].

Attention hackers

These users knowingly produce and share misinformation content. Their motivation is personal joy. Sometimes referred to as 'trolling', these users devise outlandish or divisive content and take steps to maximise attention for them. Examples include sending messages to mainstream talk shows in the hope they will read out the content on air[204], fooling high profile figures into resharing content on their social media accounts[205], and sharing conspiracy theories on unsuspecting television and radio phone-ins (known as groyping)[206].

[198] Bakir V, McStay A. 2018 Fake News and the Economy of Emotions. Digital Journalism. 6, 154-175.

[199] How the "King of Fake News" built his empire. The Hustle. 7 November 2017. See https://thehustle.co/fake-news-jestin-coler/ (accessed 4 November 2021).

[200] European Medicines Agency. 2020 COVID-19: Beware of falsified medicines from unregistered websites. See https://www.ema.europa.eu/en/news/covid-19-beware-falsified-medicines-unregistered-websites (accessed 4 November 2021).

[201] YouTube pulls ads from anti-vax conspiracy videos. The Verge. 22 February 2019. See https://www.theverge.com/2019/2/22/18236839/youtube-demonetization-anti-vaccination-conspiracy-videos-dangerous-harmful-content (accessed 4 November 2021).

[202] Deepfake used to attack activist couple shows new disinformation frontier. Reuters. 15 July 2020. See https://www.reuters.com/article/us-cyber-deepfake-activist-idUSKCN24G15E (accessed 4 November 2021).

[203] Marlow T, Miller S, Roberts J. 2020 Bots and online climate discourses: Twitter discourse on President Trump's announcement of U.S. withdrawal from the Paris Climate Agreement. Climate Policy. 21, 765-777. (https://doi.org/10.1080/14693062.2020.1870098).

[204] Phillips W. 2012 The House That Fox Built: Anonymous, Spectacle, and Cycles of Amplification. Television & New Media. 14, 494-509. (https://doi.org/10.1177%2F1527476412452799).

[205] Donald Trump retweets serial killer photo in comedian's Twitter prank. The Guardian. 29 September 2014. See https://www.theguardian.com/media/2014/sep/29/donald-trump-retweets-serial-killer-photos-in-comedians-twitter-prank (accessed 4 November 2021).

[206] The far-right plot to flood radio airwaves with racism. Vice. 18 June 2020. See https://www.vice.com/en/article/z3exp3/groyping-far-right-propaganda-tool-alt-right (accessed 4 November 2021).

These incentives can determine the veracity of a piece of content, the price for accessing it, and the experience of consuming it.

Incentives for information production and consumption

The role of incentives for content production and consumption is important to consider when examining how the online information environment operates. The overarching incentives can be categorised as content for public benefit (eg to improve the health of a population or raise awareness of the plight of others) or as content for private benefit (eg to maximise advertising revenue, shareholder value, or self-satisfaction). The NHS, BBC, and Wikipedia would be examples of organisations who produce content for public benefit. Examples of those who produce content for private benefit include newspapers, academic journals, and social media influencers.

Content producers can also fall into both categories. For example, a Wikipedian may edit articles in order to better inform others (public benefit) and to raise their own reputation within the Wikipedia community (private benefit). Similarly, an academic publisher may produce content to advance collective understanding (public benefit) and to maximise profit for shareholders (private benefit). Whilst predating the internet, these public and private incentives play a major role in shaping how and why we access, produce, and engage with content online. These incentives can determine the veracity of a piece of content, the price for accessing it, and the experience of consuming it. Coupled with the attention economy concept – that information seeks and competes for attention – they have created an online information environment dominated by pay-per-click advertising[207], incongruent headlines[208], and addictive user interfaces[209].

These incentives occur on both a macro and micro level. On a macro level, for-profit organisations produce engaging and relevant content in order to maximise online revenue (eg advertising revenue, product sales). Not-for-profit organisations produce content to inform and educate users irrespective of how many views the content receives. On a micro level, an individual user may produce content in order to maximise the amount of satisfaction (or dopamine[210]) they receive as a result of a post's engagement or they may produce content in order to raise awareness of an issue or to help their loved ones.

The public benefit incentive has encouraged the development of open data repositories[211], application programming interfaces[212], and disability-friendly user experience design[213]. Furthermore, the incentive encourages accuracy and expert opinion which can be gained via trained journalists, medical professionals, and scientific authorities.

[207] Kapoor K, Dwivedi Y, and Piercy N. 2016 Pay-Per-Click Advertising: A Literature Review. The Marketing Review. 16, 183-202. (http://dx.doi.org/10.1362/146934716X14636478977557).

[208] Chesney S, Liakata M, Poesio M, and Purver M. 2017 Incongruent headlines: Yet another way to mislead your readers. Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism. (http://dx.doi.org/10.18653/v1/W17-4210).

[209] Social media apps are 'deliberately' addictive to users. BBC News. 4 July 2018. See https://www.bbc.co.uk/news/technology-44640959 (accessed 4 November 2021).

[210] Dopamine is a type of neurotransmitter. In popular culture, it is referred to as the chemical linked to pleasure.

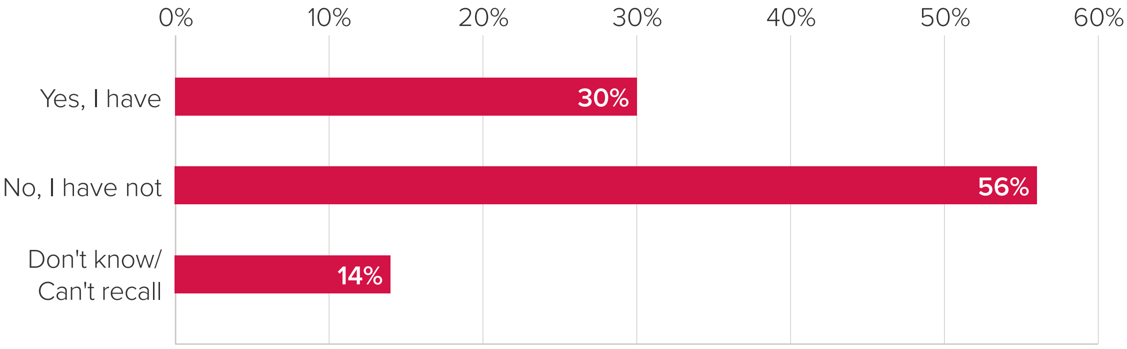
[211] HM Government. Find Open Data. See https://data.gov.uk/ (accessed 4 November 2021).

[212] Government Digital Service. API Catalogue. See https://www.api.gov.uk/gds/#government-digital-service (accessed 4 November 2021).

[213] HM Government. Accessibility – GOV.UK Design System. See https://design-system.service.gov.uk/accessibility/ (accessed 4 November 2021).

FIGURE 3

Survey results for the question: Have you ever personally shared content online, via any method, about any scientific developments (eg news articles, tweets, posts on social media, clips/videos etc)?



Source: Royal Society / YouGov, July 2021. (n= 2,019)

The private benefit incentive has led to the development of pay-per-click advertising, paywalls, and more engaging or addictive user experience design (eg infinite scrolling[214, 215]). The incentive encourages engagement which can be gained via high quality content, search engine optimisation, and clickbait[216].

How these incentives affect users' consumption and dissemination of information online can, to some extent, be controlled by search engines and social media platforms[217]. Search engines can alter how their ranking algorithms prioritise content in the results page (eg sorting by location, relevance, and credibility)[218]. Social media and video hosting platforms can control their rewards system and sharing mechanisms (eg amplifying engaging content[219], demonetising problem content[220], or prompting users to read articles before sharing[221]).

[214] No More Pages? Humanized. April 2006. Available on Web Archive.

[215] Reading, Humanized. Humanized. April 2006. Available on Web Archive.

[216] Zannettou S, Chatzis S, Papadamou K, Sirivianos M. 2018 The Good, the Bad and the Bait: Detecting and Characterizing Clickbait. 2018 IEEE Security and Privacy Workshops. (https://doi.org/10.1109/SPW.2018.00018).

[217] Royal Society roundtable with Major Technology Organisations, March 2021.

[218] Google Search Central. Advanced SEO. See https://developers.google.com/search/docs/advanced/guidelines/get-started (accessed on 4 November 2021).

[219] Five points for anger, one for 'like': How Facebook's formula fostered rage and misinformation. The Washington Post. 26 October 2021. See https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/ (accessed on 4 November 2021).

[220] YouTube Help. Advertiser-friendly content guidelines. See https://support.google.com/youtube/answer/6162278?hl=en-GB (accessed on 4 November 2021).

[221] Twitter is bringing its 'read before you retweet' prompt to all users. The Verge. 25 September 2020. See https://www.theverge.com/2020/9/25/21455635/twitter-read-before-you-tweet-article-prompt-rolling-out-globally-soon (accessed 4 November 2021).

FIGURE 4

Survey results to the question: You previously said you have shared content online about a scientific development(s). Thinking about the most recent time you did this… what were your reasons for sharing content online about scientific developments?

To counter misinformation about Covid-19 vaccines.

To share information with friends and family.

Last month I sent my grandson a link to BBC *Science Focus* magazine to help with a school project.

To create awareness.

I was surprised by what I found and wanted a friend's opinion.

My daughter is studying Marine Biology and if I see any news articles that I think would interest her I share them.

The information given by scientists from the Barrington declaration. It wasn't being shared honestly by the main stream media.

Source: Royal Society / YouGov, July 2021. (n=613)

FIGURE 5

Examples of how incentives can shape the production and consumption of online health information.

	MICRO-LEVEL		MACRO-LEVEL	
INCENTIVE	Production	Consumption	Production	Consumption
Public benefit	Individual produces content to raise awareness of the negative side-effects of a medical treatment and protect their loved ones.	Individual consumes content in order to learn how to treat the symptoms of a loved one suffering from an illness.	Public health authority produces content to encourage optimal hand-washing techniques to reduce the impact of a viral disease.	Public health authority consumes social media content to identify gaps in the public's understanding of a disease.
Private benefit	Medical professional produces content as a means of socialising with others and fostering new relationships.	Medical student consumes content in order to improve chances of passing an upcoming exam.	Wellness company produces engaging content to showcase benefits of their products and increase sales.	Insurance company consumes social media content to help amend health premiums.

How the internet facilitates access to information

The online information environment differentiates itself from the offline environment in its breadth of reach and its mass of users. In the current environment, any internet user can publish content to be accessed, read, and processed by other internet users regardless of location. This has provided significant opportunities for the improvement of scientific understanding. We identify the following three internet-enabled innovations as being especially transformative for the collective understanding of scientific issues:

Hypertext

Hypertext – online text which redirects (links) users to other content when clicked – has enabled fast and simple referencing in a manner which was not feasible prior to the advent of the online information environment[222]. These links (known as hyperlinks) are omnipresent in the online information environment, connecting content together and providing a navigation route across the internet.

The sharing of hyperlinks – a common behaviour, exclusive to the online information environment – allows users to refer others to external sources. This often occurs in online literature, social media posts, and direct messaging platforms. They provide a simple mechanism for adding credibility to an argument made online (eg an individual is more likely to convince someone to follow their medical advice if they can link them to an official webpage from a respected health authority confirming it).

[222] De Maeyer J. 2012 Towards a hyperlinked society: A critical review of link studies. New Media & Society. 15, 737-751. (https://doi.org/10.1177%2F1461444812462851).

Wikis

A wiki is a type of website which allows users to collaboratively edit its content. Wikis are often put together using a hypertext structure and pages do not have defined owners. They can be both public and private (requiring membership to view and edit). Prominent examples of wikis include the online encyclopaedia, Wikipedia, and the whistleblowing website, WikiLeaks[223].

Unlike static analogue alternatives, wikis enable information about a subject to be updated regularly and transparently. These updates can be commented on, scrutinised, and reverted by other users. On Wikipedia – which contains 56 million articles and 1.6 billion unique visitors a month – the most contentious subjects (eg COVID-19) are heavily scrutinised, with thousands of users reviewing edits[224].

Open science

The internet has enabled a movement towards open access practices in which scientific research papers are publicly and freely available to download. It represents a shift from the traditional publishing model of papers being published behind a paywall accessible to those with the financial means to do so (through attending an institution with a journal subscription or by paying for individual articles)[225].

This movement has included the rise of preprints. These are versions of academic papers published online prior to receiving formal peer review. They are hosted on preprint servers and allow for rapid dissemination and scrutiny of provisional research findings.

Open access journals divide opinion in the academic community with some (including the Royal Society) arguing that they enable the 'widest possible dissemination of research outputs'[226] and others arguing that it leads to lower quality research being published[227]. During the COVID-19 pandemic, however, many research outlets committed to open access publication of their research related to the disease – showcasing the value of open access research for emergencies[228].

[223] WikiLeaks was originally founded as a wiki but transitioned away from this in 2010. WikiLeaks gets a facelift. Mother Jones. 19 May 2010. See https://www.motherjones.com/politics/2010/05/wikileaks-assange-returns/ (accessed 4 November 2021).

[224] Royal Society roundtable with Major Technology Organisations, March 2021.

[225] The Royal Society. 2012 Science as an open enterprise. See https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf (accessed 4 November 2021).

[226] The Royal Society. Open access publishing. See https://royalsociety.org/journals/authors/which-journal/open-access/ (accessed 4 November 2021).

[227] van Vlokhoven, H. 2019 The effect of open access on research quality. Journal of Informetrics. 13, 751-756. (https://doi.org/10.1016/j.joi.2019.04.001).

[228] Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. Wellcome Trust. 31 January 2020. See https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak (accessed 4 November 2021).

At the same time, the limited availability to legitimate open access research has led to a growth in popularity of pirate websites such as Sci-Hub which provide free downloads of academic papers, access to which is often gained through illegal means such as phishing[229].

All three of these innovations are underpinned by their capabilities for sharing, collaboration, and scrutiny.

Policies adopted by major online platforms

Online platforms for user-generated content have developed a range of activities and policies over the past few years to combat the spread of misinformation, including deceptive synthetic media (eg deepfakes). Many online platforms have specific policies covering elections and COVID-19 that are more specific and far-reaching than their approaches to misinformation in general but three key themes remain central to misinformation management on social media platforms: misinformation detection, limiting the spread of misinformation, and improving digital literacy.

