FACULTY OF
PUBLIC HEALTH

Protecting and improving the health of the public
through the organised efforts of our members

Response to ‘Social media, misinformation and harmful algorithms’ inquiry call for evidence

Faculty of Public Health

This response is from the Faculty of Public Health (FPH), as developed by the Artificial Intelligence and Digital Public Health Special Interest Group. The FPH, as part of the medical Royal College arrangements, is the standard-setting body for public health in the UK and professional home for over 5,000 members of the public health workforce. We advocate on key public health issues and have a strong mandate and responsibility to ensure that the essential functions, standards and resources of a robust public health system are maintained.

Executive summary

To what extent do the business models of social media companies, search engines and others encourage the spread of harmful content, and contribute to wider social harms?

1. Most of the UK population access online content and spend a significant portion of their day online. The majority encounter potentially harmful content.
2. The revenue model of social media platforms incentivises their use of engagement-based content recommendation algorithms.
3. We focus on recommendation algorithms (content propagation) because platforms have almost complete control over what they recommend to a user.
4. Recommendation algorithms are designed to show users popular content. But what becomes popular is the results of users' collective online actions.
5. The rapid amplification and exposure to potentially harmful content is a particularly concerning feature of human interaction with engagement-based content recommendation algorithms.
6. There is growing evidence that exposure to online content translates to actual health harms at an individual and population level, with certain groups more vulnerable to harmful content exposure.
7. Social media platforms can be tools to support health, although the evidence currently skews towards negative rather than positive outcomes.
8. Our judgement is that the growing evidence of harms resulting from the amplification of potentially harmful content by ‘engagement based-recommendation algorithms’ is sufficient to recommend that a public health approach to mitigating the risk of harm from online content must be taken.

What role do generative artificial intelligence (AI) and large language models (LLMs) play in the creation and spread of misinformation, disinformation and harmful content?

9. Generative Artificial Intelligence and large language models could be a powerful public health tool, but as they become used more widely, in an unregulated context, they also pose a significant threat in propagating misinformation.

What more should be done to combat potentially harmful social media and AI content?

10. The British public strongly welcome action to tackle potentially harmful content.
11. The UK government should support evidence-based interventions to protect individuals from potentially harmful online content.
12. The UK government has a duty to protect and promote human health and therefore must strengthen regulation of social media platforms in line with other international examples.

To what extent do the business models of social media companies, search engines and others encourage the spread of harmful content, and contribute to wider social harms?

Social media platforms have transformed society by redefining how people communicate, assess information, and live together (Sindermann et al., 2024). Social media platforms allow individuals to connect within a virtual network to share, co-create, or exchange digital content such as information, messages, photos, and videos (Naslund et al., 2020). “Social media platforms” encompasses social networking sites as well as messaging services. Mainstream platforms, characterised by their large user base, include Facebook, YouTube, WhatsApp, Instagram, WeChat, TikTok, and X. It is estimated that around 58.4% of the total world population uses social media, which is 93.4% of all internet users (Sindermann et al., 2024).

Most (68%) adults online in the UK have been exposed to potentially harmful content. Most (67%) also believe that for them, the benefits of being online outweigh the risks (Ofcom, 2024). There is growing evidence that online content has the potential to both improve and to harm the health of users *and* the wider population (Kanchan & Gaidhane, 2023). We advise applying the precautionary principle and taking proportionate action to mitigate the risks arising from potentially harmful online content, especially for young users who are most at risk.

1. Most of the UK population access online content and spend a significant portion of their day online. The majority encounter potentially harmful content.

- In May 2024, 47.4 million UK adults accessed the internet.
  - They spent an average of 4 hours 20 minutes a day online. Young adults (18-24) spend an average of 6 hours 1 minute online.
  - 94% of online adults visit YouTube, 83% visit Google Search, 70% visit all top three Meta-owned platforms (Facebook, WhatsApp, and Instagram), and 48% visit Reddit.
  - 74% of online 18-24 year olds visited TikTok in May 2024.
- In June 2024, 68% of adult users have experienced potentially harmful content, this climbs to 80% for young adults 18-24.
  - The most prevalent potentially harmful content was misinformation - 39% of users aged 13+ reported encountering misinformation, and 30% said they had seen something in the past four weeks that had made them feel uncomfortable, upset or negative.
  - 26% of UK adult internet users report seeing hateful, offensive or discriminatory content.
- 20% of 8-15 year olds with a social media profile have a user profile age of at least 18, exposing them to a greater risk of seeing 'adult' content.

Data from Ofcom, 2024.

2. The revenue model of social media platforms incentivises their use of engagement based content recommendation algorithms.

