Response to the Royal Society Call for Evidence:

Technologies for Spreading and Detecting Misinformation

Submitted by Prof. Derek McAuley, Dr. Ansgar Koene and Dr. Jiahong Chen of Horizon Digital Economy Research Institute, University of Nottingham

15 August 2020

1. Horizon[1] is a Research Institute centred at The University of Nottingham and a Research Hub within the UKRI Digital Economy programme[2]. Horizon brings together researchers from a broad range of disciplines to investigate the opportunities and challenges arising from the increased use of digital technology in our everyday lives. Prof. McAuley is Director of Horizon and Principal Investigator of the EPSRC-funded DADA[3] (Defence Against Dark Artefacts) project, addressing smart home IoT network security, and its acceptability and usability issues, the ESRC-funded CaSMa[4] (Citizen-centric approaches to Social Media analysis) project to promote ways for individuals to control their data and online privacy, and the EPSRC-funded UnBias[5] (Emancipating Users Against Algorithmic Biases for a Trusted Digital Economy) project for raising user awareness and agency when using algorithmic services. Dr Koene was a lead researcher of the CaSMa and UnBias projects, is Research co-Investigator on the EPSRC-funded ReEnTrust[6] (Rebuilding and Enhancing Trust in Algorithms) project and chairs the working group for developing the IEEE P7003 Standard for Algorithm Bias Considerations. Dr Jiahong Chen is a Researcher Fellow of Horizon, working on the DADA project and a book project based on his doctoral research on regulating online advertising.

Propagation and impacts of misinformation

Q1. What impact have digital technologies had on patterns of information consumption? What evidence exists on their wider social impact?

2. The prevalent business model of major sources of online information, including news outlets, blogs, and social media, depends heavily on monetisation of granular user profiles. The "ad tech" industry has developed advanced technologies, such as programmatic trading and universal IDs, to target internet users with highly personalised content. This may change the patterns of information consumption in several ways: First, internet users may find themselves trapped in their own "echo chambers" with a feedback loop of information from like-minded groups, which may intensify the polarisation of the society.[7] Second, political campaigners may gain unfair advantage by exploiting

[1] http://www.horizon.ac.uk

[2] https://epsrc.ukri.org/research/ourportfolio/themes/digitaleconomy/

[3] https://www.horizon.ac.uk/project/defence-against-dark-artefacts/

[4] http://casma.wp.horizon.ac.uk

[5] http://unbias.wp.horizon.ac.uk

[6] https://ReEnTrust.org

[7] Tien T. Nguyen et al. "Exploring the filter bubble: the effect of using recommender systems on content diversity." *Proceedings of the 23rd international conference on World wide web*. 2014; Haim, Mario, Andreas Graefe, and Hans-Bernd Brosius. "Burst of the filter bubble? Effects of personalization on the diversity of Google News." *Digital journalism* 6.3 (2018): 330-343.

online personal data, as shown by the ICO's investigation into the Cambridge Analytica scandal.[8] Third, conspiracy and pseudoscientific theorists may find it easier to channel their messages to potentially more susceptible audiences, which is evidenced by, for example, the revelation about the possibility to target social media users based on a "vaccine controversies" category.[9]

Q2. How do digital technologies contribute to the spread of misinformation?

3. Many of the new technological phenomena are "neutral" in the sense that they can facilitate the dissemination of information regardless of the nature of such information. Social media, for example, have significantly augmented individuals' ability to create and share information, a large part of which, however, can be misinformation. The rise of marketing by online influencers has further contributed to the spread of misleading or mistaken information.[10] Deepfake is another controversial area where synthetic videos can be created to help circulate false news.[11]

Q3. What tools exist to create synthetic text, audio or visual media, and what are the likely near-term future developments of these technologies?

4. Machine learning has been applied to create machine-generated content, including misinformation. Text-based applications, such as natural language generation, for example, have been proved capable of fabricating convincing news stories.[12] Image- and video-based applications are also widely used in context less associated with the spreading of misinformation, such as smartphone camera filters, but have also raised concerns about user privacy, dignity and mental health.[13] These technologies can be easily repurposed for conducting political or personal attacks.[14]

Detection and tracing of misinformation

Q4. Which technologies can currently be used to identify or trace misinformation? What are the strengths and weaknesses of these technologies?

