Pre-print of the article accepted to Communications of the ACM – Regional Special Section on East Asia and Oceania
DOI: https://doi.org/10.1145/3588431

Human-AI Cooperation to Tackle Misinformation and Polarisation

DAMIANO SPINA, RMIT University, Australia
MARK SANDERSON, RMIT University, Australia
DANIEL ANGUS, Queensland University of Technology, Australia
GIANLUCA DEMARTINI, The University of Queensland, Australia
DANA MCKAY, RMIT University, Australia
LAUREN L. SALING, RMIT University, Australia
RYEN W. WHITE, Microsoft Research, USA

A dominant narrative of the past decade is that algorithms contribute to a misinformed and segregated society. Perhaps paradoxically, algorithms are often sought as solutions to such problems. We describe a significant emerging trend away from this techno-solutionist approach that seeks to create and understand a new paradigm: a productive interplay between algorithms and people. Two relevant test cases are being explored in our region: the first addresses a new framework to tackle misinformation by assisting fact-checkers with computational methods; the second seeks new models to understand how search engines deliver personalised search results when little or no algorithmic personalisation exists.

CCS Concepts: • **Information systems** → **Information retrieval**; **Information retrieval query processing**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → Artificial intelligence; • **Applied computing** → Law, social and behavioral sciences.

1 MISINFORMATION IN EAST ASIA AND OCEANIA

In late 2020 and early 2021, the Australian Communication and Media Authority conducted a study to analyse the state of misinformation in Australia. The findings, reported to the Australian Government in June 2021, showed that four out of five Australian adults had seen misinformation about COVID-19. They also found that online misinformation, such as the propagation of anti-vaccine narratives within the Australian community, had a direct negative impact on the trust that people place in democratic institutions and public health agencies. These narratives often originate overseas but quickly spread through local communities. The fact-checking organisations that have traditionally verified statements made by public figures or politicians in public and mainstream media now also need to monitor and debunk dramatically faster-spreading claims on social media platforms. Narratives containing misinformation are having a direct and negative impact on the way people consume information: they may influence the content we engage with, and the search terms we enter [10]. Given that an informed citizenry is a cornerstone of democracy, public decision making is at risk.

The significance of the problem was also recognised in the International Cyber and Critical Technology Engagement Strategy released by the Australian Government, which identifies digital misinformation as a clear risk to the security and safety of Australia, the Indo-Pacific region, and beyond. Countries across East Asia and Oceania introduced legislation that specifically targets so-called 'fake news' and they created voluntary codes of practice developed in partnership with the technology industry.

Authors' addresses: Damiano Spina, RMIT University, Melbourne, Australia, damiano.spina@rmit.edu.au; Mark Sanderson, RMIT University, Melbourne, Australia, mark.sanderson@rmit.edu.au; Daniel Angus, Queensland University of Technology, Brisbane, Australia, daniel.angus@qut.edu.au; Gianluca Demartini, The University of Queensland, Brisbane, Australia, demartini@acm.org; Dana McKay, RMIT University, Melbourne, Australia, dana.mckay@rmit.edu.au; Lauren L. Saling, RMIT University, Melbourne, Australia; Ryen W. White, Microsoft Research, Redmond, WA, USA, ryenw@microsoft.com.

Pre-print of the article accepted to Communications of the ACM – Regional Special Section on East Asia and Oceania
DOI: https://doi.org/10.1145/3588431

Despite such efforts, as of December 2022, out of the 122 currently verified signatories of the Poynter's International Fact-Checking Network (IFCN), only eight are in the East Asia and Oceania region: Australian Associated Press (AAP) and RMIT FactLab in Australia; Cek Fakta Liputan 6, MAFINDO, Tempo.co, and Tirto ID in Indonesia; Rappler and Verafiles Incorporated in The Philippines.[1]

Some platforms have turned to fact-checkers to help identify problematic content. However, the deluge of misinformation means that the checkers are unable to handle the large number of claims that need to be assessed. Algorithmic assistance may therefore be beneficial to help identify instances of misinformation.

2 COMPUTATIONAL METHODS TO TACKLE MISINFORMATION

In the last few years, a trend has emerged of computing professionals tackling the problem of misinformation by means of hybrid (human and artificial) intelligence methods deployed to remove misinformation from online platforms [6]. The problem is more complex than identifying and removing misinformation; such content needs to be comprehensively *managed* throughout the stages of its life cycle, from creation, to propagation, and consumption. The computer science community has already developed technologies that can help at each stage (Figure 1).