Social media platforms charge a fee to show advertisements to their users, who typically do not need to pay a monetary fee to use the platform. Social media platforms often collect large amounts of data about their users and can therefore sell two models of reach; to a mass audience, and to a highly targeted audience (Sindermann et al., 2024). Advertisers (or sponsors) are seeking maximum connection with potential customers (or people whose views and choices they wish to influence). To maximise revenue, social media platforms optimise for maximal user attention, engagement, retention, and growth - thereby maximising users' exposure to ads over time.

This revenue model (user-related advertisement-based) rewards content that is highly engaging i.e. content that drives views, clicks, and shares. Algorithms are therefore designed to select and propagate content that is predicted to be highly engaging - regardless of its accuracy or potential to harm. These "engagement-based recommendation algorithms" are commonly used to generate social media feeds (Covington et al., 2016; Narayanan, 2023).

It is important to note that social media platforms are not the only actors deriving revenue from social media platforms. There are other actors such as content creators, online retailers, brand and affiliate marketers whose revenue models are also directly affected by what information is uploaded, popularised, and propagated on social media (Sindermann et al., 2024).

3. We focus on recommendation algorithms (content propagation) because platforms have almost complete control over what they recommend to a user.

Social media platforms may propagate information using three models; subscription, network, and algorithmic (Figure 1). In the subscription model, the post reaches those who have subscribed to the poster. In the network model, it moves through a network of followers as each chooses whether or not to share the post. In the algorithmic model, users with similar “interests” (as learned by the algorithm based on their past engagement) are more likely to be shown a post than those with different online behaviours. This is an important difference because the control over what you see (and stop seeing), shifts from the user (and their friends) to the platform. The result is that users have less choice over what they see, and are more likely to be exposed to content that they didn’t choose, from people that they do not know.

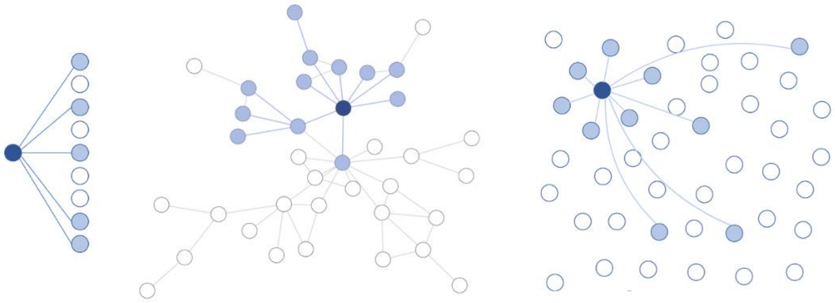

Figure 1: Reproduced from Narayanan, 2023. *Three models of information propagation: subscription, network, and algorithm, showing the propagation of one individual post.* *https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms*

4. Recommendation algorithms are designed to show users popular content. But what becomes popular is the results of users' collective online actions.

Social media platforms are complex systems where information spreads in unpredictable ways; content is propagated (or not) depending on dynamic interactions and feedback loops between platform design and human behaviour (Narayanan, 2023; Rodrigues et al., 2024)).

On the platform side, this includes things like:

- the recommendation algorithms that decide what content you see,
- content policies that determine what can be posted, and
- the overall design of the user interface.

On the human side, this involves;

- how users consume and share content,
- the behaviour of content creators, and
- the ways people interact with and respond to posts.

The spread of harmful content on social media is not inevitable. It arises from a combination of the social media business model, engagement-based recommendation algorithms, and the choices and actions of its users (Jafar et al., 2023; Narayanan, 2023).

5. The rapid amplification and repeated exposure to potentially harmful content is a particularly concerning feature of human interaction with engagement-based content recommendation algorithms.

There is evidence that action and engagement is driven more effectively by negative appeals than positive or coactive appeals (Yousef et al., 2021), that 'anger' travels through social media faster and further than 'joy' (Fan et al., 2020), and that fake information spreads faster than facts (Rodrigues et al., 2024). Potentially harmful content can reach and wrap-around users. It can reach users when it is disproportionately amplified across the platform. It can also 'wrap-around' users by being disproportionately presented to vulnerable users in the form of harmful content spirals (aka 'rabbit holes').

- Harmful content amplification: Sensational and controversial content such as misinformation, conspiracy theories, and divisive material tends to perform well under engagement-based recommendation systems, and so receive disproportionate amplification compared to health-promoting or neutral content (Rodrigues et al., 2024). Platforms that rely heavily on user-generated content can become conduits for the rapid spread of misinformation, as harmful content can go viral before it is identified and addressed. With billions of users worldwide, effectively monitoring and moderating content in all languages and contexts is a significant challenge.
- Harmful content spirals: The personalisation of content feed can reinforce users' pre-existing beliefs, amplifying misinformation and reducing exposure to diverse perspectives (Eg et al., 2023). Algorithmic recommendations can exacerbate the harm caused by harmful content by creating content spirals that expose users, who are already at a higher risk of harm, to increasingly extreme or harmful material (Stoilova et al., 2021).

