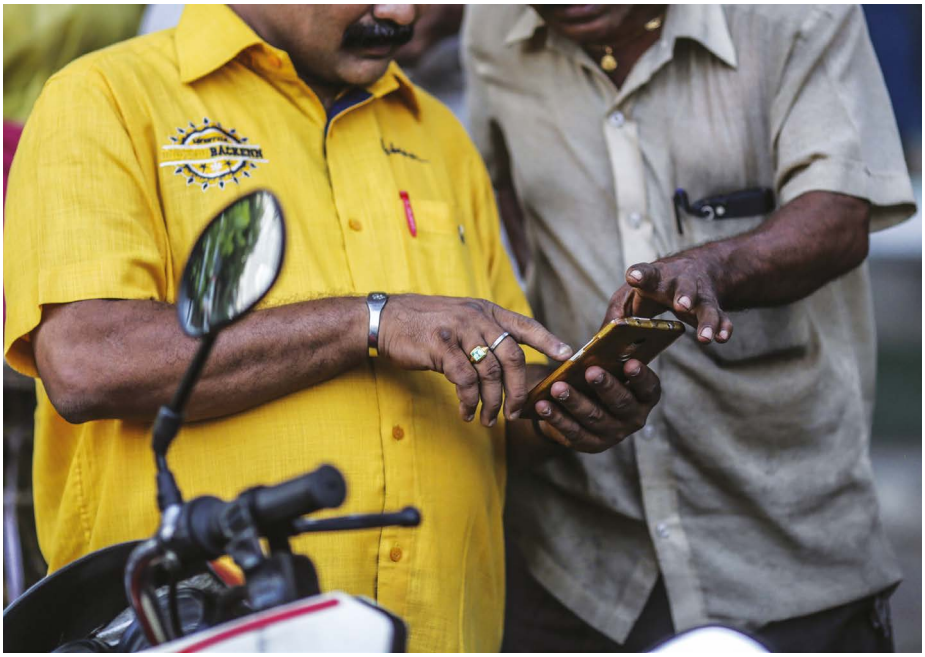
either voluntarily or in response to regulatory pressures, such as the EU's Digital Services Act, which regulates online platforms to prevent the spread of disinformation. The time to act is now.

The authors

Ullrich Ecker is a professor at the School of Psychological Science and a fellow at the Public Policy Institute, University of Western Australia, Perth, Australia. **Jon Roozenbeek** is assistant professor in psychology and security at the Department of War Studies, King's College London, UK. **Sander van der Linden** is a professor of social psychology in society at the University of Cambridge, UK. **Li Qian Tay** is a postdoctoral fellow at the School of Medicine and Psychology, Australian National University, Canberra, Australia. **John Cook** is a senior research fellow at the Melbourne Centre for Behaviour Change at the University of Melbourne, Australia. **Naomi Oreskes** is the Henry Charles Lea professor of the History of Science at Harvard University, Cambridge, Massachusetts. **Stephan Lewandowsky** is a professor of cognitive science at the University of Bristol, UK, and a guest professor at the University of Potsdam, Germany.
e-mail: ullrich.ecker@uwa.edu.au