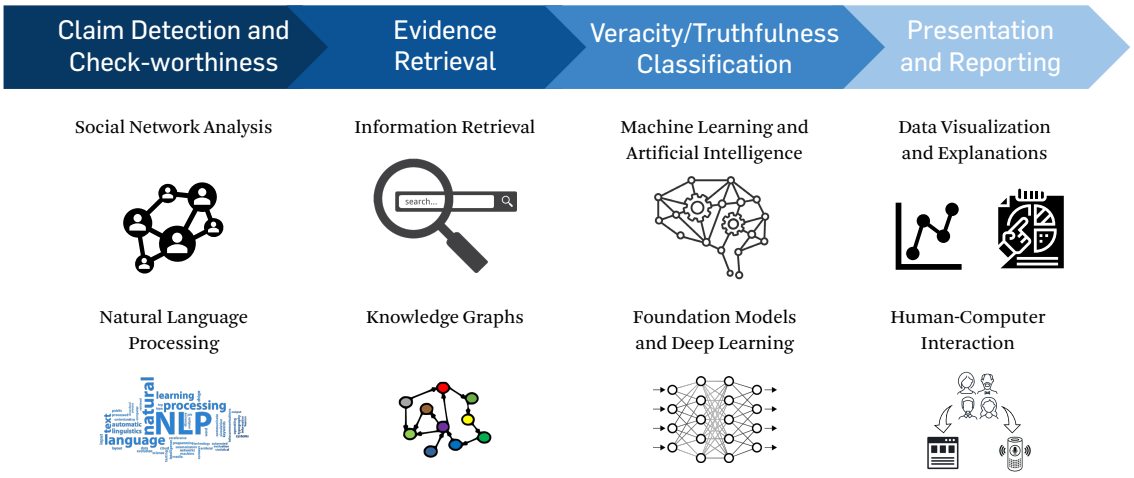

Fig. 1. Computational methods to assist in fact-checking.

A range of methods — including social network analysis, natural language processing, information retrieval (IR), knowledge graphs, machine learning and artificial intelligence, foundation models (e.g., neural transformers such as BERT and GPT) and deep learning, data visualization, explanations of machine-learned model output, and advances in human-computer interaction (e.g., new user experiences and interaction capabilities) — can assist, but not replace, humans during misinformation management processes. In all these stages, a close collaboration between experts,

[1] https://ifcncodeofprinciples.poynter.org/signatories [Accessed: 15 Dec 2022]

DOI: https://doi.org/10.1145/3588431

systems, and non-experts such as crowd workers is crucial to be able to scale up while maintaining the quality, as well as agency and accountability, of the process [6].

Human-in-the-loop fact-checking in the East Asia and Oceania region is in its infancy. While non-governmental organisations such as FirstDraft News (now the Information Futures Lab) are active in the region, more research is needed to better understand how hybrid intelligence methods can be effectively embedded into misinformation management processes without taking agency away from experts [9]. In addition to debunking misinformation, computational methods can assist in *prebunking* processes by developing effective ways to educate people about misinformation, thus enhancing digital literacy. This will provide them with the skills to identify and question unverified information online. The low number of fact-checking organisations in East Asia and Oceania makes the support of computational methods more pressing in this region. Ensuring that this algorithmic assistance is useful, though, will also require region-specific attention. What is 'fact' internationally is often counterfactual in East Asia and Oceania, so merely importing international information will not be effective. Examples of 'facts' that do not hold true in this region include seasonal issues: for much of the region summer is December–February. While having Christmas in summer is slightly countercultural, gatherings for these holidays did result in a major southern summer COVID-19 spike in 2021. Conversely, knowing that the influenza (and COVID-19) seasons occur in June and July in this part of the world is a key part of good public health advice. The region also predominantly sits close to the equator, so public health advice about sun exposure must be tailored. Most countries in the region drive on the left-hand side of the road, affecting road safety advice. There are also cultural differences within the region: many countries are predominantly Muslim, meaning Christmas is not celebrated at all (but Eid is), and that the weekend can fall on different days. Understanding the importance of COVID-19 vaccines being Halal was key to public health messaging in Melbourne, Australia. Democratic conventions are also unique: Australia has one of the highest rates of democratic participation in the world, at over 90%, a direct result of compulsory voting. Nearby New Zealand, though, also had strong democratic participation of over 80% without compulsory voting in the most recent election. Given these regional and local differences from global 'norms', local fact checking is key to ensuring an informed populace. Further, prebunking strategies have to sit alongside existing educational and cultural norms. This presents the two-pronged problem of scarce local expertise and the need for localised resources, to build and evaluate algorithmic tools and human-in-the-loop solutions for the region.