6. There is growing evidence that exposure to online content translates to actual health harms at an individual and population level, with certain groups more vulnerable to harmful content exposure.

We highlight four areas of public health concern; the direct impact of harmful content on users' mental health, the promotion of harmful products, and indirect impacts on society from the normalisation of harmful social messages or behaviours, and the erosion of trust in scientific expertise and public institutions. These highlighted harms are exemplary, not exhaustive.

At the individual level, there is evidence that exposure to specific online content correlates directly with harm:

- Analysis of the UK Millennium Cohort of 10,904 14-year olds found an association between social media use and online harassment, poor sleep, low self-esteem and poor body image; which in turn related to higher depressive symptom scores (Kelly et al., 2018).
- Facebook's released internal research found that 13% of suicidal teenagers in Britain attributed their suicidal ideation to Instagram. They also found that a significant proportion of UK teens report negative feelings traceable to Instagram - 29% not feeling good enough, 18% feeling lonely, 43% not feeling attractive, and 51% feeling like they have to create the perfect image (Jafar et al., 2023).

- Exposure to content relating to depression and self-harm can worsen users' mental health(Kostyrka-Allchorne et al., 2023; Naslund et al., 2020; Stoilova et al., 2021; Susi et al., 2023)
- Exposure to misogynist, racist or homophobic content online is also associated with poor mental health outcomes, depression, anxiety, chronic stress and poor self-esteem (Keum & Choi, 2023; Keum & Miller, 2017; Tao & Fisher, 2022).
- A systematic review of social media use and wellbeing found no significant association between excessive social media use and subjective or psychological wellbeing, but a negative association between problematic social media use and subjective and psychological wellbeing (Ansari et al., 2024)

There is evidence that social media platforms and content creators promote harmful products to users, including children (WHO, 2019).

- Exposure may occur through direct advertising or indirect allusion to consumption and may act to increase the consumption of harmful products such as ultra-processed high fat, salt, sugar foods, sugar-sweetened beverages, alcohol, tobacco, vapes, and recreational drugs (Vannucci et al., 2020).
- A 2018 systematic review of the effects of digital marketing of unhealthy commodities on people aged 12 to 30 found that digital marketing enhanced attitudes, intention to use, and current use of harmful commodities, particularly alcohol. Peer social media had a greater impact than online advertising in general (Buchanan et al., 2018).
- In a Canadian study of influencers of children aged 10-12, the rate of unhealthy food marketing instances per post was 1 food marketing instance every 0.7 YouTube posts, 10.2 TikTok posts, and 19.3 Instagram posts (Potvin Kent et al., 2024).

There is evidence that some content may contribute to the normalisation of harmful social messages or behaviours.

- Evidence suggests that exposure to some types of content, particularly when unintentionally exposed as a result of algorithmic promotion, can lead to the normalisation, desensitisation or competition in risky behaviours, to learning new risk behaviours (for example, methods for self-harm), and fosters communities of harmful practices (Kostyrka-Allchorne et al., 2023; Susi et al., 2023)
- Social media content depicting eating disorders (aka 'thinspiration') or diet and exercise culture (including 'fitspiration') in promoting or exacerbating poor body image and disordered eating through negative social comparisons, thin / fit ideal internalisation, and self-objectification (Dane & Bhatia, 2023; Griffiths et al., 2018)
- Up to 40% of posts on social media contain health misinformation related to vaccinations, eating disorders, treatments, and chronic diseases including cancer. Accessing health-related misinformation could result in delaying treatment, or engaging in harmful, expensive, and futile therapies (Fridman et al., 2023)
- Online consumption of misinformation is associated with altered perceptions of issues, negative emotions, and undermining of prosocial behaviour when offline. This can result in the translation of cyberviolence to in-person violence, as felt by public health professionals during the COVID-19 pandemic who experienced harassment, abuse, and threats on social media, as well as direct physical threats and confrontation - which were glorified and amplified on social media. Harassment and

abuse were particularly virulent for public health professionals who were women or visible minority individuals (Regehr et al., 2024).

There is evidence that social media platforms can erode public trust in scientific expertise by amplifying misinformation.

- During the COVID-19 pandemic, for instance, there was significant spread of disinformation and misinformation via social media platforms, leading to fear, hysteria, confusion, and an erosion of confidence in public health measures, False claims on social media contributed to avoidable hospitalisations and deaths worldwide (Rodrigues et al., 2024).
- Anti-vaccination content and conspiracy theories led to vaccine hesitancy and undermined critical public health messaging, demonstrating how digital platforms can generate population-level health risks (Vidgen et al., 2021). A randomised controlled trial in the UK found that compared to seeing factual information, being shown misinformation decreased intention to receive the COVID-19 vaccine by 6.2% (Loomba et al., 2021).