These initiatives, whilst important to note, are rarely subjected to transparent, independent assessments meaning that it is difficult to assess and verify their impact.

Misinformation detection

To counter misinformation, combinations of in-house automated technology, third-party fact-checking and human reviews are employed by social media platforms.

- Automated monitoring technologies are used by many of the platforms to detect and remove policy-violating content. In instances where uncertainty and ambiguity surround content, further human reviews are conducted. This manual labelling helps to improve the quality of the platform's AI systems.
- Third-party fact-checkers review content independently of the social media companies to decide upon rating options.
- Community-driven approaches are seen across several platforms to tackle the issue of misinformation. For example, Reddit has minimal site-wide policies, with communities ('subreddits') self-authoring and self-policing further policies on top of these. Meanwhile, Twitter is piloting a programme called 'Birdwatch', which allows users to identify information they believe to be misleading[230].

[229] Himmelstein *et al.* 2018 Sci-Hub provides access to nearly all scholarly literature. Meta-Research: A Collection of Articles. (https://doi.org/10.7554/eLife.32822).

[230] Coleman K. 2021 Introducing Birdwatch, a community-based approach to misinformation. Twitter. 25 January 2021. See https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation (accessed 4 November 2021).

Limiting the spread of misinformation

There is wariness amongst online platforms about the removal of misinformation content[231]. Concerns for public perception and safety are played off against concerns surrounding the free expression of opinions.

Given these issues, platforms have been more willing to limit the spread of misinformation through methods such as minimising the prominence of such items in newsfeeds and adding warning labels, than they have been to remove items of misinformation outright[232]. Other tools include limiting the number of times a message can be forwarded and redirecting searches prone to misinformation towards authoritative sources. Addressing misinformation within advertising policies is another important angle for platforms.

Improving digital literacy

Accessible information relating to an online platform's terms of service is critical for educating users on appropriate online behaviour and is emphasised in the UK Government's strategy for combating online harms[233]. To deliver greater transparency and accountability, mandatory transparency reporting is proposed to be introduced. Platforms are increasingly disclosing top-line data in the form of these transparency reports on their efforts to tackle misinformation and empower users to make informed choices online.

Informational labels on posts with links to trusted websites are increasingly common across platforms. In addition to this, platforms are increasingly developing trustworthy information hubs on their platforms. For example, Facebook have developed hubs with credible science-based information from trusted experts and promote news literacy about COVID-19 and climate science[234]. Similarly, Twitter has integrated a specific tab in users' feeds for authoritative content about COVID-19[235]. Video-sharing platforms such as TikTok and YouTube have been working with content creators to generate authoritative and engaging resources to combat misinformation about vaccines[236. 237].

[231] Royal Society roundtable with Major Technology Organisations, March 2021.

[232] Full Fact. 2020 Fighting the causes and consequences of bad information. See https://fullfact.org/blog/2020/apr/full-fact-report-2020/ (accessed 4 November 2021).

[233] HM Government. 2021 Draft Online Safety Bill. See https://www.gov.uk/government/publications/draft-online-safety-bill (accessed 4 November 2021).

[234] Connecting people with credible climate change information. Facebook. 18 February 2021. See https://about.fb.com/news/2021/02/connecting-people-with-credible-climate-change-information/ (accessed 4 November 2021).

[235] Coronavirus: Staying safe and informed on Twitter. Twitter. 12 January 2021. See https://blog.twitter.com/en_us/topics/company/2020/covid-19 (accessed 4 November 2021).

[236] Morgan K. 2020 Taking action against COVID-19 vaccine misinformation. TikTok. 15 December 2020. See https://newsroom.tiktok.com/en-gb/taking-action-against-covid-19-vaccine-misinformation (accessed 4 November 2021).

[237] Graham G. 2021 New health content is coming to YouTube. YouTube Official Blog. 13 January 2021. See https://blog.youtube/news-and-events/new-health-content-coming-youtube/ (accessed 4 November 2021).

CASE STUDY 2

5G misinformation.

Misinformation about 5G telecommunications technology has attracted significant concern in recent years after baseless conspiracy theories linking it to the spread of COVID-19[238] went viral, exacerbating prior concerns about the technology held by groups such as the International Electromagnetic Fields Scientist Appeal[239]. In April 2020, the UK Government and trade body, Mobile UK, issued statements condemning the theories after phone masts were set on fire in multiple locations across the country[240]. The CEO of British Telecom reported that there had been 40 incidents of attacks on their staff, including one of their engineers being stabbed and hospitalised[241].

The World Health Organization[242] and the International Commission on Non-Ionizing Radiation Protection[243] have stated that no adverse health effects have been causally linked with exposure to wireless technologies and that no consequences for public health are anticipated provided overall exposure remains below international guidelines. To explore the spread of misinformation about telecommunications, the Royal Society held a roundtable on the topic with experts from academia, industry, and government. The key themes from this discussion are outlined below.

A range of different groups are involved in creating and spreading misinformation around 5G. One of the distinguishing aspects of 5G information from other topics, such as climate change, is that while the scientific advice is predominantly distributed at a global level, standards and regulation are often set nationally or locally. Meanwhile, the variety of government departments and private sector organisations involved drives confusion among some in the public over who is responsible for what, and which voices are trustworthy, especially as press coverage has tended to associate 5G with negative language concerning government intent[244]. ▶

[238] Temperton J. 2020 How the 5G coronavirus conspiracy theory tore through the internet. Wired. 6 April 2020. See https://www.wired.co.uk/article/5g-coronavirus-conspiracy-theory (accessed 4 November 2021).

[239] Kelley E, Blank M, Lai H, Moskowitz J, Havad M. 2015 International Appeal: Scientists call for protection from non-ionizing electromagnetic field exposure. European Journal of Oncology. 20, 180-182.

[240] Mast fire probe amid 5G coronavirus claims. BBC News 4 April 2020. See https://www.bbc.co.uk/news/uk-england-52164358 (accessed 4 November 2021).

[241] 77 cell phone towers have been set on fire so far due to a weird coronavirus 5G conspiracy theory. Business Insider. 6 May 2020. See https://www.businessinsider.com/77-phone-masts-fire-coronavirus-5g-conspiracy-theory-2020-5?r=US&IR=T (accessed 4 November 2021).

[242] World Health Organization. Radiation: 5G mobile networks and health. See https://www.who.int/news-room/q-a-detail/radiation-5g-mobile-networks-and-health (accessed 4 November 2021).

[243] International Commission on Non-Ionizing Radiation Protection. 5G Radiofrequency – RF EMF. See https://www.icnirp.org/en/applications/5g/5g.html (accessed 4 November 2021).

[244] Mansell R, Plantin J. 2020 Urban futures with 5G: British press reporting. London School of Economics. See http://eprints.lse.ac.uk/id/eprint/105801 (accessed 4 November 2021).

CASE STUDY 2 (continued)

Generally, people trust public bodies on 5G, with a marked tendency to trust more local voices, with local government significantly more trusted than national government[245]. However, according to the survey commissioned for this report, a significant minority, 15%, believe the technology is harmful to human health[246].

Genuine uncertainty also drives much public misunderstanding around 5G. A distinguishing feature of the topic is that among scientific actors involved in discussion of 5G, there are genuine disagreements about the balance of risk versus reward, predominantly stemming not from a disagreement about the underlying science, but different attitudes to risk assessment. The difficulty of communicating such uncertainty to the public and key decision-makers is exacerbated by the fact that terminology used within the scientific community, such as how 'safe' or 'proven safe' might be used, maps poorly onto how those terms are used in wider society. Furthermore, there are historic cases where the safety of technologies have been overstated.

Distrust of telecoms technologies generally has been far less prevalent where the direct personal benefits of technologies were more evident. This is reflected in the fact that concerns are far more frequently ascribed to base stations than personal handsets. Consent is an important factor in this. During the initial wave of uptake of mobile handsets, there was markedly greater concern among those required to carry them for work purposes.

There is also significance to the ways that exploration of potential harms can be perceived. For example, in the 1990s the UK Department for Trade and Industry put out a research contract on how to measure electromagnetic frequencies close to the head, to assess whether there was a risk of brain cancer from mobile phone use. The research was reported by the media, and the very fact that the research was being carried out led to the potential risk becoming the focal point of a health scare[247].

[245] European 5G survey: Europeans are positive, but disinformation looms and citizens call for action. Ipsos MORI. 16 October 2020. See https://www.ipsos.com/en/european-5g-survey-2020 (accessed 4 November 2021).

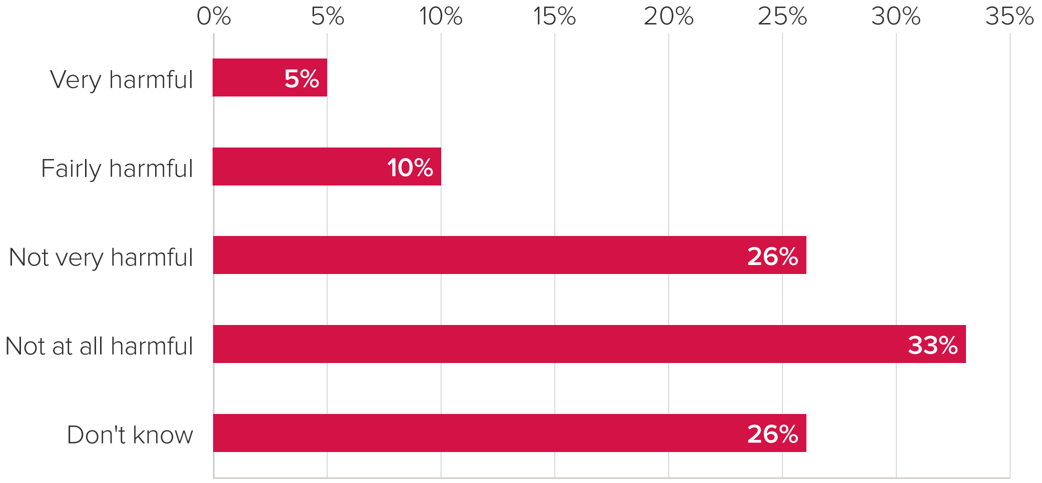
[246] 5% believe 5G technology is very harmful to human health, however a further 10% believe it is fairly harmful to human health. Royal Society / YouGov, July 2021.

[247] Stilgoe J. 2007 The (co-)production of public uncertainty: UK scientific advice on mobile phone health risks. Public understanding of science (Bristol, England). 16, 45-61. (https://doi.org/10.1177/0963662506059262).

An additional challenge is that speculative health concerns lie low on the list of priorities both of health bodies, such as Public Health England, where there are more pressing health concerns with a greater case for receiving research funding, and within architects of the digital economy, where the pressing needs for wider spectrum bands in use outweighs the investigation of unsubstantiated claims. Even where spectrum bands were considered, but ultimately rejected, as with the initial band for 5G in Europe, health concerns specific to that band attached to the label of '5G' persisted and are still found in 5G misinformation even where it uses a completely different spectrum band[248].

FIGURE 6

Survey results for the question: In general, how harmful, if at all, do you think 5G technology is to human physical and/or mental health?



Source: Royal Society / YouGov, July 2021. (n=2,019)

[248] European Parliamentary Research Services. 2021 Health impact of 5G. See https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2021)690012 (accessed 4 November 2021).

Chapter three
Techniques for creating and countering misinformation

Left
D. Fairy offering flowers to Iris, one of a series of photographs depicting the Cottingley Fairies, taken by Elsie Wright and by Frances Griffith, in 1920. CC BY-NC-ND 4.0 www.collection.sciencemuseumgroup.org.uk

Techniques for creating and countering misinformation

Synthetic content

Synthetic digital and online content has caused some concern in public conversations about the online information environment, defined by the US Federal Bureau of Investigation as 'the broad spectrum of generated or manipulated digital content, which includes images, video, audio, and text'[249]. Although manipulation of this kind was possible prior to the advent of the internet, it has become increasingly easier to achieve with modern techniques. Synthetic content has both positive and negative applications in the online information environment. We identify the following eight major types of synthetic content:

Bots

Bots are pre-programmed online accounts that engage with and respond to online content in an automated fashion. They take many forms in the online information environment. Chatbots can act as customer service operators for large companies (eg retail banks) or as false personas on social media platforms. Voice assistants recognise and respond verbally to spoken requests made by a user to a smart device. Bot crawlers undertake basic administrative tasks for owners such as indexing web pages or countering minor acts of vandalism on wikis[250]. Traffic bots exist to inflate the number of views for a piece of online content to boost revenue derived from online advertising[251].

Positive applications of bots include their use to counter misinformation[252], to support people with disabilities[253], and to disseminate news updates[254]. Negative applications include the use of bots to deceptively influence public opinion[255], to suppress news stories[256], and to abuse people[257].

[249] Federal Bureau of Investigation, Cyber Division. 2021 Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations. See https://www.ic3.gov/Media/News/2021/210310-2.pdf (accessed 4 November 2021).

[250] Royal Society roundtable with Major Technology Organisations, March 2021.

[251] Zeifman I. 2014 Bot Traffic Report: Just the droids you were looking for. Imperva. 18 December 2014. See https://www.imperva.com/blog/bot-traffic-report-2014/ (accessed 4 November 2021).

[252] Carmi E, Musi E. 2021 How to empower citizens to fight fake news during Covid-19. University of Liverpool. See https://www.liverpool.ac.uk/coronavirus/blog/februaryposts/how-to-fight-fake-news-during-covid/ (accessed 4 November 2021).

[253] Pradhan A, Mehta K, Findlater L. 2018 "Accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 459, 1-13. (https://doi.org/10.1145/3173574.3174033).

[254] Lokot T, Diakopoulos N. 2015 News bots: Automating news and information dissemination on Twitter. Digital Journalism. 4, 682-699. (https://doi.org/10.1080/21670811.2015.1081822).

[255] Shao *et al.* 2018 The spread of low-credibility content by social bots. Nature Communications. 9, 4787. (https://doi.org/10.1038/s41467-018-06930-7).