3 THE FILTER BUBBLE MYTH

While there is evidence that polarisation in society dramatically escalated with the introduction of broadband Internet, the cause is not well understood. Filter bubbles, formed by algorithms delivering personalised content that reinforces a particular worldview have become an incredibly popular explanation. The conceptualisation of the filter bubble was coined in a book by the same name by Eli Pariser [7]. However, the filter bubble concept may distract from the deeper epistemic causes of polarisation. Pariser's book, cited thousands of times, makes the case, but provides limited evidence of bubbles being formed by search engines. Empirical studies indicate a lack of such bubbles, going further to suggest that search platforms actually increase exposure to contrary viewpoints. Cross-disciplinary teams of computer scientists, media specialists, information scientists, industry researchers, and psychologists are working together on the issue of search personalisation through novel experimentation, which has better revealed the role that search engines play in polarisation.

The Australian Search Experience [3], a project carried out by the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (ADM+S), is a data donation study where over 1,000 people across Australia were recruited to examine whether search engines returned different kinds of results across the cohort. Participants

Pre-print of the article accepted to Communications of the ACM – Regional Special Section on East Asia and Oceania
DOI: https://doi.org/10.1145/3588431

Table 1. Sample of query variants crowdworkers generated, drawn from the UQV test collection [2].

Description of Information Need	Example Query Variants
A group of local farmers has been protesting outside your energy utility's offices, complaining about a plan to build a wind farm on hills near their properties. Their placards say that the disadvantages (cons) outweighed the advantages (pros). Until now you had always assumed that wind power was a good thing; now you are not so sure, and decide to find out more.	• advantages and disadvantages of wind power • cons of wind power • is wind power good • negative effects of using wind power • pros and cons of wind power • information about wind power • engineering principles of wind turbines

installed a web browser plugin that issued periodic queries to well-known search engines using the participants' accounts. Search results were scraped by the plugin and returned to researchers for examination. The queries were drawn from a pre-defined list of common searches on a range of topics that spanned political, controversial, and everyday categories. While the research is still ongoing, initial findings indicate that, although search results were found to be contextualised to particular specific geographic locales, algorithmic personalisation in search engines may be less extensive than was suggested by previous filter bubble research. This leads to the question: if search is largely homogeneous, where is information polarisation coming from?

One possible answer lies in work by the IR community in our region examining the impact of query variation in search. While users of search engines have been studied for decades, recent experiments where a large number of people are asked what query they would use when seeking to satisfy a common information need have found an astonishing range of distinct queries [2]. To illustrate, Table 1 lists a sample drawn from over 50 query variants found when 100 crowdworkers were asked how they would search for information about wind power. Such extensive variations were recorded across a diverse set of 100 topics. The results of the experiment are packaged in a test collection that captures this user query variability (UQV).

Query variations were found to have a significant impact on search engine performance. Wide variations in the queries submitted to commercial search engines were identified [1] and detailed statistical analysis found that variations in queries had a substantially larger effect on search results than any change in the workings of a search algorithm [5]. The Australian Associated Press recently debunked a social media post with a false claim about the lifespan of wind farm generators.[2] One can see in the sample queries shown in Table 1 that different queries appear to reflect different attitudes to the topic. It is natural to wonder whether misinformation influences the way people choose the keywords that they type into a search engine.

4 OPPORTUNITIES TO ADDRESS POLARISATION IN SEARCH ENGINES

This collection of findings suggest that polarisation in search is being driven not by algorithms, but by searchers [1]. The research trend highlights a critical oversight in search engine algorithm design: understanding how search algorithms react to and potentially alleviate this user variation. The research challenges of such work include:

[2] https://www.aap.com.au/factcheck/wind-turbine-lifespan-claim-generates-misinformation [Accessed: 10 Feb 2023]

- Understanding the reasons for the variation people show when searching (e.g., demographics, search habits, domain knowledge, cognitive biases, how people are prompted to search),
- Exploring if and how people are influenced by others to search in particular ways, and
- Determining how search algorithms can be adapted to better handle the variation.

Initial results suggest that the way people construct queries is informed by established searching habits, although other factors, such as existing knowledge, biases, prompts, etc., most likely also contribute. Questions of how people are influenced to search need to be examined. Here, misinformation seems to play a crucial role and collaborations with fact-checking organisations in the region [4, 8] are helping to better understand how people formulate their queries when they encounter misinformation and interventions (e.g., verified content produced by fact-checkers). When examining search algorithms, the roles and responsibilities of search engines will be questioned: most would agree that search engines should return the most reliable content, but should they intervene in trying to change user views arising from polarising queries? Answers to such questions need to be approached in a nuanced way because confronting people with views too distant to their own could prove to be alienating. Other key research issues include investigating whether queries resulting from a disinformation campaign can be reliably detected, whether search engines could and should detect the sole pursuit of confirmatory information and evidence of confirmation biases, and whether search results can be tailored to support reflection and understanding of those with beliefs different to one's own.