Critically, these harmful impacts are not uniformly distributed. Vulnerable populations including young users, individuals with pre-existing mental health conditions, women, and marginalized ethnic or sexual identity groups, experience disproportionately higher exposure to and/or harms from harmful content (Abreu & Kenny, 2018; Kelly et al., 2018; Keum & Choi, 2023; Kostyrka-Allchorne et al., 2023; Winstone, 2024).

7. Social media platforms can be tools to support health, although the evidence currently skews towards negative rather than positive outcomes.

It is important to note that most online adults (67%) believe that for them, the benefits of being online outweigh the risks (Ofcom, 2024). From a public health perspective, social media platforms also have the potential to be tools for improving health (Jafar et al., 2023; Kanchan & Gaidhane, 2023; Kostyrka-Allchorne et al., 2023)

- Social media can be used to deliver timely public health messaging or interactive health programmes to support preventive health behaviours. Studies during COVID-19 found that social media users were more likely to be aware of methods to prevent disease spread, and more likely to follow pandemic health rules (Jafar et al., 2023).
- Social media can also have positive effects on mental health, social wellbeing, resilience, and loneliness by providing the opportunity for meaningful social interactions, to maintain social networks, and to build social capital (Jafar et al., 2023). A meta-analysis of 58 studies found that use of social networking sites was associated with higher social capital when used to facilitate contact and interaction among people who already know each other offline (Liu et al., 2016)
- Social media can be used to provide free access to interventions shown to improve health. For example, a 2021 Cochrane review found that programmes on social media may help people to become more physically active and may improve people’s wellbeing (Petkovic et al., 2021).

However, these positive outcomes are undermined when the overarching incentives of the current digital ecosystem prioritises engagement over content quality.

8. A public health approach to mitigating the risk of harm from online content must be taken.

The evidence on the health and social impacts of social media is mixed, and further research is needed to fully characterise the mechanism and extent of the effects of different types of content for different groups of users. However, the need to generate more robust evidence does not outweigh the need for action. Our judgement is that the growing evidence of health and social harm resulting from the amplification of potentially harmful content by 'engagement-based recommendation algorithms' is sufficient to recommend that a public health approach to mitigating the risk of harm from online content must be taken.

What role do generative artificial intelligence (AI) and large language models (LLMs) play in the creation and spread of misinformation, disinformation and harmful content?

9. Generative Artificial Intelligence (GenAI) and large language models (LLMs) could be a powerful public health tool but as they become used more widely, in an unregulated context, they also pose a significant threat in propagating misinformation

Creation and Amplification of Harmful Content

GenAI enables the rapid creation of harmful content, including misinformation, deepfakes, and biased narratives (Sippy et al., 2024). For example, LLMs have demonstrated the ability to generate highly convincing disinformation for political manipulation (Hackenburg et al., 2024; A. R. Williams et al., 2024). There is no reason that similar tactics won't be used in health. This technology is also leveraged to produce derivative harmful content that mirrors and amplifies existing sensational material, compounding its reach and impact. This is more so where there are swarms of interest and advertising opportunities around that content. Although GenAI can be used by botnets to spread counter-information that is shaped for maximum accessibility by hyperlocal (place/other) communities, botnets linked to GenAI are also used to propagate harmful (mis)information for malevolent purposes online (A. Williams, 2024; Yang & Menczer, 2024).

Bias and Targeting

AI systems are susceptible to embedding and amplifying biases, which can disproportionately target marginalised groups, eroding trust, and exacerbating social inequalities. Algorithms may disproportionately expose vulnerable populations to harmful content, perpetuating cycles of harm and mistrust (Capraro et al., 2024; European Union Agency for Fundamental Rights, 2022).

Opportunities for Countermeasures

Despite these threats, GenAI also offers opportunities for mitigating harm. AI-driven tools can enhance public health efforts by identifying and countering biases, disseminating accurate information tailored to specific communities, and fostering digital resilience (Vicari & Komendatova, 2023). There is scope for security services and public health agencies to collaborate more at national level on agent-based approaches to health protecting and promoting GenAI and countermeasures linked to national security technologies for tackling misinformation, such as botnet neutralisation (World Economic Forum, 2024).

However, a proactive regulatory approach by the government is essential to optimise the potential of these technologies without exposing the population to harm. At a minimum, greater surveillance in this area and funding for research to improve understanding should be a priority.

What more should be done to combat potentially harmful social media and AI content?

It is time to act to reduce the potential for social media to cause harm to individuals and to society. Action can be taken to interrupt the pathway of harm at three places; the platform, the content, and the individual.

- Actions at the platform level should target the propagation of harmful content by engagement-based recommendation algorithms.
- Actions at the content level should aim to reduce the prevalence of potentially harmful content by preventing it from being uploaded to platforms, and by increasing removal of harmful content from platforms.
- Actions at the individual level should aim to alter users' behaviour on or use of social media.