[256] Diakopoulos N, 2018 The bots beat: How not to get punked by automation. Columbia Journalism Review. 3 April 2018. See https://www.cjr.org/tow_center/bots-manipulate-trends.php (accessed 4 November 2021).

[257] Daniel F, Cappiello C, Benatallah, B. 2019 Bots acting like humans: Understanding and preventing harm. IEEE Internet Computing. 23, 40-49. (https://doi.org/10.1109/MIC.2019.2893137).

Text

Text can be subjected to various forms of manipulation. A common example of text being manipulated is the creation and dissemination of fabricated social media posts. This involves an individual using image-editing techniques (which can involve mainstream word processors rather than sophisticated image-editing software) to erase the original text from a target's social media post and replace it with new text[258]. This is then screenshotted and disseminated as 'evidence' of the target taking a false stance on an issue.

Another form of manipulation involves creating bots which are capable of natural language generation. These bots can understand the contents of text and automatically generate replies which appears as natural as a reply generated by a human. This technique is often used to create false social media personas and has also been used to generate false public policy consultation responses[259].

Images

Editing images to generate misinformation content dates back more than a century. Notable examples include the 1901 image 'Trick photograph of man with two heads'[260], the 1902 painting of General Ulysses S Grant[261], and the Soviet Union 'Great Purge' photographs from the 1920s and 1930s[262]. This practice, now broadly referred to as 'photoshopping', can be done via an increasingly wide range of accessible platforms and applications, and therefore by a growing number of internet users. These techniques can be used to create false images[263] or fabricated quotations[264].

Image editing techniques can also be used to create internet memes, especially the kind of humorous content that is shared rapidly and becomes viral. They usually consist of captions superimposed onto an image. Researchers have found that memes helped spread a conspiracy theory linking COVID-19 to 5G technology[265].

[258] Fake Jeremy Corbyn tweet spreads after the London Bridge attack. Full Fact. 30 November 2019. See https://fullfact.org/online/fake-jeremy-corbyn-did-not-tweet-london-bridge-attacker-was-murdered-police/ (accessed 4 November 2021).

[259] New York State Office of the Attorney General Letitia James. 2021 Fake Comments: How U.S. Companies & Partisans Hack Democracy to Undermine Your Voice. See https://ag.ny.gov/sites/default/files/oag-fakecommentsreport.pdf (accessed 4 November 2021).

[260] Library of Congress. Trick photograph of man with two heads. See https://www.loc.gov/resource/cph.3a15713/ (accessed 4 November 2021).

[261] Library of Congress. Civil War Glass Negatives and Related Prints. See https://www.loc.gov/pictures/collection/cwp/ (accessed 4 November 2021).

[262] Blakemore E. 2018 How photos became a weapon in Stalin's Great Purge. The History Channel. 20 April 2018. See https://www.history.com/news/josef-stalin-great-purge-photo-retouching (accessed 4 November 2021).

[263] Zhao B, Zhang S, Xu C, Sun Y, Deng C. 2021 Deep fake geography? When geospatial data encounter artificial intelligence. Cartography and Geographic Information Science. 48, 338-351. (https://doi.org/10.1080/15230406.2021.1910075).

[264] Fact check: Trump did not call Republicans "the dumbest group of voters". Reuters. 28 May 2020. See https://www.reuters.com/article/uk-factcheck-trump-republicans-meme-idUSKBN2342S5 (accessed 4 November 2021).

[265] Fischer S, Snyder A. How memes became a major vehicle for misinformation. Axios. 23 February 2021. See https://www.axios.com/memes-misinformation-coronavirus-56-2c3e88be-237e-49c1-ab9d-e5cf4d2283ff.html (accessed 4 November 2021).

Miscontextualised content

Genuine, unedited content can be shared without context to provide a misleading narrative. This is made easier in the online information environment as content can be disseminated between people without intermediaries (eg news outlets, government officials). This has been referred to as 'malinformation'[266]. Examples include sharing real images and claiming that they represent something that they do not[267]. They can also involve sharing images of different events from a different date to create a false narrative and discredit targets[268].

Shallowfakes

A form of malinformation content, shallowfakes refer to videos which have been presented out of context or crudely edited[269]. Examples of shallowfakes have included videos which have been edited to portray intoxication[270], an act of aggression[271], and a false interview response[272]. These effects are achieved by using video-editing software or smartphone applications to change the speed of video segments or crop together clips in order to omit relevant context.

Deepfakes

Originating from a Reddit user[273] who shared edited videos of celebrity faces swapped into pornographic videos, deepfakes refer to novel audio and/or visual content generated using artificial intelligence techniques such as generative adversarial networks[274] (GANs). GANs involve two neural networks competing against each other – one creating false content and the other trying to detect it. The GANs can be trained using images, sounds, and videos of the target. The result is convincingly edited 'new' audio and/or visual content.

[266] Council of Europe. 2017 Information Disorder: Toward an interdisciplinary framework for research and policy making. See https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c (accessed 4 November 2021).

[267] White Helmets 'staging fake attacks' in Syria? We sort fact from fiction. France 24. 14 May 2018. See https://observers.france24.com/en/20180514-white-helmets-allegations-fact-fiction (accessed 4 November 2021).

[268] Fazio L. 2020 Out-of-context photos are a powerful low-tech form of misinformation. The Conversation. 14 February 2020. See https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959 (accessed 4 November 2021).

[269] Johnson B. 2019 Deepfakes are solvable – but don't forget that "shallowfakes" are already pervasive. MIT Technology Review. March 2019. See https://www.technologyreview.com/2019/03/25/136460/deepfakes-shallowfakes-human-rights/ (accessed 4 November 2021).

[270] No, this is not a real video of US House Speaker Pelosi slurring her words onstage – the footage has been doctored to slow down her speech. AFP Fact Check. 24 May 2019. See https://factcheck.afp.com/no-not-real-video-us-house-speaker-pelosi-slurring-her-words-onstage-footage-has-been-doctored-slow (accessed 4 November 2021).

[271] Bauder D, Woodward C. 2018 Expert: Acosta video distributed by White House was doctored. Associated Press. 9 November 2018. See https://apnews.com/article/entertainment-north-america-donald-trump-us-news-ap-top-news-c575bd1cc3b1456cb3057ef670c7fe2a (accessed 4 November 2021).

[272] This video has been edited to make Keir Starmer appear confused at the end. Full Fact. 7 November 2019. See https://fullfact.org/online/keir-starmer-gmb-facebook/ (accessed 4 November 2021).

[273] Cole S. 2017 AI-assisted fake porn is here. Motherboard. 11 December 2017. See https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn (accessed 4 November 2021).

[274] Goodfellow *et al.* 2014 Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. 2, 2672-2680.

Deepfakes can involve portraying individuals doing or saying things which they never did or said. They can also involve the generation of a new 'person' – a still image of a novel face to be used in the creation of a fabricated online persona[275]. Research has found that the majority of the thousands of deepfakes currently in existence are of a pornographic nature[276], however other examples have included deepfakes of politicians[277], campaigners[278], celebrities[279], and the Queen[280].

Fake news websites

Websites which are dedicated to publishing false content or undertake little to no verification of the facts in their content can be significant sources of misinformation content in the online information environment. The goal of these websites is to generate money through advertising revenue with some websites paying writers directly into a Google AdSense account[281]. For some people, earning money from fake news websites is a living[282]. These websites often portray themselves as legitimate sources of information with some replicating the format of local news websites, sometimes naming themselves after towns and cities[283].

Inauthentic groups, pages, and accounts

Misinformation content can be spread through the creation or acquisition of social media groups, pages, and accounts. This can occur in an automated or manual fashion. Examples of automated inauthentic accounts include YouTube pages which generate video content about popular topics in order to increase subscribers before switching to share misinformation content[284]. Examples of manual techniques include purchasing admin rights for existing pages in order to disseminate misinformation content and creating accounts to use as a mechanism for boosting content[285, 286].

> For some people, earning money from fake news websites is a living.

275. This Person Does Not Exist. See https://thispersondoesnotexist.com/ (accessed 4 November 2021).

276. Deeptrace. 2019 The state of deepfakes: Landscape, threats, and impact. See https://regmedia.co.uk/2019/10/08/deepfake_report.pdf (accessed 4 November 2021).

277. The fake video where Johnson and Corbyn endorse each other. BBC News. 12 November 2019. See https://www.bbc.co.uk/news/av/technology-50381728 (accessed 4 November 2021).

278. Deepfake Greta Thunberg to become Channel 4 TikTok star for Earth Day. Channel 4. 22 April 2021. See https://www.channel4.com/press/news/deepfake-greta-thunberg-become-channel-4-tiktok-star-earth-day (accessed 4 November 2021).

279. Harrison E. 2021 Shockingly realistic Tom Cruise deepfakes go viral on TikTok. The Independent. 26 February 2021. See https://www.independent.co.uk/arts-entertainment/films/news/tom-cruise-deepfake-tiktok-video-b1808000.html (accessed 4 November 2021).

280. Deepfake Queen to deliver Channel 4's Alternative Christmas Message. Channel 4. 24 December 2020. See https://www.channel4.com/press/news/deepfake-queen-deliver-channel-4s-alternative-christmas-message (accessed 4 November 2021).

281. How the "King of Fake News" built his empire. The Hustle. 7 November 2017. See https://thehustle.co/fake-news-jestin-coler/ (accessed 4 November 2021).

282. Miller C. 2018 Meeting Kosovo's clickbait merchants. BBC News. 10 November 2018. See https://www.bbc.co.uk/news/technology-46136513 (accessed 4 November 2021).

283. Silverman C. 2020 These fake local news sites have confused people for years. We found out who created them. BuzzFeed News. 6 February 2020. See https://www.buzzfeednews.com/article/craigsilverman/these-fake-local-news-sites-have-confused-people-for-years (accessed 4 November 2021).

284. Graphika. 2021 Spamouflage Breakout. See https://graphika.com/reports/spamouflage-breakout/ (accessed 4 November 2021).

285. Meta. October 2020 Coordinated Inauthentic Behavior Report. See https://about.fb.com/news/2020/11/october-2020-cib-report/ (accessed 4 November 2021).

286. Graphika. 2021 Spamouflage Breakout. See https://graphika.com/reports/spamouflage-breakout/ (accessed 4 November 2021).

Techniques used to promote misinformation

Approaches by actors aiming to promote misinformation constantly evolve and have varying levels of sophistication. These approaches can involve complex uses of digital technology (eg generating deepfakes) as well as old-fashioned techniques (eg buying influence). We outline four key approaches which can arise in the online information environment.

Content manipulation

This approach involves physically altering genuine content to present a different narrative. This is not to be confused with the concept of 'spin' which relates to a biased interpretation of genuine content[287].

Content manipulation in the online information environment mainly involves editing text, images, and video (as discussed in the previous section) however, it can also involve 'poisoning' datasets and altering medical records. For example, in 2019, a team of researchers from Ben Gurion University found that they could use a computer virus to alter medical scan images to display fake tumours[288]. These forms of attacks can be particularly effective on semi-supervised machine learning models (which learn from a small subset of labelled data and a large set of unlabelled data) as the unlabelled dataset receives minimal review and can be more easily poisoned[289]. This can lead to machine learning models misclassifying data and presenting false outcomes.

System gaming

Disinformation in the online information environment has been described as an 'arms race' between those spreading misinformation and those countering it[290]. This is exemplified by the actions of some internet users to 'game' the counter-misinformation policies of social media platforms.

In response to the rise of social media botnets[291] (networks of social media bots controlled by an individual or organisation), social media platforms introduced policies which would reduce the reach of content or accounts which received high engagement from bots. Following this, the policy has been exploited to instead reduce the reach of legitimate journalistic reporting[292]. This was achieved by having bots engage with the accounts of target journalists.

Other mechanisms of system gaming, include downrating or uprating content[293], hijacking trending topics[294], and flooding the notifications of target users[295].

[287] Andrews L. 2006 Spin: from tactic to tabloid. Journal of Public Affairs. 6, 31-45. (https://doi.org/10.1002/pa.37).

[288] Mirsky Y, Mahler T, Shelef I, Elovici Y. 2019 CT-GAN: Malicious tampering of 3D medical imagery using deep learning. Proceedings of the 28th USENIX Security Symposium. (https://arxiv.org/abs/1901.03597).

[289] Carlini N. 2021 Poisoning the unlabelled dataset of semi-supervised learning. Proceedings of the 30th USENIX Security Symposium. (https://arxiv.org/abs/2105.01622).

[290] Jack S. 2020 Facebook's Zuckerberg defends actions on virus misinformation. BBC News. 21 May 2020. See https://www.bbc.co.uk/news/business-52750162 (accessed 4 November 2021).

[291] Foster J. 2015 The rise of social media botnets. Dark Reading. 7 July 2015. See https://www.darkreading.com/attacks-breaches/the-rise-of-social-media-botnets (accessed 4 November 2021).

[292] Royal Society roundtable with Major Technology Organisations, March 2021.

[293] This is also known as 'review bombing' or 'vote brigading'.

[294] Michael K. 2017 Bots without borders: how anonymous accounts hijack political debate. The Conversation. 24 January 2017. See https://theconversation.com/bots-without-borders-how-anonymous-accounts-hijack-political-debate-70347 (accessed 4 November 2021).

[295] The surprising new strategy of pro-Russia bots. BBC News. 12 September 2017. See https://www.bbc.co.uk/news/blogs-trending-41203789 (accessed 4 November 2021).

Enticement

With an objective to eventually influence a large audience, disinformation actors can generate seemingly benign accounts, popularise them, and then utilise them for malign purposes. This can be achieved by creating accounts which publish popular videos or memes in order to entice a large online following[296]. Once a sufficient following has been gained, these accounts will then publish content intended to influence users.

Another route to enticement is to create authoritative-appearing pages or community spaces and use them to disseminate content designed to influence users for political or financial purposes. This has been a particular problem during the coronavirus pandemic, with some websites using the NHS acronym to promote questionable remedies to COVID-19[297]. The practice has also been found to have been adopted by governments, including the UK Government's counter-terrorism programme[298]. The Mueller investigation into Russian interference in the 2016 US Presidential Election found that creating community Facebook groups was a tactic adopted by the Internet Research Agency[299].

