REPHRAIN

Protecting citizens online

Call for evidence: Misinformation and trusted voices

This is a submission from the REPHRAIN centre. Specifically, the following researchers contributed to the formulation of this response (in alphabetical order): Prof Madeline Carr, Dr Andrés Domínguez, Dr Matthew Edwards, Dr Ola Michalec, Prof Awais Rashid, Yvonne Rigby, Dr Adam Sutton.

September 2022

UK Research and Innovation · University of Bristol · UCL · The University of Edinburgh · King's College London · University of Bath

Call for evidence: Misinformation and trusted voices

Introduction

Thank you for an opportunity to provide our response to this call for evidence. We are writing on behalf of REPHRAIN, the National **Re**search Centre on **P**rivacy, **H**arm **R**eduction and **A**dversarial **I**nfluence O**n**line. REPHRAIN is the UK's world-leading interdisciplinary community focused on the protection of citizens online. As a UKRI-funded National Research Centre, we boast a critical mass of over 100 internationally leading experts at 13 UK institutions working across 37 diverse research projects and 23 founding industry, non-profit, government, law, regulation and international research centre partners. As an interdisciplinary and engaged research group, we work collaboratively on addressing the three following missions:

- Delivering privacy at scale while mitigating its misuse to inflict harms
- Minimising harms while maximising benefits from a sharing-driven digital economy
- Balancing individual agency vs. social good.

We are addressing this consultation since our researchers have extensive expertise in d/misinformation, trust, content moderation and cognate areas. This is a submission from the REPHRAIN centre. Specifically, the following researchers contributed to the formulation of this response (in alphabetical order): Prof Madeline Carr, Dr Andrés Domínguez, Dr Matthew Edwards, Dr Ola Michalec, Prof Awais Rashid, Yvonne Rigby, Dr Adam Sutton. We are happy to arrange a follow up meeting to provide details of our work in progress in the area.

1. **Where do you seek authoritative information to make up your mind about matters of national debate (such as vaccines and climate change)?**

The authoritativeness of sources is shaped by many social factors such as credentialled expertise, reputation, historical context, public understanding of science and social acceptance. There is a large body of research in science and technology studies around knowledge production and science legitimacy in the post truth era (e.g., Collins et al 2017; Grundmann 2017; Sismondo, 2017).

For matters of national debate, we expect knowledge sources to be sufficiently independent for them to be credible. In the case of public health and climate change, for example, we recommend seeking expertise from the Academy and independent bodies. Some examples of authoritative sources are: The Lancet, BMJ, WHO (public health). IPCC (Climate change).

We also stress that in the case of scientific knowledge production, it is important to look for comprehensive and multidisciplinary reviews of the evidence from high-quality and impartial scientific bodies, such as major scientific journals.

Finally, it is important to distinguish between m/disinformation (which can be counteracted with moderation and reliance of authoritative sources, I.e., a claim that climate change is caused by sunspots) and complex and multifaceted debates (I.e., implementing effective climate action in particular areas while balancing it with justice concerns and available budget). In the second case, 'making up one's mind' should not necessarily be the goal due to the contingent and

dynamic nature of such debates. We, therefore, recommend curating discussion environments, where people can be challenged without aggression and where they can afford to change their minds without shame. These environments ought to embrace plurality, humility and opening up expertise to a wide range of stakeholders who would learn from each other's inherently partial perspectives (Leach et al., 2010; Haraway, 1988).

2. **Are you able to "do your own research" on matters of national debate?**

We highly support that members of the public should be able to 'do their own research' - sieve through numerous publications, compare differing views or seek answers through multiple means. A suggestion that the above should not be possible is profoundly anti-scientific and anti-democratic.

3. **Is the provision of authoritative information responsive enough to meet the challenge of misinformation that is spread on social media?**

Authoritative information is one of the key resources to combat the spread of misinformation on social media. One way this is done is by ML algorithms which are trained on authoritative corpus of data and either help human moderation or automate decisions around banning or downranking (see here https://www.rephrain.ac.uk/clariti/). We caution however that there are risks with the use of automation as ML models are susceptible not only to errors and uncertainty, but to implicit assumptions around the authoritativeness of knowledge sources. This is particularly the case when these systems are black boxed by technology companies. Oversight and human intervention are needed to avoid problems such as falsely categorising misinformation which could undermine people's trust in public information or reinforce false beliefs. There is currently limited scrutiny or self-reporting of how platforms use these tools and the design decisions that underpin them. We analysed the above issues in a recent study and are happy to share a copy of the manuscript (e.g., through a follow up email or meeting).

For the readers interested in the process of building ML to detect misinformation, we published a paper (available here: https://cpb-eu-w2.wpmucdn.com/blogs.bristol.ac.uk/dist/1/670/files/2022/06/CLARITI-MuMiN-paper-June-2022.pdf) where we released and described the dataset that we built to train a ML model to verify social media claims. The dataset contains a rich variety of social media information (tweets, replies, users, images, articles, hashtags), spans 21 million tweets belonging to 26 thousand Twitter threads, each of which have been semantically linked to 13 thousand fact-checked claims across dozens of topics, events and domains, in 41 different languages, spanning more than a decade. The dataset itself can be accessed here https://mumin-dataset.github.io/ for those who would like to continue research in that area.