10. The British public strongly welcome action to tackle potentially harmful content.

The need to act is understood and supported by the British public - a large-scale, nationally representative survey of 2000 adults in the UK found that participants, across all demographic groups, were strongly in favour of action against online harms - less than 1% thought that there should be no consequences for platforms that fail to deal with harmful content (The Alan Turing Institute, 2023).

- 79% thought social media platforms should ban or suspend users who create harmful content.
- 73% thought social media platforms should remove harmful content.
- 69% thought that it should be easier to report harmful content to the platform.
- 68% thought paid promotion of harmful content should be stopped.
- 52% thought harmful content should be made more difficult to find.
- 72% thought the government should issue large fines to social media platforms that fail to deal with harmful content, while 66% were in favour of launching legal proceedings against these platforms, and 63% in favour of publicly naming and shaming them.
- 60% thought platforms should be forced to systematically report how much harmful content they host and how they are attempting to combat it.
- 54% thought platforms should be forced to introduce strict age verification processes.

11. The UK government should support evidence-based interventions to protect individuals from potentially harmful online content.

A recent review from Johansson and colleagues (2022) synthesises the current approaches to tackling online misinformation and their efficacy (Johansson et al., 2022). The UK should implement targeted public health interventions that balance innovation with the imperative to protect individuals from harm.

Enhancing digital and algorithmic literacy among the public will be a critical component of empowering individuals to safely consume social media and navigate misinformation and harmful content (Winstone, 2024). Digital literacy was found by an Alan Turing Institute analysis to reduce susceptibility to misinformation, even after controlling for sociodemographic and socioeconomic traits. It recommends exploring digital literacy as an immediate priority (Vidgen et al., 2021).

There is some evidence for interventions designed to curb the spread of misinformation such as fact checking labels and accuracy prompts, as well as correcting misleading claims with an alternative narrative, especially corrections from a trusted source (Enock et al., 2024).

Facilitating cross-sector collaboration and research is critical to addressing the systemic drivers of digital harms. Open data sharing agreements would enable robust analysis of algorithmic risks and foster transparency, while joint initiatives could accelerate the development of tools to monitor and mitigate harmful content. For instance, collaborative efforts could focus on designing AI-powered systems to detect and flag misleading content, coupled with evidence-based counter-information campaigns to correct false narratives before they spread widely.

Investment in further research will be crucial to understand the nuanced relationship between algorithmic content promotion and health outcomes, and the impact of exposure to harmful content on health inequalities. Research should prioritise preventive measures, such as the development of algorithms that promote health-positive content, as well as harm reduction strategies that limit user exposure to damaging material.

12. The UK government has a duty to protect and promote human health and therefore must strengthen regulation of social media platforms in line with other international examples.

Regulatory frameworks should be expanded to include obligations for platforms to protect and promote public health. The UK regulations can be further bolstered by existing regulations proposed globally; WHO report on Ethics and Governance of Artificial Intelligence for Health (WHO, 2024), moreover, the EU report on Artificial Intelligence in Healthcare (European Parliament, 2022). Both these reports offer relevant frameworks for addressing the impact of social media on population health and well-being.

While the WHO report focuses on large multi-modal models (LMM) of AI in healthcare, it does highlight critical upstream ethical considerations and guidance of the challenges of social media misinformation and harmful algorithms that can be applied to non-healthcare social media related fields.

We urge consideration of the following recommendations from these reports to be applied to social media developers/companies (WHO pages 12-22, EU pages 3-52):

- **Protect human autonomy and safety:** Implement safeguards against AI-driven manipulation and harmful content, ensuring human oversight and accountability for algorithmic decisions that impact public discourse and individual safety.
- **Promote human well-being, safety and public interest:** Regulate social media algorithms to prioritise public health and safety over profit maximisation, minimising the spread of misinformation and harmful content that can negatively impact mental and physical health.
- **Ensure transparency, explainability and intelligibility:** Advocate for transparency from social media companies regarding their algorithms, enabling public

understanding of how information is filtered and presented, and how this can influence opinions and behaviours.

- **Foster responsibility and accountability:** Establish clear accountability mechanisms for social media platforms regarding the impact of their algorithms on society, including their role in spreading misinformation and amplifying harmful content.
- **Ensure inclusiveness and equity:** Mitigate algorithmic biases that can perpetuate discrimination and inequality, ensuring fair and equitable access to information and opportunities for all social groups.
- **Promote AI that is responsive and sustainable:** Encourage the development of social media algorithms that are responsive to societal needs and evolving challenges, contributing to a healthy and sustainable information ecosystem.