Buying influence

Despite contravening the policies of major social media platforms[300], the buying and selling of accounts and pages can be a technique adopted by disinformation actors wishing to reach a large audience without needing to build one from scratch[301]. This activity involves approaches being made to the administrators or owners of popular social media pages with an offer to purchase the page, followed by negotiation, and transfer of administrator rights for the page[302].

Alternatively, disinformation actors may pay to advertise on social media platforms[303]. This involves creating social media accounts or pages and then paying for adverts which can be hyper-targeted to users based on demographic data (eg location, gender, age, interests, political leanings). Unlike traditional advertising (eg on billboards), social media adverts are often approved using an automated system rather than manual review[304], meaning that it can be harder to identify disinformation content being placed in adverts.

[296] Graphika. 2020 The Case of the Inauthentic Reposting Activists. See https://graphika.com/reports/the-case-of-the-inauthentic-reposting-activists/ (accessed 4 November 2021).

[297] ASA ruling on Go-Vi Ltd. Advertising Standards Authority. 23 June 2021. See https://www.asa.org.uk/rulings/go-vi-ltd-a20-1085562-go-vi-ltd.html (accessed 4 November 2021).

[298] Cobain I. 2019 'This Is Woke': The media outfit that's actually a UK counter-terror programme. Middle East Eye. 15 August 2019. See https://www.middleeasteye.net/news/revealed-woke-media-outfit-thats-actually-uk-counterterror-programme (accessed 4 November 2021).

[299] US Department of Justice. 2019 Report On The Investigation Into Russian Interference In The 2016 Presidential Election. See https://www.justice.gov/archives/sco/file/1373816/download (accessed 4 November 2021).

[300] Instagram. Community Guidelines. See https://help.instagram.com/477434105621119 (accessed 4 November 2021).

[301] Confessore N, Dance G, Harris R, Hansen M. 2018 The Follower Factory. The New York Times. 27 January 2018. See https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html (accessed 4 November).

[302] Viral Accounts. Sell your Facebook fanpage. See https://viralaccounts.com/sell/facebook-fanpage/ (accessed 4 November 2021).

[303] Stamos A. 2017 An update on information operations on Facebook. Meta. 6 September 2017. See https://about.fb.com/news/2017/09/information-operations-update/ (accessed 4 November 2021).

[304] Facebook. Advertising Policies. See https://www.facebook.com/policies/ads/ (accessed 4 November 2021).

Tools and approaches for countering misinformation

In response to the challenges of online misinformation, an anti-misinformation ecosystem has emerged across the public, private, and third sector. This ecosystem consists of investigative journalists; academic researchers; technology companies; educational charities; lawyers; government bodies; and internet users. Each constituent part of the ecosystem plays an essential role in creating a healthier online information environment. Their activities are centred around helping the public understand how misinformation works, detecting misinformation content, correcting viral mistruths, limiting damage, delivering justice, and preventing misinformation content.

Automated detection systems

It is estimated that millions of messages are submitted online by internet users every minute[305]. On WhatsApp, 42 million messages are sent per minute[306]. These high quantities make it unfeasible for there to be a manual review of each piece of shared digital content. As a result, major platforms have developed automated review systems which can detect, flag, and address problematic content[307]. These systems are not fully automated and often require human intervention at the training stage[308] (eg tagging training datasets) and for appeals (where sanctioned users challenge a decision made against them). These systems can be applied to detect illegal content (eg child abuse images, violent content), harmful content (eg hate speech, health misinformation), and specific types of content (eg copyrighted music, deepfakes).

Automated detection systems can be applied for various purposes including:

- Blocking content at the point of upload
- Removing content shortly after upload
- Flagging problematic content
- Adding context and resources to content

These systems are imperfect and have a number of limitations. The Center for Democracy and Technology[309] identify the following five limitations:

i. Natural language processing (NLP) tools perform best when they are trained and applied in specific domains and cannot necessarily be applied with the same reliability across different contexts.

ii. Decisions based on automated social media content analysis risk further marginalising and disproportionately censoring groups that already face discrimination (by amplifying social biases).

[305] Jenik C. 2020 A Minute on the Internet in 2020. Statista. 21 September 2020. See https://www.statista.com/chart/17518/data-created-in-an-internet-minute/ (accessed 4 November 2021).

[306] *Ibid.*

[307] Royal Society roundtable with Major Technology Organisations, March 2021.

[308] Tubaro P, Casilli A, Coville M. 2020 The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. Big Data & Society. 7. (https://doi.org/10.1177%2F2053951720919776).

[309] Center for Democracy and Technology. 2017 Mixed messages? The limits of automated social media content analysis. See https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/ (accessed 4 November 2021).

iii. NLP tools require clear, consistent definitions of the type of speech to be identified; policy debates around content moderation and social media mining tend to lack such precise definitions[310].

iv. The relatively low accuracy and intercoder reliability achieved in natural language processing studies warn strongly against widespread application of the tools to social media content moderation.

v. Even state-of-the-art NLP tools remain easy to evade and fall far short of humans' ability to parse meaning from text.

A further limitation is the impact that training automated tools can have on the mental health of workers who are annotating or reviewing harmful content[311].

Despite these issues, automated detection systems remain a significant part of the anti-misinformation ecosystem and are used as a key performance indicator by social media platforms to assess the quality of their response to misinformation content[312].

Emerging anti-misinformation sector

A small sector of anti-misinformation organisations has formed over recent years and has become an important part of the anti-misinformation ecosystem. These are part of the wider 'safety tech' sector and include organisations working on automated detection systems, fact-checking, user-initiated protection services, and support for human moderators[313]. A number of these organisations work in cooperation with major social media platforms to combat scientific misinformation[314] with these partnerships featuring in promotional advertising campaigns[315].

The sector, which is yet to mature and contains many start-ups, is vulnerable to funding challenges[316]. Due to the nature of their work, it can be undesirable to take funding from governments or technology companies. The sector also lacks consistent definitions and approaches on misinformation content[317], however initiatives such as the Poynter Institute's International Fact-Checking Network seek to address this by promoting best practice in the field[318]. The Poynter Institute have also identified the business model and sustainability of fact-checking organisations as a key challenge[319].

[310] This limitation was also raised during the Royal Society roundtable with Safety Tech Organisations.

[311] Elliott V, Parmar T. 2020 "The despair and darkness of people will get to you". Rest of World. 22 July 2020. https://restofworld.org/2020/facebook-international-content-moderators/ (accessed 4 November 2021).

[312] Royal Society roundtable with Major Technology Organisations, March 2021.

[313] HM Government. 2021 Safer technology, safer users: The UK as a world-leader in Safety Tech. See https://www.gov.uk/government/publications/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech/safer-technology-safer-users-the-uk-as-a-world-leader-in-safety-tech (accessed 4 November 2021).

[314] Royal Society roundtable with Major Technology Organisations, March 2021.

[315] Collins K. 2020 Facebook partners with Full Fact to help people spot fake news. CNET. 29 June 2020. See https://www.cnet.com/tech/services-and-software/facebook-partners-with-full-fact-to-help-people-spot-fake-news/ (accessed 4 November 2021).

[316] Royal Society roundtable with Safety Technology Organisations, March 2021.

[317] *Ibid*.

[318] Poynter Institute. The International Fact-Checking Network. See https://www.poynter.org/ifcn/ (accessed 4 November 2021).

[319] Mantzarlis A. 2016 There's been an explosion of international fact-checkers, but they face big challenges. Poynter Institute. 7 June 2016. See https://www.poynter.org/fact-checking/2016/theres-been-an-explosion-of-international-fact-checkers-but-they-face-big-challenges/ (accessed 4 November 2021).

Other organisations in the sector focus on developing automated detection systems, advising marketing agencies on where not to place adverts, tracking the spread of misinformation content, and creating trust ratings for news websites[320].

Provenance enhancing technology

Focusing instead on the origins of content rather than its value, organisations developing provenance enhancing technologies aim to equip information consumers with the means to help them decide whether a piece of content is genuine and not manipulated. This is achieved by applying the content's metadata (eg sender, recipient, time stamp, location) to determine who created it, how it was created, and when it was created[321].

This is the primary aim of the Coalition for Content Provenance and Authenticity[322] (an initiative led by Adobe, ARM, the BBC, Intel, Microsoft, TruePic, and Twitter) which is developing a set of technical specifications on content provenance. If sufficiently enabled, platforms would be able to better address or label problematic content and information consumers will be able to determine the veracity of a claim, image, or video[323].

APIs for research

Taking advantage of the openness of the internet and the benefits of collective intelligence[324], some organisations (eg Twitter and the NHS) enable application programming interfaces (APIs). These APIs are a mechanism which make it possible for one entity to have access to data held by another entity. In the online information environment, APIs can enable researchers to analyse the spread of content across social media platforms and can enable third party providers access to high quality resources. However, these APIs have been criticised for being overly restrictive[325] following the Cambridge Analytica scandal[326]. The NHS API has been used to provide answers to health queries on third-party voice assistants[327] and the Twitter API has been used to detect, understand, and counter misinformation[328].

[320] Royal Society roundtable with Safety Technology Organisations, March 2021.

[321] Royal Society roundtable with Major Technology Organisations, March 2021.

[322] Coalition for Content Provenance and Authenticity. See https://c2pa.org/ (accessed 4 November 2021).

[323] Royal Society roundtable with Major Technology Organisations, March 2021.

[324] Nesta. 2020 The future of minds and machines: How artificial intelligence can enhance collective intelligence. See https://www.nesta.org.uk/report/future-minds-and-machines/ (accessed 4 November 2021).

[325] Bruns A. 2019 After the 'APIcalypse': social media platforms and their fight against critical scholarly research. Information, Communication & Society. 22, 1544-1566. (https://doi.org/10.1080/1369118X.2019.1637447).

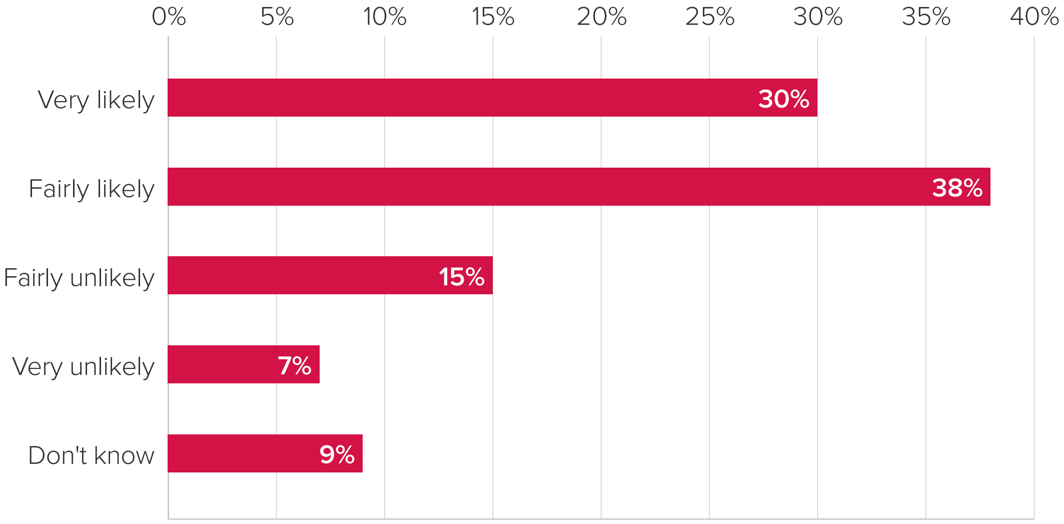
[326] Cadwalladr C, Graham-Harrison E. 2018 Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. 17 March 2018. See https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election (accessed 4 November 2021).

[327] Lake E. 2019 How we are talking to Alexa. NHS Digital. 25 July 2019. See https://digital.nhs.uk/blog/transformation-blog/2019/how-we-are-talking-to-alexa (accessed 4 November 2021).

[328] Twitter. About Twitter's APIs. See https://help.twitter.com/en/rules-and-policies/twitter-api (accessed 4 November 2021).

FIGURE 7

Survey results for the question: Please imagine that you saw a scientific claim (eg on social media, a news article etc) that you found to be suspicious or surprising. How likely or unlikely would you be to fact-check it? (By 'fact-check', we mean confirming the accuracy of any statistics or claims included in a piece of content.)



Source: Royal Society / YouGov, July 2021. (n=2,050)

Unlike Twitter, which is more public and open by default, Facebook is what is known as a 'walled garden'[329] – a controlled data ecosystem with restricted access. As such, it is more complex to carry out research on Facebook, although there are tools (eg CrowdTangle) which provide access to publicly shared content[330]. If successfully actioned, Recommendations 6 (social media data access) and 7 (best practice tools and guidance) should help develop a more open environment benefiting both researchers and emerging online platforms.

[329] McCown F, Nelson M. 2009 What happens when Facebook is gone?Proceedings of the 9th ACM/IEEE-CS Joint conference on Digital libraries. 251-254. (https://doi.org/10.1145/1555400.1555440).

[330] Dotto C. 2020 How to analyze Facebook data for misinformation trends and narratives. First Draft. 7 May 2020. See https://firstdraftnews.org/articles/how-to-analyze-facebook-data-for-misinformation-trends-and-narratives/ (accessed 4 November 2021).

Active bystanders

Active bystanders are individuals who intervene when they see problematic actions or content, unlike passive bystanders who witness problematic content but choose not to intervene[331]. Active bystanders in the online information environment are users who intervene when they witness problematic content (eg abuse or disinformation). In the context of misinformation, an intervention could be to directly respond to a message, or it could involve reporting a user to the social media platform. It may also involve forwarding content to a third party (eg a public health body, a fact checker, or the police) or downrating content in order for it to not feature prominently in users' social media feeds.

Platforms and public bodies have put in measures to make it simpler for internet users to be active bystanders through the use of reporting tools[332].