5. Current approaches to detecting misinformation can be largely categorised as text-based, image-based and profiled-based. Text-based strategies identify misinformation by analysing lexical, semantic and statistical features.[15] Image-based solutions identify deepfakes by examining biological signals[16] or inter-frame dissimilarities.[17] Profile-based approaches identify misinformation spreaders

[8] https://ico.org.uk/media/action-weve-taken/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf

[9] https://www.theguardian.com/technology/2019/feb/15/facebook-anti-vaccination-advertising-targeting-controversy

[10] Catalina Goanta and Gerasimos Spanakis. "Influencers and Social Media Recommender Systems: Unfair Commercial Practices in EU and US Law." (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3592000

[11] https://www.bbc.co.uk/news/business-51204954

[12] https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction

[13] https://www.nbcnews.com/tech/security/face-swapping-app-takes-china-making-ai-powered-deepfakes-everyone-n1049501; https://www.bbc.co.uk/news/business-50152053; https://www.theguardian.com/commentisfree/2019/jun/29/deepnude-app-week-in-patriarchy-women

[14] https://www.fireeye.com/blog/threat-research/2020/02/information-operations-fabricated-personas-to-promote-iranian-interests.html; https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell

[15] Dinesh Kumar Vishwakarma and Chhavi Jain. "Recent State-of-the-art of Fake News Detection: A Review." *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020.

[16] Umur Aybars Ciftci, Ilke Demir and Lijun Yin. "Fakecatcher: Detection of synthetic portrait videos using biological signals." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[17] Irene Amerini et al. "Deepfake video detection through optical flow based CNN." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.

by distinguishing abnormal user behavioural patterns.[18]

Q5. What technological advancements in the next ten years could improve the ability to identify or trace misinformation?

6. Provenance has been held out as a solution to this problem as a means to clearly identify trustworthy information, and hence, by implication, everything else should be untrusted. The World Wide Web Consortium had a programme of work on Provenance culminating in 2013[19] which is widely ignored – one issue is that it tries to revert the unstructured web to the mentality of databases and predefined schemas. Rather, for provenance, we must draw the lessons from web search and adopt statistical means to define probabilistic provenance graphs. The challenge will be that a significant number of the original sources of misinformation are not openly available on the web, so a global system would require the collaboration of many platform providers and the federation of provenance information.

Q6. What role could these technologies play in building a trustworthy information environment?

7. The algorithmic editorial processes that are at the heart of much social media are currently statistical optimisations with a primary goal to drive profit. It would not be surprising to find that this is the antithesis of provenance, so challenges will include whether citizens have any faith that asking for "provenance ordered search results" or equivalent has not been tampered with based on commercial considerations.

Q7. Are there any current regulatory or policy barriers to the successful development or deployment of detection technologies?

8. Currently in the UK, there is no primary legislation prohibiting the publication or circulation of misinformation. Nor is there a general obligation for platforms to monitor or remove misinformation. Equally, however, the law does not stop platforms from taking measures to address misinformation, although such measures must be fully in line with human rights standards, especially with respect to freedom of speech. The adoption of detection technologies by online platforms are facing two major regulatory barriers: The lack of economic incentives[20] and the legal uncertainty of what counts as misinformation[21]. The Government's Online Harms White Paper proposes to address these issues by introducing a statutory duty of care,[22] but the approach is subject to criticisms about the unclear definition of "harm".[23]

[18] Liang Wu et al. "Misinformation in social media: definition, manipulation, and detection." *ACM SIGKDD Explorations Newsletter* 21.2 (2019): 80-90.

[19] https://www.w3.org/TR/prov-overview/

[20] https://www.theguardian.com/technology/2016/nov/15/facebook-fake-news-us-election-trump-clinton

[21] https://www.pewresearch.org/internet/2017/10/19/the-future-of-truth-and-misinformation-online/

[22] https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper

[23] https://uhra.herts.ac.uk/handle/2299/21431