Both WHO and EU recommendations indicate shared concerns between healthcare and social media, emphasising the need for:

- Human accountability and regulation to mitigate risks and ensure ethical social media implementation.
- Foster trust and understanding of how the algorithms function.
- Continuous monitoring and evaluation of systems to address biases, safety issues and the ever-changing societal needs.

International examples, such as the WHO framework on AI ethics, can inform these efforts. Transparency in algorithmic decision-making and content prioritisation processes must be mandated to ensure accountability. By adhering to these WHO guidelines, the government can foster responsible social media ecosystems that will positively impact population health and well-being, maximizing benefits while mitigating ethical risks and ensuring public digital literacy, health and safety.

Several countries have implemented social media restrictions in response to concerns about its harmful impacts, signalling a growing recognition of the risks posed by unregulated digital environments. Multiple EU member state countries have implemented their own laws relating to social media misinformation/disinformation. These include Germany`s Network Enforcement Act (NetzDG; enacted 30. June 2017) and France's law against the manipulation of information (2018-1202 law; enacted 22. December 2018), coupled together by the 2022 EU Code of Practice on Disinformation and the other initiatives of the European Digital Media Observatory (EDMO), with EU countries participating in such initiatives (European Commission, 2024).

The anticipated outcomes include accurate health related information and regaining trust in public health institutions. Australia’s vote to ban social media for children under-16 was a threshold moment for policymakers. While it is too early to fully understand the long-term effects of these measures, they represent an important opportunity to generate evidence on the potential health benefits of stronger regulations.

The actions of these nations underscore the seriousness with which governments worldwide are addressing the health impacts of social media, serving as a critical signal that the UK must also consider significant and proactive measures to protect public health.

References

1. Abreu, R. L., & Kenny, M. C. (2018). Cyberbullying and LGBTQ Youth: A Systematic Literature Review and Recommendations for Prevention and Intervention. *Journal of Child & Adolescent Trauma*, *11*(1), 81–97. https://doi.org/10.1007/s40653-017-0175-7
2. Ansari, S., Iqbal, N., Asif, R., Hashim, M., Farooqi, S. R., & Alimoradi, Z. (2024). Social Media Use and Well-Being: A Systematic Review and Meta-Analysis. *Cyberpsychology, Behavior and Social Networking*, *27*(10), 704–719. https://doi.org/10.1089/cyber.2024.0001
3. Buchanan, L., Kelly, B., Yeatman, H., & Kariippanon, K. (2018). The Effects of Digital Marketing of Unhealthy Commodities on Young People: A Systematic Review. *Nutrients*, *10*(2), 148. https://doi.org/10.3390/nu10020148
4. Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A. C., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., … Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS Nexus*, *3*(6), pgae191. https://doi.org/10.1093/pnasnexus/pgae191
5. Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*, 191–198. https://doi.org/10.1145/2959100.2959190
6. Dane, A., & Bhatia, K. (2023). The social media diet: A scoping review to investigate the association between social media, body image and eating disorders amongst young people. *PLOS Global Public Health*, *3*(3), e0001091. https://doi.org/10.1371/journal.pgph.0001091
7. Eg, R., Demirkol Tønnesen, Ö., & Tennfjord, M. K. (2023). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, *9*, 100253. https://doi.org/10.1016/j.chbr.2022.100253
8. Enock, F. E., Bright, J., Stevens, F., Johansson, P., & Margetts, H. Z. (2024). *How do people protect themselves against online misinformation? Attitudes, experiences and uptake of interventions amongst the UK adult population* (SSRN Scholarly Paper No. 4918950). Social Science Research Network. https://doi.org/10.2139/ssrn.4918950
9. European Commission. (2024, October 22). *The 2022 Code of Practice on Disinformation | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
10. European Parliament. (2022). *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts.* Publications Office. https://data.europa.eu/doi/10.2861/568473
11. European Union Agency for Fundamental Rights. (2022). *Bias in algorithms: Artificial intelligence and discrimination*. Publications Office of the European Union. https://data.europa.eu/doi/10.2811/25847
12. Fan, R., Xu, K., & Zhao, J. (2020). *Weak ties strengthen anger contagion in social media* (No. arXiv:2005.01924). arXiv. https://doi.org/10.48550/arXiv.2005.01924
13. Fridman, I., Johnson, S., & Elston Lafata, J. (2023). Health Information and Misinformation: A Framework to Guide Research and Practice. *JMIR Medical Education*, *9*, e38687. https://doi.org/10.2196/38687
14. Griffiths, S., Murray, S. B., Krug, I., & McLean, S. A. (2018). The Contribution of Social Media to Body Dissatisfaction, Eating Disorder Symptoms, and Anabolic Steroid Use Among Sexual Minority Men. *Cyberpsychology, Behavior and Social Networking*, *21*(3), 149–156. https://doi.org/10.1089/cyber.2017.0375