Community moderation

Community moderators (individuals with administrative control over an online forum) can play an important role in ensuring healthy online discourse[333]. However, there are significant limitations for their role in addressing misinformation content as moderators are faced with the challenge of deciding what is and is not misinformation. These moderators are often untrained volunteers, although in recent years there have been moves to pay moderators[334] and provide formal training[335]. Rules for moderation can be decided by the moderator themselves, or by the community being moderated.

Anti-virals

An emerging trend in the online information environment is a shift away from public discourse to private, more ephemeral, messaging[336]. This creates a challenge for platforms, researchers, and journalists who want to analyse the spread of misinformation as it is highly complex and resource intensive to attempt to study. One solution which has been adopted is to restrict the virality of messages, regardless of whether or not it contains problematic content. This solution has been implemented by WhatsApp with users being unable to easily forward a message which has been through five chat sessions already[337]. This is achieved by tracking how many times a message has been forwarded with the messages remaining encrypted[338].

[331] MIT. Active bystanders: Definition and philosophy. See https://web.mit.edu/bystanders/definition/index.html (accessed 4 November 2021).

[332] World Health Organization. How to report misinformation online. See https://www.who.int/campaigns/connecting-the-world-to-combat-coronavirus/how-to-report-misinformation-online (accessed 4 November 2021).

[333] Matias J. 2019 The Civic Labor of Volunteer Moderators Online. Social Media & Society. 5. (https://doi.org/10.1177%2F2056305119836778).

[334] Upwork. Forum Moderator jobs. See https://www.upwork.com/freelance-jobs/forum-moderation/ (accessed 4 November 2021).

[335] Facebook. Learning Labs. See https://www.facebook.com/community/learning-labs/ (accessed 4 November 2021).

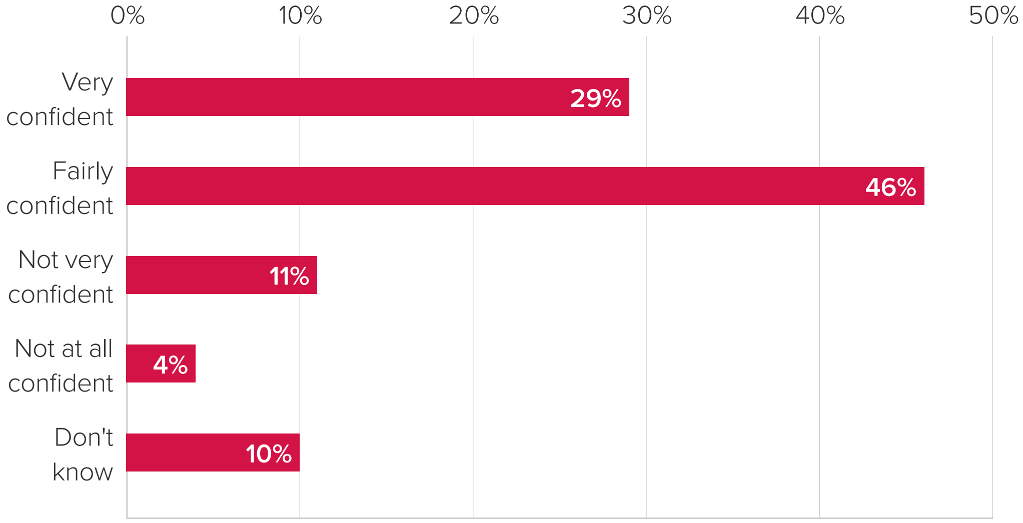
[336] Royal Society roundtable with Major Technology Organisations, March 2021.

[337] WhatsApp. About forwarding limits. See https://faq.whatsapp.com/general/chats/about-forwarding-limits/?lang=en (accessed 4 November 2021).

[338] Royal Society roundtable with Major Technology Organisations, March 2021.

FIGURE 8

Survey results for the question: In general, how confident, if at all, would you feel challenging a scientific claim made by a friend or family member that you felt was suspicious (ie asking them where they heard this, giving a different side of the debate, telling them it may be incorrect, etc)?



Source: Royal Society / YouGov, July 2021. (n=2,019)

This ‘anti-viral’ approach has been adopted by Twitter for a different challenge – that of misinformation content shared by politicians or other high-profile figures and institutions. Twitter’s approach involves disabling engagement with tweets which contain misleading content (in contravention with Twitter’s policies) from high profile figures or with high engagement[339].

This is part of a wider movement within the company to introduce friction into the platform. Another example is their introduction of a prompt for users to read articles before retweeting them[340]. They claim the prompt led to a third of users (who were about to retweet an article) to read the article first and half of users decided to cancel their retweets[341].

[339] Gadde V, Beykpour K. 2020 Additional steps we’re taking ahead of the 2020 US Election. Twitter. 9 October 2020. See https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes (accessed 4 November 2021).

[340] Twitter is bringing its ‘read before you retweet’ prompt to all users. The Verge. 25 September 2020. See https://www.theverge.com/2020/9/25/21455635/twitter-read-before-you-tweet-article-prompt-rolling-out-globally-soon (accessed 4 November 2021).

[341] Royal Society roundtable with Major Technology Organisations, March 2021.

[342] Nesta. 2020 The future of minds and machines: How artificial intelligence can enhance collective intelligence. See https://www.nesta.org.uk/report/future-minds-and-machines/ (accessed 4 November 2021).

Collective intelligence

The open nature of the internet reduces the barriers to 'collective intelligence', or the enhanced insights gained from people working together[342]. Examples of collective intelligence in the online information environment include Wikipedia (multiple editors refining encyclopaedic articles), Waze (drivers reporting potholes, traffic light cameras, hazards etc.), and Quora (with users collaborating on question and answers). This can be a particularly useful tool against misinformation content and is akin to the principle of academic peer review, with outputs being evaluated by others.

On Wikipedia, volunteers work together to edit and verify each other's articles with the most controversial topics being the most heavily scrutinised (with thousands of editors)[343]. On social media platforms (eg Facebook, Twitter, TikTok), users are able to report problematic content to the company so that they can be addressed. Building on this, Twitter has announced plans to introduce an initiative called 'Birdwatch' in which users are able to write notes and provide context to tweets which they consider to be misleading[344]. This use of collective intelligence can reduce the reliance of platforms on paid content moderators and automated detection systems.

Traditional institutions have also been making use of collective intelligence techniques. An example of this is the Trusted News Initiative (TNI) led by the public service broadcaster, the BBC. The TNI operates an early warning system of rapid alerts in which partners warn each other about the spread of disinformation content. Partners in the TNI include media outlets such as the AFP, CBC, and the Financial Times and technology companies such as Google, Facebook, Microsoft, and Twitter[345].

[343] Royal Society roundtable with Major Technology Organisations, March 2021.

[344] Coleman K. 2021 Introducing Birdwatch, a community-based approach to misinformation. Twitter. 25 January 2021. See https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation (accessed 4 November 2021).

[345] Trusted News Initiative (TNI) steps up global fight against disinformation with new focus on US presidential election. BBC Media Centre. 13 July 2020. See https://www.bbc.co.uk/mediacentre/latestnews/2020/trusted-news-initiative (accessed 4 November 2021).

Chapter four
Trustworthy institutions

Left
Library Walk, Manchester
© George-Standen.

Trustworthy institutions

Simply being trustworthy is not sufficient, institutions must also demonstrate and communicate their trustworthiness in order to earn trust.

Institutions have a key role to play in ensuring the production, maintenance and communication of good quality scientific information.

Scientific enquiry is a complex endeavor that depends on the collaboration of many people with a wide variety of specialisms and expertise. Institutions (including universities, publishers, archives and learned societies) are essential in facilitating and enabling this collaboration. These institutions provide guarantees as to the quality of research they process, enabling others to trust in and build upon that work[346]. They also supply the expertise necessary to preserve and curate the outputs of research, such as data sets, experimental results, and papers.

Beyond the system of scientific research, the summarisation and accurate communication of complex topic areas requires trustworthy actors skilled in conveying technical and scientific information in accessible ways. While misinformation about science is not a new phenomenon[347], the increasing visibility of such misinformation to a broad online audience, and the ability of misinformation actors to connect and reinforce each other, means that there is a need for actors that can be trusted to have a visible presence.

Trust can only be earned through trustworthy behavior. Drawing on the work of Onora O'Neill, institutions which are trustworthy are those that demonstrate they are reliable, competent and honest[348]. Simply being trustworthy is not sufficient, institutions must also demonstrate and communicate their trustworthiness in order to earn trust.

Institutions which consistently act in such a way are likely to find that people place their trust in them. It is important to note that, for many people, trust in establishment institutions may have been damaged over time due to negative individual and collective experiences with them.

This distinction between trust and trustworthiness is especially important with regards to science. Science describes a set of methods that deal with investigating uncertainty and comprises a dynamic set of processes, not a static body of knowledge. The body of scientific knowledge changes over time.

Trustworthiness is domain specific. Institutions develop reliability over time to particular fields, and develop appropriate expertise to those fields. Institutions that are trustworthy with regards to a particular branch of science, or any other field of knowledge, may not be able to adequately replicate those in other fields.

Digital technologies pose novel challenges for trustworthy institutions. They offer many opportunities in terms of connecting, synthesising and communicating scientific information. At the same time, while many of the fundamental issues pertaining to misinformation have long histories, the ways that the internet has changed how information is shared, edited and compiled poses new challenges in how trustworthiness is communicated.

[346] Kerasidou A. 2021 Trustworthy Institutions in Global Health Research Collaborations. In: Laurie, Dove, Ganguli-Mitra, McMillan, Postan, Sethi & Sorbie (eds.) The Cambridge Handbook of Health Research Regulation.

[347] Sleigh C. 2021 Fluoridation of drinking water in the UK: A case study of misinformation before social media. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[348] O'Neill O. 2018 Linking Trust to Trustworthiness, International Journal of Philosophical Studies, 26:2, 293-300.

Decontextualisation

Traditionally, an important technique for assessing the quality of information has been to evaluate the information's provenance and sources. Regulation has often supported this, as in the cases of publishing, or broadcasting, or print advertising, where an imprint is required, and there may be penalties for false or defamatory content. The nature of online information sharing means that such stamps of authority (and such potential consequences) are often absent, and content shared without credentials and without risk. Even where they do exist, content is frequently excerpted or edited in such a way that the details of authorship, or the endorsement of trustworthy institutions, is lost or mischaracterised in the process.

Strategies for addressing the damage decontextualisation causes can take the form of preserving and emphasising provenance, such as through the use of provenance enhancing technologies[349] such as blockchain to create an immutable paper trail of changes, or the promotion on online platforms of content produced by trustworthy institutions, as has been used to boost public health messaging during the COVID-19 pandemic. Alternatively, strategies can be pursued that improve the ability of people interpret non-provenance-based markers of authority, such as the quality and transparency of data-handling[350].

Changing and tracking what institutions say

The challenges of communicating inherently dynamic and uncertain scientific topics are compounded by the changeable nature of online material. Much material on the internet is ephemeral, being rapidly uploaded, consumed and discarded. Web pages are regularly edited and updated. Preserving a trail of what different websites, individuals or institutions have said, and how that has changed over time, is an important component of assessing and communicating trustworthiness. However, current attempts to capture the archival data to track this are limited by lack of supporting legislation, addressed in more detail below.

Quantity of information

While no exact figures are available, the number of peer-reviewed scientific journals in the world has been estimated to be approximately 30,000. An estimated 2.5 million scientific papers are published every year, and the number of publishing scientists was estimated in 2018 to be growing by 4 – 5% a year over the previous decade[351]. The COVID-19 pandemic has exacerbated this trend, with Nature estimating over 100,000 articles published on the pandemic in 2020 alone, not including the extensive use of non-reviewed pre-print articles made available[352]. One of the largest science publishers saw an increase of 58% in submissions in the same year[353].

While there are obvious benefits to an increase in the amount of scientific research being made available, the sheer quantity of work poses its own challenges. It becomes increasingly difficult for any one researcher to keep track of all the updates in their own specialism, let alone those in adjacent fields.

[349] See Chapter Three: Techniques for countering misinformation.

[350] Blastland M, Freeman A, van der Linden S, Marteau T, Spiegelhalter D. 2020 Five Rules for Evidence Communication. Nature. 587, 362-364.

[351] The Royal Society. Over 350 years of scientific publishing. See https://royalsociety.org/journals/publishing-activities/publishing350/ (accessed 4 November 2021).

[352] Else H. 2020 How a torrent of COVID science changed research publishing — in seven charts. Nature. 588. 533.

[353] *Ibid*.

The quality of research varies and it can be challenging to take time to evaluate the quality of such a variety of publications[354]. Participants at workshops hosted at the Royal Society have remarked on both the rarity and value of cross-disciplinary approaches to scientific misinformation in different fields.

This creates a demand for hallmarks of trustworthiness and institutions that can help navigate such a broad space. It also increases the importance for authoritative and trustworthy syntheses of evidence, such as those offered by the Cochrane reports in the field of medical science designed to keep practitioners up-to-date with theoretical developments in their field. It also opens up the possibility for advanced data-scraping approaches that can rapidly and accurately collate relevant information from large data sets.

Hallmarks of trustworthiness

What do reliability, competence and honesty look like in practice?

The UK Statistics Authority has formally recognised the importance of trustworthy institutions in its Code of Practice for Statistics, designed to build public confidence in statistics produced by government[355]. The trustworthiness of the people and institutions that handle statistics is the first of three pillars, the others being quality of data and methodology, and demonstrable social value. The Code offers a useful example of the practical operationalisation of principles of trustworthiness. It does this through creating expectations for statistics-handling organisations to demonstrate impartiality, transparent decision making, and appropriate skills and governance capabilities.

Within the scientific system, open science seeks to utilise transparency in a similar way to produce high quality science, guarantee competency, and improve communication. Open science makes scientific papers readily accessible to all audiences, while also offering transparency of the underlying data, and enabling a broader array of reviewer comments on papers. As well as offering transparency, open science approaches give greater scope for the production of replication and null result experiments, which are important parts of the scientific process in reinforcing existing knowledge[356].