Other comments

We would like to highlight the following work of REPHRAIN members in the area of misinformation and trusted voices

- Alexandros Efstratiou and Emiliano De Cristofaro identify links between misinformation and polarisation on social media. They argue that when misinformation proliferates, this

happens because the social media environment enables adherence to misinformation by reducing, rather than increasing, the psychological cost of doing so. Source: https://arxiv.org/abs/2206.15237 (pre-print submitted for peer review)

- Papadamou et al characterise and detect pseudoscientific misinformation on YouTube based on 6.6K videos related to COVID-19, the Flat Earth theory, as well as the anti-vaccination and anti-mask movements. Using crowdsourcing, they annotated them as pseudoscience, legitimate science, or irrelevant and trained a deep learning classifier to detect pseudoscientific videos with an accuracy of 0.79. They quantified user exposure to this content on various parts of the platform and how this exposure changes based on the user's watch history. They found that YouTube suggests more pseudoscientific content regarding traditional pseudoscientific topics (e.g., flat earth, anti-vaccination) than for emerging ones (like COVID-19). At the same time, these recommendations are more common on the search results page than on a user's homepage or in the recommendation section when actively watching videos. Finally, the paper sheds light on how a user's watch history substantially affects the type of recommended videos. Source: https://ojs.aaai.org/index.php/ICWSM/article/view/19329
- Paudel et al designed a multi-platform computational pipeline geared to identify social media posts discussing (known) conspiracy theories. In their research, they used 189 conspiracy claims collected by Snopes, 66k posts and 277k comments on Reddit, and 379k tweets discussing them. Then, they studied how conspiracies are discussed on different Web communities and which ones are particularly influential in driving the discussion about them. The analysis sheds light on how conspiracy theories are discussed and spread online, while highlighting multiple challenges in mitigating them. Source: https://arxiv.org/abs/2111.02187 (pre-print submitted for peer review)
- In NEWS project (https://www.rephrain.ac.uk/news/), researchers are investigating the potential for predicting personality from news consumption. Personality information can be used as part of adversarial manipulation of an individual, making them more vulnerable to attacks on their reasoning process – whether the adversary here is a criminal phishing for card details or a marketing company trying to sell a product of dubious quality (see a case of Cambridge Analytica). REPHRAIN researchers are currently building a model to uncover a microtargeting algorithm in action. This work will feed into the design of privacy enhancing interventions that can detect when political message content is matched for consumption by particular personalities and inform users when material they are reading is suspiciously tailored to their own personality. We are happy to provide further information on request.
- Finally, in MANIPU project (https://www.rephrain.ac.uk/manipu/), researchers are working toward a philosophical analysis of manipulation on social media, using a case study of interventions into democratic elections. We are happy to discuss the work in progress and release further information in due course.

References

Collins, Harry, Robert Evans, and Martin Weinel. 2017. ‘STS as Science or Politics?’ *Social Studies of Science* 47 (4). SAGE Publications Ltd: 580–86. doi:10.1177/0306312717710131.

Efstratiou, Alexandros, and Emiliano De Cristofaro. "Adherence to Misinformation on Social Media Through Socio-Cognitive and Group-Based Processes." arXiv preprint arXiv:2206.15237 (2022).

Grundmann, Reiner. 2017. ‘The Problem of Expertise in Knowledge Societies’. *Minerva* 55 (1): 25–48. doi:10.1007/s11024-016-9308-7.

Haraway, Donna. 1988. "Situated knowledges: The science question in feminism and the privilege of partial perspective." Feminist theory reader. Routledge. 303-310.

Leach, Melissa, Ian Scoones, and Andrew Stirling. 2010. "Governing epidemics in an age of complexity: Narratives, politics and pathways to sustainability." Global Environmental Change 20, no. 3 (2010): 369-377.

Nielsen, Dan S. and Ryan McConville. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’22), July 11–15, 2022, Madrid, Spain. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3477495. 3531744

Papadamou, K., Zannettou, S., Blackburn, J., Cristofaro, E. D., Stringhini, G., & Sirivianos, M. (2022). “It Is Just a Flu”: Assessing the Effect of Watch History on YouTube’s Pseudoscientific Video Recommendations. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 723-734. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/19329

Paudel, P., Blackburn, J., De Cristofaro, E., Zannettou, S., & Stringhini, G. (2021). Soros, child sacrifices, and 5G: understanding the spread of conspiracy theories on web communities. arXiv preprint arXiv:2111.02187.

Sismondo, Sergio. 2017. ‘Post-Truth?’ *Social Studies of Science* 47 (1). SAGE Publications Ltd: 3–6. doi:10.1177/0306312717692076.