15. Hackenburg, K., Tappin, B. M., Röttger, P., Hale, S., Bright, J., & Margetts, H. (2024). *Evidence of a log scaling law for political persuasion with large language models* (No. arXiv:2406.14508). arXiv. https://doi.org/10.48550/arXiv.2406.14508
16. Jafar, Z., Quick, J. D., Larson, H. J., Venegas-Vera, V., Napoli, P., Musuka, G., Dzinamarira, T., Meena, K. S., Kanmani, T. R., & Rimányi, E. (2023). Social media for public health: Reaping the benefits, mitigating the harms. *Health Promotion Perspectives*, *13*(2), 105–112. https://doi.org/10.34172/hpp.2023.13
17. Johansson, P., Enock, F., Hale, S., Vidgen, B., Bereskin, C., Margetts, H., & Bright, J. (2022). *How can we combat online misinformation? A systematic overview of current interventions and their efficacy* (No. arXiv:2212.11864). arXiv. https://doi.org/10.48550/arXiv.2212.11864
18. Kanchan, S., & Gaidhane, A. (2023). Social Media Role and Its Impact on Public Health: A Narrative Review. *Cureus*, *15*(1), e33737. https://doi.org/10.7759/cureus.33737
19. Kelly, Y., Zilanawala, A., Booker, C., & Sacker, A. (2018). Social Media Use and Adolescent Mental Health: Findings From the UK Millennium Cohort Study. *eClinicalMedicine*, *6*, 59–68. https://doi.org/10.1016/j.eclinm.2018.12.005
20. Keum, B. T., & Choi, A. Y. (2023). Profiles of online racism exposure and mental health among Asian, Black, and Latinx emerging adults in the United States. *International Review of Psychiatry (Abingdon, England)*, *35*(3–4), 310–322. https://doi.org/10.1080/09540261.2023.2180346
21. Keum, B. T., & Miller, M. J. (2017). Racism in digital era: Development and initial validation of the Perceived Online Racism Scale (PORS v1.0). *Journal of Counseling Psychology*, *64*(3), 310–324. https://doi.org/10.1037/cou0000205
22. Kostyrka-Allchorne, K., Stoilova, M., Bourgaize, J., Rahali, M., Livingstone, S., & Sonuga-Barke, E. (2023). Review: Digital experiences and their impact on the lives of adolescents with pre-existing anxiety, depression, eating and nonsuicidal self-injury conditions – a systematic review. *Child and Adolescent Mental Health*, *28*(1), 22–32. https://doi.org/10.1111/camh.12619
23. Liu, D., Ainsworth, S. E., & Baumeister, R. F. (2016). A Meta-Analysis of Social Networking Online and Social Capital. *Review of General Psychology*, *20*(4), 369–391. https://doi.org/10.1037/gpr0000091
24. Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1
25. Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms*. Knight First Amendment Institute. http://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms
26. Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. *Journal of Technology in Behavioral Science*, *5*(3), 245–257. https://doi.org/10.1007/s41347-020-00134-x
27. Ofcom. (2024). *Online Nation 2024 report*.
28. Petkovic, J., Duench, S., Trawin, J., Dewidar, O., Pardo, J. P., Simeon, R., DesMeules, M., Gagnon, D., Roberts, J. H., Hossain, A., Pottie, K., Rader, T., Tugwell, P., Yoganathan, M., Presseau[a], J., & Welch[a], V. (2021). *Behavioural interventions delivered through interactive social media for health behaviour change, health outcomes, and health equity in the adult population—Petkovic, J - 2021 | Cochrane Library*. https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD012932.pub2/inform