A key part of transparency contributing to trustworthiness is being clear about the objectives being pursued. Science researchers and communicators can act as simple informers trying to accurately represent the current state of knowledge and uncertainty in their particular field, or as persuaders trying to actively effect a change in an audience's thought or behaviour (which could include fellow researchers). While both are valid, it is important to distinguish between them. Suspicion that actor motivation is being hidden is often cited as a leading reason for an actor losing trust[357].

[354] The Royal Society and the Academy of Medical Science. 2018 Evidence synthesis for policy. See https://acmedsci.ac.uk/file-download/36366486 (accessed 4 November 2021).

[355] UK Statistics Authority. Code of Practice for Statistics. See https://code.statisticsauthority.gov.uk/ (accessed 4 November 2021).

[356] In praise of replication studies and null results. Nature. 25 February 2020. See https://www.nature.com/articles/d41586-020-00530-6 (accessed 4 November 2021).

[357] Ipsos MORI. Global Trends 2020. See https://www.ipsosglobaltrends.com/ (accessed 4 November 2021).

The importance of curatorship

The scientific system depends on accurate stores of data and information that are accessible for others to use for research. Maintenance of such stores requires skilled curatorship – trusted libraries and archives are important institutions in this regard. Libraries are also important locations for the teaching of media and information literacy skills, especially for the adult population. However, while libraries and archives have started to adapt to the proliferation of online information, their ability to do this is heavily circumscribed by lack of up-to-date legislation.

Although organisations such as the British Library began collecting and archiving websites around 15 years ago, in 2013 the UK Government introduced new regulations that required digital publications to be systematically preserved as part of something known as legal deposit. Legal deposit has existed in English law since 1662 and obliges publishers to place at least one copy of everything they publish in the UK and Ireland – from books to music and maps – at a designated library.

Since it was extended to include digital media, the six designated legal deposit libraries in the UK have accumulated around 700 terabytes of archived web data as part of the UK Web Archive, growing by around 70 terabytes every year. The libraries automatically collect – or crawl – UK websites at least once a year to gather a snapshot of what they contain, while some important websites such as news sites are collected daily. They also collect ebooks, electronic journals, videos, pdfs and social media posts – almost everything that is available in a digital format.

Access to this material is extremely limited. Due to the current legislative framework, historic pages for only around 19,000 or so websites (out of an estimated 4 million) can be accessed through the Web Archive's online portal[358]. These are sites where their creators have given explicit permission to allow open access to their content, but contacting every UK website in this way is almost impossible. For the rest, even though access is permitted and the material is held digitally, researchers must travel to one of nine named sites in person. These sites are inefficiently distributed around the country, with only one access point in England outside of the London-Cambridge-Oxford triangle. The framework also permits only one researcher to use a piece of material at any one time, an arbitrary limitation when it comes to digital access.

This framework for access is now out-of-date to how people access and use data, and severely limits the value that trustworthy libraries and archives are able to offer. Opening up the Web Archive would allow it to be mined at scale for high quality information using modern text analysis methods, helping address the challenges posed by the sheer quantity of material. It would enable researchers, businesses, journalists and anyone else with an interest to uncover trends or information hidden in web pages from the past.

The frameworks governing electronic legal deposit need to be reviewed and reformed to allow wider access. Part of such review will involve considering the data held in these legal deposits and available on-site that remains commercially valuable, such as newspaper archives. Rather than act as a barrier to access, systems such as micropayments – like those to authors of books borrowed from libraries already – could be applied to material held in commercial archives.

> The six designated legal deposit libraries in the UK have accumulated around 700 terabytes of archived web data as part of the UK Web Archive, growing by around 70 terabytes every year.

[358] The British Library. Legal deposit and web archiving. See: https://www.bl.uk/legal-deposit/web-archiving# (accessed 4 November 2021).

CASE STUDY 3

Water fluoridation misinformation.

Following trials and studies throughout the 1940s and 1950s which found that adding fluoride to drinking water could reduce the incidence of tooth decay, a nationwide debate was ignited in the UK surrounding the merits of its introduction[359]. The UK Government's report on the efficacy of fluoridation[360], published in 1962, was intended to rebut rumours about the dangers of fluoride which had been circulating through local communities in the 1950s and early 1960s, however debates on the topic have continued to the present day.

During the peak of conversation in the mid-to-late 1960s, newspapers were filled with letters about fluoride and numerous pamphlets were published by the National Pure Water Association[361] (established by a member of the House of Lords to campaign against water fluoridation). The primary concern was that sodium fluoride (the form added to water) was poisonous. Throughout the first half of the 20th century, sodium fluoride was known as an insecticide and household poison used to kill rats and other types of vermin[362].

There had also been national news stories about livestock fluoride poisoning in Scotland[363]. The combination of these stories, in addition to the presence of fluoride in the 'Piltdown Man' being central to debunking a fraudulent claim of a fossilised human[364] (demonstrating that fluorides accumulate in human bones), contributed to hesitation towards its addition to drinking water.

Similar to contemporary examples of scientific misinformation, anti-fluoridation concerns were rooted more in doubt than testable hypotheses which could be debunked with proof. Furthermore, from the mid-1960s, concerns in anti-fluoridation literature began shifting away from scientific arguments, towards protests of authoritarianism and mass-medication[365]. The National Pure Water Association played a significant role in funding the composition and distribution of literature against fluoridation, which took the form of leaflets and pamphlets handed to people in the streets or sent to local councillors. Beyond literature, misinformation about fluoride was spread through meetings (advertised in newspapers) and through at least one film on the topic[366].

[359] Sleigh C. 2021 Fluoridation of drinking water in the UK: A case study of misinformation before social media. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[360] Ministry of Health. 1962 The conduct of the fluoridation studies in the United Kingdom and the results achieved after five years. London, UK.

[361] National Pure Water Association. About us. See http://www.npwa.org.uk/about-us/ (accessed 4 November 2021).

[362] Sleigh C. 2021 Fluoridation of drinking water in the UK: A case study of misinformation before social media. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[363] *Ibid.*

[364] De Groote *et al.* 2016 New genetic and morphological evidence suggests a single hoaxer created 'Piltdown Man'. Royal Society Open Science. 3. (https://doi.org/10.1098/rsos.160328).

[365] Sleigh C. 2021 Fluoridation of drinking water in the UK: A case study of misinformation before social media. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[366] *Ibid.*

The British Housewives' League were also active in campaigning against fluoridation, citing concerns about government intervention in food production and the effects they believed fluoride may have on foetuses in the womb[367, 368].

In 1967, 110 out of the 203 local health authorities in England and Wales had decided in favour of fluoridating water supplies, and 73 against. By 1969 however, due to the strength of objection, the only councils to do so were Birmingham and Watford[369]. At present, the NHS estimate around 5.8 million people in England receive fluoridated water[370]. In September 2021, the Chief Medical Officers of England, Northern Ireland, Scotland, and Wales jointly recommended the fluoridation of drinking water in the UK[371]. This recommendation is supported by the UK Government, who intend to make it simpler to expand fluoridation schemes[372].

Various online campaigns countering the implementation of water fluoridation schemes and challenging scientific consensus about its safety actively produce and disseminate content on both major and fringe platforms. The Facebook page 'Moms Against Fluoridation' has 135,000 followers and anti-fluoride videos on YouTube have amounted hundreds of thousands of views. Numerous anti-fluoride videos are also present on fringe online platforms such as BrandNewTube, Bitchute, and Odysee.

[367] Whipple A. 2010 'Into every home, into every body': Organicism and Anti-Statism in the British Anti-Fluoridation Movement, 1952-60. 20th century British history. 21, 330-349. (https://doi.org/10.1093/tcbh/hwq016).

[368] Sleigh C. 2021 Fluoridation of drinking water in the UK: A case study of misinformation before social media. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[369] *Ibid*.

[370] NHS. Community water fluoridation in England. See https://www.nhs.uk/conditions/fluoride/ (accessed 4 November 2021).

[371] HM Government. 2021 Statement on water fluoridation from the UK Chief Medical Officers. See https://www.gov.uk/government/publications/water-fluoridation-statement-from-the-uk-chief-medical-officers/statement-on-water-fluoridation-from-the-uk-chief-medical-officers (accessed 4 November 2021).

[372] Badshah N. 2021 Fluoride will be added to UK drinking water to cut tooth decay. The Guardian. 23 September 2021. See https://www.theguardian.com/society/2021/sep/23/fluoride-will-be-added-to-uk-drinking-water-to-cut-tooth-decay (accessed 4 November 2021).

Chapter five
The future landscape

Left
© xijian.

The future landscape

We expect to witness a variety of emerging trends over the coming decade which will positively and negatively affect the online information environment. These include a closing in of the online world towards more private and paid communications, higher and more literate access to the internet, a splintered version of the internet, and an increased focus on climate inactivism.

Proliferation of paywalls

In response to falling revenues, several 'legacy' media outlets have adopted paywalls (a mechanism which requires internet users to pay a subscription fee to read content on news websites)[373]. It is a trend which has also been adopted by individual journalists and 'new media' outlets using paid membership platforms such as Patreon and Substack[374]. On these platforms, online news consumers are commodified with paid supporters or subscribers receiving exclusive content[375].

If this trend continues to grow, we may see a strengthening of both legacy and new media outlets with a reduced reliance on pay-per-click advertising revenue. This reduced reliance would likely to lead to journalists focusing more on the quality of content instead of quantity[376] of engagement.

However, the inelastic global demand for free online news content[377] presents a challenge for a healthy online information environment. The migration of trained journalists and established media outlets to paywalled content leaves a vacuum for potentially lower quality and less robust media outlets to fill. This two-tiered, fragmented system of online news consumption is likely to further complicate efforts to minimise the consumption of misinformation content in future. This should, therefore, further underline the necessity to support public service media. In the UK, outlets such as the BBC, Channel 4, and the ITV regions play a vital role in providing freely available, high-quality, news for those who wish to find it.

Rise of encrypted and ephemeral messaging

A continued growth in ephemeral, story-based, content creation will reduce accessibility and visibility for researchers wishing to study the online information environment. Coupled with an increased consumer awareness around online privacy[378], we are likely to witness a reduction in the amount of data available for analysis by researchers. This is likely to weaken our collective understanding of online information consumption as well as the ability of platforms and regulators to counter the sharing of misinformation content.

[373] Edge M. 2019 Are UK newspapers really dying? A financial analysis of newspaper publishing companies. Journal of Media Business Studies. 16, 19-39. (https://doi.org/10.1080/16522354.2018.1555686).

[374] laser M. 2020 Journalists getting paid: How online platforms are boosting income for writers. Knight Foundation. 3 September 2020. See https://knightfoundation.org/articles/journalists-getting-paid-how-online-platforms-are-boosting-income-for-writers/ (accessed 4 November 2021).

[375] Hunter A. 2016 "It's like having a second full-time job": Crowdfunding, journalism, and labour. Journalism Practice. 10, 217-232. (https://doi.org/10.1080/17512786.2015.1123107).

[376] Tow Center for Journalism. 2017 The Traffic Factories: Metrics at Chartbeat, Gawker Media, and The New York Times. See https://www.cjr.org/tow_center_reports/the_traffic_factories_metrics_at_chartbeat_gawker_media_and_the_new_york_times.php (accessed 4 November 2021).

[377] Reuters Institute for the Study of Journalism. 2021 Digital News Report. See https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021 (accessed 4 November 2021).

[378] ICO's annual report reveals increased public awareness of privacy and information rights issues. UK Information Commissioner's Office. 20 July 2018. See https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2018/07/annual-report-2017-18/ (accessed 4 November 2021).

Influence longevity

Unique to the online information environment, influencers can retain their reach and platform over a prolonged, if not permanent, basis. Prior to the emergence of the online information environment, the influence of scientists and other experts would be limited to their specific field, with their voices platformed by invitation. In the online information environment, experts can generate significant reach (via online followers or subscribers) at specific moments yet retain it permanently.

During the COVID-19 pandemic, several epidemiologists have effectively become social media influencers, gaining tens of thousands of followers online[379]. Following the end of the pandemic, these influencers will retain their platform and reach to share content about epidemiology or opinions about other subjects which they may lack expertise in. It is not yet clear what the long-term implications of this phenomenon will be, which could be positive or negative dependent on the behaviours of the individual experts.

Increasing attention on the role of audio and visual misinformation content

The increasing popularity of podcasts[380] and online video[381] for news and opinion presents novel challenges for researchers and regulators interested in the spread of misinformation content. Examples of discredited scientific theories being broadcast in these arenas include the promotion of race science on popular podcasts[382, 383], and climate denialism on YouTube[384]. Following the popularity of audio social media platform, Clubhouse[385] (a platform involving users speaking audibly in chatrooms), and the introduction of audio features to established social media platforms[386], online discourse may increasingly shift away from being text to audio[387].

Recent announcements by Facebook that they will be focusing efforts on building a virtual and augmented reality platform called the 'Metaverse' is likely to also contribute to an increased need to focus on audio and visual misinformation content[388].

In the online information environment, experts can generate significant reach at specific moments yet retain it permanently.

[379] Ohlheiser A. 2020 Doctors are now social media influencers. MIT Technology Review. 26 April 2020. See https://www.technologyreview.com/2020/04/26/1000602/covid-coronavirus-doctors-tiktok-youtube-misinformation-pandemic/ (accessed 4 November 2021).

[380] Audio on demand: the rise of podcasts. Ofcom. 30 September 2019. See https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/rise-of-podcasts (accessed 4 November 2021).

[381] Ofcom. 2021 News consumption in the UK. See https://www.ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption (accessed 4 November 2021).

[382] Turkheimer E, Harden K, Nisbett R. 2017 Charles Murray is once again peddling junk science about race and IQ. Vox. 18 May 2017. See https://www.vox.com/the-big-idea/2017/5/18/15655638/charles-murray-race-iq-sam-harris-science-free-speech (accessed 4 November 2021).

[383] Evans G. 2018 The unwelcome revival of 'race science'. The Guardian. 2 March 2018. See https://www.theguardian.com/news/2018/mar/02/the-unwelcome-revival-of-race-science (accessed 4 November 2021).

[384] Avaaz. 2020 Why is YouTube Broadcasting Climate Misinformation to Millions? See https://secure.avaaz.org/campaign/en/youtube_climate_misinformation/ (accessed 4 November 2021).