5 THE ROAD AHEAD

Detailing these two case studies shows that a richer engagement between humans and machines ensures more effective outcomes in the management of information and a better understanding of how online information is sought. The work described here represents an important emerging trend in our region, one that has impacts far beyond this geographic area. It also presents a number of grand challenges in deploying these assistive technologies at a massive scale and realising human-AI cooperation in practice.[3] Computing professionals will need to continue collaborating with other disciplines to make and integrate advances in critical areas such as fairness, accountability, transparency, explainability, and the safety of human-AI cooperation. Misinformation and its exposure will only grow in the coming years, as will the adversarial uses of computational methods to generate and spread disinformation narratives. Polarisation will persist as long as we fail to understand the causes of query variation in search engine engagement and fail to develop more robust search algorithms capable of handling that variation. As a community, we must meet these challenges head on. Understanding and supporting the interplay of humans and algorithmic systems will ultimately lead to better outcomes for all.

ACKNOWLEDGMENTS

The authors would like to thank Axel Bruns, Luke Gallagher, Timothy Graham, James Meese, Stefano Mizzaro, Quoc Viet Hung Nguyen, Abdul Karim Obeid, and Falk Scholer for their contributions and feedback towards this work. This research is partially supported by the Australian Research Council (DE200100064, CE200100005, IC200100022). The authors acknowledge the Traditional Custodians of Country throughout Australia and their connections to land, sea, and community. We pay our respect to their Ancestors and Elders past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today.

[3] As also highlighted in the ACM Technology Policy Council's Statement on Principles for Responsible Algorithmic Systems: https://www.acm.org/binaries/content/assets/public-policy/final-joint-ai-statement-update.pdf [Accessed: 10 Feb 2023]

Pre-print of the article accepted to Communications of the ACM – Regional Special Section on East Asia and Oceania
DOI: https://doi.org/10.1145/3588431

REFERENCES

[1] Marwah Alaofi, Luke Gallagher, Dana Mckay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. 2022. Where Do Queries Come From?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*. Association for Computing Machinery, 2850–2862. https://doi.org/10.1145/3477495.3531711

[2] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2016. UQV100: A Test Collection with Query Variability. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*. Association for Computing Machinery, 725–728. https://doi.org/10.1145/2911451.2914671

[3] Axel Bruns. 2022. *Australian Search Experience Project: Background Paper.* Technical Report. ARC Centre of Excellence for Automated Decision-Making and Society. https://doi.org/10.25916/k7py-t320

[4] Assunta Cerone, Elham Naghizade, Falk Scholer, Devi Mallal, Russell Skelton, and Damiano Spina. 2020. Watch 'n' Check: Towards a Social Media Monitoring Tool to Assist Fact-Checking Experts. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 607–613. https://doi.org/10.1109/DSAA49011.2020.00085

[5] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. 2021. Topic Difficulty: Collection and Query Formulation Effects. *ACM Trans. Inf. Syst.* 40, 1, Article 19 (Sept. 2021), 36 pages. https://doi.org/10.1145/3470563

[6] Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.* 43, 3 (2020), 65–74. http://sites.computer.org/debull/A20sept/p65.pdf

[7] Eli Pariser. 2011. *The Filter Bubble: What the Internet is Hiding from You.* Penguin UK.

[8] Lauren L. Saling, Devi Mallal, Falk Scholer, Russell Skelton, and Damiano Spina. 2021. No One Is Immune to Misinformation: An Investigation of Misinformation Sharing by Subscribers to a Fact-checking Newsletter. *PLOS ONE* 16, 8 (Aug. 2021), 1–13. https://doi.org/10.1371/journal.pone.0255702

[9] T.J. Thomson, Daniel Angus, Paula Dootson, Edward Hurcombe, and Adam Smith. 2022. Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities. *Journalism Practice* 16, 5 (2022), 938–962. https://doi.org/10.1080/17512786.2020.1832139

[10] Elad Yom-Tov, Susan Dumais, and Qi Guo. 2014. Promoting Civil Discourse Through Search Engine Diversity. *Social Science Computer Review* 32, 2 (2014), 145–154. https://doi.org/10.1177/0894439313506838