ation

29. Potvin Kent, M., Bagnato, M., Amson, A., Remedios, L., Pritchard, M., Sabir, S., Gillis, G., Pauzé, E., Vanderlee, L., White, C., & Hammond, D. (2024). #junkfluenced: The marketing of unhealthy food and beverages by social media influencers popular with Canadian children on YouTube, Instagram and TikTok. *The International Journal of Behavioral Nutrition and Physical Activity*, *21*(1), 37. https://doi.org/10.1186/s12966-024-01589-4
30. Regehr, K., Regehr, C., Goel, V., Sato, C., Lyons, K., & Rudzicz, F. (2024). From the Screen to the Streets: Technology-Facilitated Violence Against Public Health Professionals. *Journal of Loss and Trauma*, *0*(0), 1–26. https://doi.org/10.1080/15325024.2024.2406509
31. Rodrigues, F., Newell, R., Rathnaiah Babu, G., Chatterjee, T., Sandhu, N. K., & Gupta, L. (2024). The social media Infodemic of health-related misinformation and technical solutions. *Health Policy and Technology*, *13*(2), 100846. https://doi.org/10.1016/j.hlpt.2024.100846
32. Sindermann, C., Scholz, R. W., Löchner, N., Heinzelmann, R., & Montag, C. (2024). The Revenue Model of Mainstream Social Media: Advancing Discussions on Social Media Based on a European Perspective Derived from Interviews with Scientific and Practical Experts. *International Journal of Human–Computer Interaction*, *40*(23), 8107–8123. https://doi.org/10.1080/10447318.2023.2278292
33. Sippy, T., Enock, F., Bright, J., & Margetts, H. Z. (2024). *Behind the Deepfake: 8% Create; 90% Concerned. Surveying public exposure to and perceptions of deepfakes in the UK* (No. arXiv:2407.05529). arXiv. https://doi.org/10.48550/arXiv.2407.05529
34. Stoilova, M., Edwards, C., Kostyrka-Allchorne, K., Livingstone, S., & Sonuga-Barke, E. (2021). *Adolescents' mental health vulnerabilities and the experience and impact of digital technologies: A multimethod pilot study*. London School of Economics and Political Science and King's College London, London, UK. https://doi.org/10.18742/pub01-073
35. Susi, K., Glover-Ford, F., Stewart, A., Knowles Bevis, R., & Hawton, K. (2023). Research Review: Viewing self-harm images on the internet and social media platforms: systematic review of the impact and associated psychological mechanisms. *Journal of Child Psychology and Psychiatry*, *64*(8), 1115–1139. https://doi.org/10.1111/jcpp.13754
36. Tao, X., & Fisher, C. B. (2022). Exposure to social media racial discrimination and mental health among adolescents of color. *Journal of Youth and Adolescence*, *51*(1), 30–44. https://doi.org/10.1007/s10964-021-01514-z
37. The Alan Turing Institute. (2023). *Tracking experiences of online harms and attitudes towards online safety interventions*. The Alan Turing Institute. https://www.turing.ac.uk/news/publications/experiences-online-harms
38. Vannucci, A., Simpson, E. G., Gagnon, S., & Ohannessian, C. M. (2020). Social media use and risky behaviors in adolescents: A meta-analysis. *Journal of Adolescence*, *79*, 258–274. https://doi.org/10.1016/j.adolescence.2020.01.014
39. Vicari, R., & Komendatova, N. (2023). Systematic meta-analysis of research on AI tools to deal with misinformation on social media during natural and anthropogenic hazards and disasters. *Humanities and Social Sciences Communications*, *10*(1), 1–14. https://doi.org/10.1057/s41599-023-01838-0
40. Vidgen, B., Taylor, H., Pantazi, M., Anastasiou, Z., Inkster, B., & Margetts, H. (2021). *Understanding vulnerability to online misinformation*.
41. WHO. (2019). *Monitoring and restricting digital marketing of unhealthy products to children and adolescents: Report based on the expert meeting on monitoring of digital marketing of unhealthy products to children and adolescents: Moscow,*

Russian Federation, June 2018 (No. WHO/EURO:2019-3592-43351-60815). Article WHO/EURO:2019-3592-43351-60815. https://iris.who.int/handle/10665/346585

42. WHO. (2024). *Ethics and Governance of Artificial Intelligence for Health: Large Multi-Modal Models. WHO Guidance* (1st ed). World Health Organization.
43. Williams, A. (2024). *Online misinformation: How generative AI and LLMs are changing the game*. The Alan Turing Institute. https://www.turing.ac.uk/blog/online-misinformation-how-generative-ai-and-llms-are-changing-game
44. Williams, A. R., Burke-Moore, L., Chan, R. S.-Y., Enock, F. E., Nanni, F., Sippy, T., Chung, Y.-L., Gabasova, E., Hackenburg, K., & Bright, J. (2024). *Large language models can consistently generate high-quality content for election disinformation operations* (No. arXiv:2408.06731). arXiv. https://doi.org/10.48550/arXiv.2408.06731
45. Winstone. (2024). *Churchill Fellowship: Teaching algorithmic literacy to young people to promote positive social media use*. University of Bristol. https://research-information.bris.ac.uk/en/projects/churchill-fellowship-teaching-algorithmic-literacy-to-young-peopl
46. World Economic Forum. (2024, June 14). *How AI can also be used to combat online disinformation*. World Economic Forum. https://www.weforum.org/stories/2024/06/ai-combat-online-misinformation-disinformation/
47. Yang, K.-C., & Menczer, F. (2024). Anatomy of an AI-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, *4*. https://doi.org/10.51685/jqd.2024.icwsm.7
48. Yousef, M., Dietrich, T., & Rundle-Thiele, S. (2021). Social Advertising Effectiveness in Driving Action: A Study of Positive, Negative and Coactive Appeals on Social Media. *International Journal of Environmental Research and Public Health*, *18*(11), 5954. https://doi.org/10.3390/ijerph18115954