[385] Clubhouse: The social audio app. See https://www.joinclubhouse.com/ (accessed 4 November 2021).

[386] Spaces is here, let's chat. Twitter, 3 May 2021. See https://blog.twitter.com/en_us/topics/product/2021/spaces-is-here (accessed 4 November 2021).

[387] Basu T. 2021 The future of social networks might be audio, MIT Technology Review. 25 January 2021. See https://www.technologyreview.com/2021/01/25/1016723/the-future-of-social-networks-might-be-audio-clubhouse-twitter-spaces/ (accessed 4 November 2021).

[388] Founder's Letter 2021. Meta. 28 October 2021. See https://about.fb.com/news/2021/10/founders-letter/ (accessed 4 November 2021).

This presents significant challenges for research and content moderation, particularly for live audio and video content, as audiovisual content is more complex to codify and analyse. Furthermore, rapid developments in audio[389] and video deepfakes may lead to more convincing misinformation content and undermine faith in genuine content[390]. Survey experiments commissioned for this report (see Box 2) suggest that most people struggle to identify a deepfake video, even when prompted.

Immature and under-resourced social media platforms

The rapid rise of competitors to established social media platforms (eg Facebook) presents a challenge of maturity and capability for emerging platforms to address misinformation content. As raised in the Society's roundtable with major technology organisations, emerging (yet popular) social media platforms will be expected to handle similar quantities of misinformation content without the benefit of having years of data to train their employees or their automated detection systems. An example is the video-sharing platform, TikTok, which launched in 2016 and has since been downloaded 3 billion times[391]. It is the first application not owned by Facebook to reach this milestone.

This challenge of immaturity, however, applies to all new social media platforms – those with thousands of users as well as those with millions (or billions). Furthermore, not all emerging social media platforms will be able to generate sufficient levels of financial resource to be able to employ human content moderators or to train automated detection systems. Should this occur, we may see a cycle of misinformation discourse in which the same problems continue to re-emerge but on new platforms.

[389] Neural Voice Cloning with a Few Samples. Baidu Research. 20 October 2018. See http://research.baidu.com/Blog/index-view?id=81 (accessed 4 November 2021).

[390] Chesney R, Citron D. 2018 Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. 107 California Law Review. (https://dx.doi.org/10.2139/ssrn.3213954).

[391] TikTok becomes the first non-Facebook mobile app to reach 3 billion downloads globally. SensorTower. 14 July 2021. See https://sensortower.com/blog/tiktok-downloads-3-billion (accessed 4 November 2021).

BOX 2

Survey on people’s capacity for deepfake detection.

To gain insights on the public’s ability to detect deepfake video content, we commissioned a survey experiment using short clips from publicly available videos of the actor Tom Cruise. This included a deepfake video of the actor published on TikTok in February 2021[392]. The video is considered to be of the highest quality published online and was created by VFX artist, Chris Ume, to demonstrate the technical possibilities involved with AI-generated video content[393]. Other videos used in the survey were a mixture of promotional content and media interviews with Cruise.

Two experiments were conducted with a representative sample of UK-based respondents (n=1,093)[394]. The first experiment focused on people’s capacity to detect a deepfake in a natural setting. In other words, do people spot something amiss when they encounter a deepfake without a content warning? Participants in this experiment were randomised into either control or treatment, with the latter watching the deepfake alongside four authentic videos and the former simply watching five authentic videos. In the second experiment, participants were forewarned that at least one of the videos they would see is a deepfake. They were then asked to select the clip(s) they believe to be the deepfake and to say whether they found the choice obvious.

The first experiment found that in a natural digital setting, without a content warning, people are no more likely to notice something out of the ordinary when they view a deepfake video than when they view normal, authentic videos. In the second experiment, the content warning increased the detection rate of the deepfake from 11% to 22%. Though this represents a significant improvement, the vast majority of participants were still unable to select the deepfake.

The implications of these results are twofold. First, without content warnings, it appears that people are neither alert to the presence of deepfakes nor able to tell them apart from authentic videos. Second, while content warnings may increase the deepfake detection rate, the vast majority of people still struggle to detect them.

[392] TikTok. deeptomcruise. See https://www.tiktok.com/@deeptomcruise?lang=en (accessed 4 November 2021).

[393] Hern A. 2021 ‘I don’t want to upset people’: Tom Cruise deepfake creator speaks out. *The Guardian*. 5 March 2021. See https://www.theguardian.com/technology/2021/mar/05/how-started-tom-cruise-deepfake-tiktok-videos (accessed 4 November 2021).

[394] Lewis A, Vu P, Duch R. 2021 Deepfake detection and content warnings: Evidence from two experiments. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

Increasing state-sponsored weaponisation of disinformation

The idea of the online information environment as an arena for inter-state warfare[395] has been widely accepted as a reality and is a considered to be a significant threat by the United Kingdom's intelligence services[396]. The more invested and skilled state actors become at online disinformation campaigns, the more challenging the online information environment will become. These military-grade disinformation campaigns benefit from state funding and continue to affect democratic institutions and private companies[397].

As state-sponsored disinformation matures and becomes more sophisticated, it is likely that current mitigations (eg active bystanders, automated detection) will become less effective. Being continuously aware of the changing nature of this threat will be essential as will a focus on building up the defences of citizens. As recommended in GCHQ's 2020 report on artificial intelligence, this may include investing in mechanisms for detecting deepfakes, blocking botnets, and identifying sources of disinformation[398].

Better digital and information literacy

The digital literacy of the global population should improve over time as more and more users gain access to the internet and become accustomed to its nature. Digital literacy – defined by the American Library Association[399] as the ability to use information and communication technologies to find, evaluate, create, and communicate information – can decrease the perceived accuracy of false news content[400] and should be considered as an essential skill for citizens of all ages.

The UK Government's Online Media Literacy Strategy[401] is designed to meet this need. Its successful implementation – coupled with Ofcom's focus on this[402] – has the potential to result in a more resilient population and could lessen the impact of misinformation content. However, it is important for this strategy to consider all people, of all ages, and be regularly reviewed as the nature of the online information environment evolves over time.

[395] Miller C. 2018 Inside the British Army's secret information warfare machine. Wired. 14 November 2018. See https://www.wired.co.uk/article/inside-the-77th-brigade-britains-information-warfare-military (accessed 4 November 2021).

[396] Director General Ken McCallum gives annual threat update 2021. MI5. 14 July 2021. See https://www.mi5.gov.uk/fa/node/863 (accessed 4 November 2021).

[397] Rana M, O'Neill S. 2020 Russians spread fake news over Oxford coronavirus vaccine. The Times. 16 October 2021. See https://www.thetimes.co.uk/article/russians-spread-fake-news-over-oxford-coronavirus-vaccine-2nzpk8vrq (accessed 4 November 2021).

[398] GCHQ. 2020 AI for National Security – Foreign State Disinformation, Pioneering a New National Security. See https://www.gchq.gov.uk/files/GCHQAIPaper.pdf (accessed 4 November 2021).

[399] American Library Association. Digital literacy. See https://literacy.ala.org/digital-literacy/ (accessed 4 November 2021).

[400] Guess *et al.* 2020 A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proceedings of the National Academy of Sciences July 2020. 117, 15536-15545. (https://doi.org/10.1073/pnas.1920498117).

[401] HM Government. 2021 Online Media Literacy Strategy. See https://www.gov.uk/government/publications/online-media-literacy-strategy (accessed 4 November 2021).

[402] Ofcom. Making Sense of Media. See https://www.ofcom.org.uk/research-and-data/media-literacy-research (accessed 4 November 2021).

Increased internet access

The target for global internet access, set by the International Telecommunications Union (ITU) and UNESCO, is for penetration to reach 75% by 2025[403]. The ITU's current estimate for global penetration is 53%. This rise, if achieved, is likely to increase the amount of misinformation content on the internet. This could arise through an expansion of new media outlets with weak standards or through an increase in disinformation actors. This global outlook is vital, as harmful misinformation content can be exported from one nation's citizens to another's. For example, the COVID-19 pandemic has seen viral online conspiracy theories originate from individuals across the world including the US, Canada, China, Russia, France, and Iran[404].

Furthermore, collective intelligence – covered earlier as a tool for combating misinformation – can also be applied by malicious actors. For example, the increased quality of deepfakes is, in part, due to creators sharing techniques and guides on mainstream platforms such as GitHub[405] or on niche, dedicated blogs. As more people gain internet access, the collective intelligence capabilities for misinformation content production is likely to advance.

New internet protocols

Competing ideas for how the infrastructure of the internet should operate could limit the ability of platforms and policymakers to shape the online information environment in future. These discussions, which take place amongst global entities such as the International Telecommunications Union, the Internet Governance Forum (both of which are UN bodies), and the Internet Engineering Task Force, are centred on the internet's protocols. These protocols are the rules which dictate how information is transmitted between two or more entities[406]. The nature of these protocols is heavily influenced by governments and major technology companies[407].

Currently, the majority of internet traffic is underpinned by the Transmission Control Protocol and the Internet Protocol (TCP/IP) which originated in the 1970s. These protocols ensure that data is sent to the correct recipient and 'understood' upon receipt. Furthermore, they embed an 'end-to-end' principle with the packaging and processing of data happening at each end of the network (the sender and the recipient)[408]. In theory, attempts to exert control over how data is transmitted over the internet is difficult to achieve as its nature is highly decentralised. In practice, governments, technology companies, and internet service providers have developed other mechanisms for filtering content and blocking access[409].

[403] International Telecommunications Union. 2020 The State of Broadband 2020: Tackling Digital Inequalities. See https://www.itu.int/dms_pub/itu-s/opb/pol/S-POL-BROADBAND.21-2020-PDF-E.pdf (accessed 4 November 2021).

[404] The 'superspreaders' behind COVID-19 conspiracy theories. Al Jazeera. 15 February 2021. See https://www.aljazeera.com/news/2021/2/15/the-superspreaders-behind-covid-19-conspiracy-theories (accessed 4 November 2021).

[405] DeepFaceLab. See https://github.com/iperov/DeepFaceLab (accessed 4 November 2021).

[406] Cerf V, Cain E. 1983 The DoD Internet Architecture Model. Computer Networks. 7, 307-318. (https://doi.org/10.1016/0376-5075(83)90042-9).

[407] Demos. 2021 Good Foundations: Why democracies should care about the wiring of the Internet. See https://demos.co.uk/project/good-foundations-why-democracies-should-care-about-the-wiring-of-the-internet/ (accessed 4 November 2021).

[408] *Ibid*.

[409] National Cyber Security Centre. The UK public sector DNS service. See https://www.ncsc.gov.uk/news/uk-public-sector-dns-service (accessed 4 November 2021).

Proposals for new protocols, being developed by governments[410, 411] and other organisations[412] could lead to different versions of the internet, a phenomenon which has been referred to as the 'splinternet'[413]. In these versions, the state's capacity to intervene with the transmission of information over the internet is either enhanced (greater visibility and control) or reduced (less visibility and control). Should the splinternet become a reality, global cooperation on maintaining the health of the online information environment will be hindered and further complicated. Internet users' access to high quality, accurate scientific information may be cut off and researchers' understanding of how misinformation content is disseminating online may be limited.

Hype and sensationalism

As the forces of the attention economy continue to govern the online information environment, it is important to recognise that the production and consumption of scientific articles will continue to be affected. One side-effect of this is a phenomenon known as 'sensationalism' or 'hype' in which the potential benefits or risks from a scientific development are exaggerated[414]. Accusations of this practice have been levelled at climate change activists[415], investors in artificial intelligence[416], and stem cell therapies[417].

Not to be confused with scientific misconduct (the fabrication, falsification, or misrepresentation of results[418]), hype in science involves exaggerating the pros or cons of a particular development.

[410] Gross A, Murgia M. 2020 China and Huawei proposed reinvention of the internet. Financial Times. 27 March 2020. See https://www.ft.com/content/c78be2cf-a1a1-40b1-8ab7-904d7095e0f2 (accessed 4 November 2021).

[411] Sherman, J. 2020 Russia is trying something new to isolate its internet from the rest of the world. Slate. 25 September 2020. See https://slate.com/technology/2020/09/russia-internet-encryption-protocol-ban.html (accessed 4 November 2021).

[412] Mozilla. Firefox DNS-over-HTTPS. See https://support.mozilla.org/en-US/kb/firefox-dns-over-https (accessed 4 November 2021).

[413] O'Hara K, Hall W. 2019 The dream of a global internet is edging towards destruction. Wired. 24 December 2019. See https://www.wired.co.uk/article/internet-fragmentation (accessed 4 November 2021).

[414] Intemann K. 2020 Understanding the problem of "hype": Exaggeration, values, and trust in science. Canadian Journal of Philosophy. 1-16. (https://doi.org/10.1017/can.2020.45).

[415] Webster B. 2020 Baseless climate warnings wiped from Extinction Rebellion film. The Times. 27 August 2020. See https://www.thetimes.co.uk/article/baseless-climate-warnings-wiped-from-extinction-rebellion-film-vskns9f6k (accessed 4 November 2021).

[416] Horgan J. 2020 Will artificial intelligence ever live up to its hype? Scientific American. 4 December 2020. See https://www.scientificamerican.com/article/will-artificial-intelligence-ever-live-up-to-its-hype/ (accessed 4 November 2021).

[417] Montague J. 2019 The 'unwarranted hype' of stem cell therapies. BBC Future. 21 August 2019. See https://www.bbc.com/future/article/20190819-the-unwarranted-hype-of-stem-cell-therapies-for-autism-ms (accessed 4 November 2021).

[418] UK Research Integrity Office. Misconduct in research. See https://ukrio.org/publications/code-of-practice-for-research/3-0-standards-for-organisations-and-researchers/3-16-misconduct-in-research/ (accessed 4 November 2021).

The inherent dangers involved with issues like climate change coupled with the greater virality of 'fear-arousing sensationalism' in the online information environment[419] means it should not be surprising to see more hype and sensationalism in scientific conversations in future.

Advocates for the use of hype, argue that it can advance interest in a particular topic[420], attract funding for further research[421], and provoke action[422]. However, it comes with the risk of undermining trust in science if expectations are not met[423].

Resistance to the fluoridation of drinking water

The fluoridation of drinking water is likely to become significantly affected by the spread of online misinformation in the short to medium term. The topic shares many of the characteristics associated with misinformation about vaccines or 5G. These include i) a push from government for widespread implementation[424], ii) historic fears over its effects on children[425], and iii) criticism from seemingly credible voices[426].

There are, already, numerous active online anti-fluoride campaign groups. On Facebook, there are many accounts focused on the topic. These include Moms Against Fluoridation (135,000 followers), the Fluoride Action Network (85,000 followers), and The Girl Against Fluoride (23,000 followers). On YouTube, a number of anti-fluoride videos have gained hundreds of thousands of views[427]. Anti-fluoride views have also been promoted on major platforms by celebrities including the renown podcast host, Joe Rogan, and the political commentator, Bill Maher[428, 429].

[419] Ali *et al.* 2019 Viruses going viral: Impact of fear-arousing sensationalist social media messages on user engagement. Science Communication. 41, 314-338. (https://doi.org/10.1177%2F1075547019846124).

[420] Culliford E. 2019 Facebook, Microsoft launch contest to detect deepfake videos. Reuters. 5 September 2019. See https://www.reuters.com/article/us-facebook-microsoft-deepfakes-idUSKCN1VQ2T5 (accessed 4 November 2021).

[421] Chubb J, Watermeyer R. 2016 Academics admit feeling pressure to embellish possible impact of research. The Conversation. 16 March 2016. See https://theconversation.com/academics-admit-feeling-pressure-to-embellish-possible-impact-of-research-56059 (accessed 4 November 2021).

[422] The Guardian view on the Extinction Rebellion protests: of course they're an inconvenience. The Guardian. 10 October 2019. See https://www.theguardian.com/commentisfree/2019/oct/10/the-guardian-view-on-the-extinction-rebellion-protests-of-course-theyre-an-inconvenience (accessed 4 November 2021).

[423] Intemann K. 2020 Understanding the problem of "hype": Exaggeration, values, and trust in science. Canadian Journal of Philosophy. 1-16. (https://doi.org/10.1017/can.2020.45).

[424] HM Government. 2021 Statement on water fluoridation from the UK Chief Medical Officers. See https://www.gov.uk/government/publications/water-fluoridation-statement-from-the-uk-chief-medical-officers/statement-on-water-fluoridation-from-the-uk-chief-medical-officers (accessed 4 November 2021).

[425] Sleigh C. 2021 Fluoridation of drinking water in the UK: A case study of misinformation before social media. The Royal Society. See https://royalsociety.org/topics-policy/projects/online-information-environment

[426] Letter to Right Honourable Boris Johnson on water fluoridation from 3 scientists. Fluoride Action Network. 27 September 2021. See https://fluoridealert.org/content/letter-to-boris-johnson-on-water-fluoridation/ (accessed 4 November 2021).

[427] The 'fluoridealert' channel has 1.3m cumulative views, including two individual videos with over 300,000 views each.

[428] 'Joe Rogan – Why is Fluoride in the Water!?'. The Joe Rogan Experience (YouTube channel). 25 July 2018. See https://www.youtube.com/watch?v=VOyukrkF0N4 (accessed 4 November 2021).

[429] 'Bill Maher on Portland's Fluoride Vote: I Would Have Voted No as Well'. FluorideAlert (YouTube channel). 25 May 2013. See https://www.youtube.com/watch?v=apKkLTZUcIY (accessed 4 November 2021).2021.

With the UK Government announcing plans to simplify the expansion of water fluoridation[430] and with decisions being made by local authorities, we expect to see a growth in online campaigns which seek to challenge established science on the matter and resist moves to approve its implementation.

Promotion of climate inactivism

Climate change denialism has been one of the most prominent examples of scientific misinformation, and one that many online platforms have put in place policies to address, as discussed earlier.

The basic science of human-caused climate change is widely accepted with UK society[431]. While not to be taken for granted, action on climate change is perceived as socially desirable, and there exists a broad political consensus on the need to intervene to achieve net-zero. In this context, there is a significant risk in directly addressing the extreme fringe of climate deniers. Given public concern, conferring legitimacy or using resources by engaging such people may be counterproductive[432].

The more serious area to consider misinformation is not the debate over the basic science, but rather how that science is now acted upon. Misinformation that targets the implementation of mitigation or adaptation measures have received less focus, but are of increasing concern given the significant financial and commercial implications associated with these measures. Pre-emptive attention and resources should be given to understanding the drivers of disinformation challenging the operationalisation of climate change mitigation measures.

Further research on the economics of climate misinformation will be important to help develop a more detailed understanding of the material incentives for the most active producers of misinformation content.

[430] Badshah N. 2021 Fluoride will be added to UK drinking water to cut tooth decay. The Guardian. 23 September 2021. See https://www.theguardian.com/society/2021/sep/23/fluoride-will-be-added-to-uk-drinking-water-to-cut-tooth-decay (accessed 4 November 2021).

[431] Lewandowsky S. 2020 Climate Change Disinformation and How to Combat It. Annual Review of Public Health. 42, 1-21. (https://doi.org/10.1146/annurev-publhealth-090419-102409).

[432] Royal Society workshop on Climate Change and Misinformation, July 2020.

Conclusion

Left
© Kkolosov.

Conclusion

Institutions will need to consider how they can best compete in the online attention economy.

The online information environment remains a relatively young innovation still in its formative years, but already has important consequences for how people engage with and use scientific information. Its role, so far, in democratising access to knowledge and transforming the way people produce content has been both a source of great excitement and great concern amongst policymakers. Seamless mechanisms for sharing, collaboration, and scrutiny have had major benefits for society's collective understanding of scientific topics. However, the risk of people acting on misinformation content, causing harm to themselves or others, is a genuine concern which needs to be proactively mitigated across society.

As the online information environment matures, it is essential that people are provided with easy access to good quality information and are made resilient against future attempts to misinform them. As this report highlights, the challenges of scientific misinformation are unlikely to disappear and will continue to evolve. Misinformation about climate change is shifting from denialism towards inactivism and historic debates about water fluoridation appear to be re-emerging. Content manipulation is likely to become more, not less, sophisticated. Online discourse will become increasingly private and harder to analyse. New social media competitors will rapidly arise with little to no experience in addressing harmful content. These challenges will require building resilience within platforms and the people who use them. This concept of 'building resilience' underpins the recommendations of this report, focusing more on proactive steps rather than after-the-event interventions.

Furthermore, as the nature of science is a process of learning and discovery, established consensus can sometimes change over time. Science operates on the 'edge of error', and a key facet of science is its ability to self-correct as new evidence is established. Clearly and carefully communicating scientific uncertainties, new developments, and the grounds for any change in scientific consensus is critical. To do so effectively, institutions will need to consider how they can best compete in the online attention economy whilst at the same time generating trust across society.

The analysis, insights, and recommendations within this report are all geared towards achieving this goal, ensuring that we retain the best qualities of the internet, and understanding how the environment will continue to evolve. The guiding objective for policymakers should be to build the capacity for individuals to make well-informed decisions, regardless of how the landscape and nature of the online information environment changes over time.

Appendix

Appendix

Index of figures and boxes

Working Group members

The members of the Working Group involved in this report are listed below. Members acted in an individual and not a representative capacity, and declared any potential conflicts of interest. Members contributed to the project on the basis of their own expertise and good judgement.

Chair	
Professor Frank Kelly CBE FRS	Emeritus Professor of the Mathematics of Systems, University of Cambridge

Members	
Professor Michael Bronstein	DeepMind Professor of Artificial Intelligence, University of Oxford
Dr Vint Cerf ForMemRS	Vice President and Chief Internet Evangelist, Google
Professor Lilian Edwards	Professor of Law, Innovation, and Society, University of Newcastle
Professor Derek McAuley	Professor of Digital Economy, University of Nottingham
Professor Gina Neff	Professor of Technology and Society, University of Oxford; Executive Director, Minderoo Centre for Technology and Democracy
Professor Rasmus Kleis Nielsen	Professor of Political Communication, University of Oxford; Director, Reuters Institute for the Study of Journalism
Professor Sir Nigel Shadbolt FRS FREng	Professor of Computing Science, University of Oxford; Executive Chair, Open Data Institute
Professor Melissa Terras	Professor of Digital Cultural Heritage, University of Edinburgh; Director, Centre for Data, Culture & Society

Royal Society staff

Royal Society secretariat	
Areeq Chowdhury	Senior Policy Adviser and Project Lead
Dr Natasha McCarthy	Head of Policy, Data
Jessica Montgomery	Senior Policy Adviser (until August 2020)
Timothy Rees Jones	Policy Adviser
Amy Walsh	UKRI placement

Reviewers

This report has been reviewed by an independent panel of experts approved by the Science Policy Committee of the Royal Society. The reviewers were not asked to endorse the conclusions or recommendations of the report, but to act as independent referees of its technical content and presentation. They acted in a personal and not a representative capacity. The Royal Society gratefully acknowledges the contribution of the reviewers.

Reviewers
Professor Sir Mark Walport FMedSci FRS
Professor Diane Coyle CBE
Dr Ananyo Bhattacharya

Acknowledgments

The Royal Society would like to thank all those who contributed to the development of this project, in particular through attendance at the following events:

- Royal Society workshop on Vaccines and Misinformation, July 2020.
- Royal Society workshop on Climate Change and Misinformation, August 2020.
- Royal Society roundtable on Telecoms and Misinformation, November 2020.
- Royal Society roundtable with Major Technology Organisations, March 2021.
- Royal Society roundtable with Safety Technology Organisations, March 2021
- Royal Society workshop on Horizon Scanning Scientific Misinformation, March 2021.

Survey methodologies

Survey on people's capacity to detect deepfake video content

The Royal Society worked with Andrew Lewis from the Nuffield Centre for Experimental Social Sciences (CESS), based at the University of Oxford, to design survey experiments to test people's capacity to detect deepfake video content. Data analysis was undertaken by Patrick Vu from Brown University. The survey used publicly available clips of the American actor, Tom Cruise and was approved by the CESS Ethics Committee.

A representative sample of UK participants (n=1,093) were recruited using the sample provider, Lucid. Once redirected to the Qualtrics-based survey, they were randomised into either control or treatment. In order to assure data quality, participants in both conditions were subjected to an attention check, on the basis of which inattentive participants are screened from the sample. This was done in line with research showing a rise in inattentive users on online survey platforms and the need to include attention screens to obtain high-quality responses. After passing the check, participants then watched a video of Cruise being interviewed to help familiarise them with the actors' face, voice, and mannerisms.

Two experiments were conducted. In Experiment 1, participants in the treatment condition viewed a series of five short videos of Cruise including the deepfake, with the order of presentation randomised (ie some participants viewed the deepfake first, some last, and so on). In the control condition, participants simply viewed five genuine videos of Cruise. All participants then answered a series of questions about their perceptions of the videos they've watched, including the outcome question of interest: "Did anything about these videos strike you as being out of the ordinary?"

Experiment 2 employed a similar design to the first, however in this condition participants were informed that one of the five videos they are to watch is a manipulated deepfake. After viewing the videos, participants were asked to identify the video they believe to be fake and given the chance to provide a brief explanation of how they reached their conclusion. The binary of detecting or not detecting the correct video was the outcome of interest, so there was no need for a control / treatment condition in this design and instead all participants in experiment 2 followed the same protocol.

The combination of these two designs allowed us to compare the public's capacity for detecting manipulated content in either a natural digital environment (experiment 1), or a digital environment that heightens individuals' awareness of potentially fake content (experiment 2). The latter also more directly addressed the question of detection per se, given participants were specifically looking for the manipulated video.

Royal Society / YouGov survey

The Royal Society worked with YouGov to design a survey exploring the public's behaviour towards scientific information online and the prevalence of beliefs in scientific 'misinformation'. The survey was conducted using an online interview administered to members of the YouGov Plc UK panel of 800,000+ individuals who have agreed to take part in surveys. Emails were sent to panellists selected at random from the base sample. The e-mail invited them to take part in a survey and provided a generic survey link. Once a panel member clicked on the link they were sent to the survey that they are most required for, according to the sample definition and quotas. (The sample definition could be "GB adult population" or a subset such as "GB adult females").

Invitations to surveys don't expire and respondents can be sent to any available survey. The responding sample is weighted to the profile of the sample definition to provide a representative reporting sample. The profile is normally derived from census data or, if not available from the census, from industry accepted data.

Commissioned literature reviews

- Sleigh C. 2021 *Fluoridation of drinking water in the UK: A case study of misinformation before social media*. The Royal Society. See https://www.royalsociety.org/online-information-environment
- Arguedas A, Robertson C, Fletcher R, Nielsen R. 2021 *Echo chambers, filter bubbles, and polarisation*. The Royal Society. See https://www.royalsociety.org/online-information-environment
- Cabrera Lalinde I. 2021 *How misinformation affected the perception of vaccines in the 20th century based on the examples of polio, pertussis and MMR vaccines*. The Royal Society. See https://www.royalsociety.org/online-information-environment
- Röttger P, Vedres B. *The Information Environment and its Effects on Individuals and Groups*. The Royal Society. See https://www.royalsociety.org/online-information-environment

The Royal Society is a self-governing Fellowship of many of the world's most distinguished scientists drawn from all areas of science, engineering, and medicine. The Society's fundamental purpose, as it has been since its foundation in 1660, is to recognise, promote, and support excellence in science and to encourage the development and use of science for the benefit of humanity.

The Society's strategic priorities emphasise its commitment to the highest quality science, to curiosity-driven research, and to the development and use of science for the benefit of society. These priorities are:

- Promoting excellence in science
- Supporting international collaboration
- Demonstrating the importance of science to everyone

For further information
The Royal Society
6 – 9 Carlton House Terrace
London SW1Y 5AG

T +44 20 7451 2500
E science.policy@royalsociety.org
W royalsociety.org

Registered Charity No 207043

9 781782 525677

ISBN: 978-1-78252-567-7
Issued: January 2022 DES7656