1. Lewandowsky, S. *et al. Technology and Democracy: Understanding the Influence of Online Technologies on Political Behaviour and Decision-Making* (Publications Office of the European Union, 2020).
2. Lewandowsky, S. *et al. Curr. Opin. Psychol.* **54**, 101711 (2023).
3. van der Linden, S. *et al. Using Psychological Science to Understand and Fight Health Misinformation: An APA Consensus Statement* (American Psychological Association, 2023).
4. Kozyreva, A. *et al. Nature Hum. Behav.* https://doi.org/10.1038/s41562-024-01881-0 (2024).
5. Adams, Z., Osman, M., Bechlivanidis, C. & Meder, B. *Perspect. Psychol. Sci.* **18**, 1436–1463 (2023).
6. Bretter, C. & Schulz, F. *Proc. Natl Acad. Sci. USA* **120**, e2217716120 (2023).
7. Freiling, I., Krause, N. M. & Scheufele, D. A. *AMA J. Ethics* **25**, 228–237 (2023).
8. Krause, N. M., Freiling, I. & Scheufele, D. A. *Ann. Am. Acad. Polit. Soc. Sci.* **700**, 112–123 (2022).
9. Uscinski, J., Littrell, S. & Klofstad, C. *Curr. Opin. Psychol.* **57**, 101789 (2024).
10. van Doorn, M. *Inquiry* https://doi.org/10.1080/0020174X.2023.2289137 (2023).
11. Oreskes, N. *Why Trust Science?* (Princeton Univ. Press, 2019).
12. Vickers, P. *Identifying Future-Proof Science* (Oxford Univ. Press, 2022).
13. Lewandowsky, S. *et al.* Preprint at PsyArXiv https://doi.org/10.31234/osf.io/q7vbr (2024).
14. Ecker, U. K. H. *et al. Nature Rev. Psychol.* **1**, 13–29 (2022).
15. Henkel, I. *Destructive Storytelling: Disinformation and the Eurosceptic Myth That Shaped Brexit* (Palgrave Macmillan Cham, 2021).
16. Tay, L. Q., Lewandowsky, S., Hurlstone, M. J., Kurz, T. & Ecker, U. K. H. *Commun. Psychol.* **2**, 4 (2024).
17. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. *Sci. Adv.* **8**, eabo6254 (2022).
18. Bailard, C. S., Porter, E. & Gross, K. *Harv. Kennedy Sch. Misinformation Rev.* https://doi.org/10.37016/mr-2020-109 (2022).
19. Kozyreva, A. *et al. Proc. Natl Acad. Sci. USA* **120**, e2210666120 (2023).

The authors declare competing interests; see go.nature.com/3x5jqal for details.



DHIRAJ SINGH/BLOOMBERG/GETTY

Social media is an important channel for political communication and discussion in India.

How prevalent is AI misinformation? What our studies in India show so far

Kiran Garimella & Simon Chauchard

A sample of two million messages received by people through WhatsApp groups in rural India sheds light on the type and spread of political content generated by artificial intelligence.

The ability of generative artificial intelligence (genAI) to create seemingly authentic text, images, audio and video is causing rising concern worldwide. As many nations head to the polls this year, the potential of such content to fuel misinformation and manipulate political narratives is a troublesome addition to the list of challenges that election authorities face.

Over the past 12 months, several high-profile examples of deepfakes produced by genAI have emerged. These include images of US Republican presidential candidate Donald Trump in the company of a group of African American supporters, an audio recording of a Slovakian politician claiming to have rigged an upcoming election and a depiction of an explosion outside the US Department of Defense's Pentagon building.

But despite the immense attention that such deepfakes garner, systematic studies on the impact of genAI on misinformation are limited. This is mainly because access to comprehensive data is restricted owing to privacy concerns and limitations imposed by social-media platforms. Many crucial questions remain unanswered.

For instance, it's unclear how prevalent AI-produced 'fake news' is — few concrete data are available on how often AI technologies are used to create deceptive content. Nor do researchers yet understand the extent to which AI-generated false information is being shared on social media, or the impact it has on public opinion. These uncertainties underscore the challenges of tackling AI-driven misinformation and highlight the need for more robust research and monitoring to grasp the

full implications of these technologies.

For more than five years now, we have been working in India to understand how political parties are using the messaging platform WhatsApp to spread misinformation[1,2]. Over the past year, we have collected data from app users in rural India through a privacy-preserving, opt-in method of data donation[3].

WhatsApp is an important channel for political communication and discussion in India[4], as it is in countries such as Brazil and Kenya. Our research includes many users who are new to the Internet, and who constitute a significant proportion of the online population in rural India. Such users are likely to be less familiar with the capabilities of genAI and potentially more susceptible to its influence than seasoned users. Here we share some findings from our ongoing research.

How common is genAI content?

We chose to focus our study on Uttar Pradesh, India's largest state, with a population exceeding 220 million. We approached nearly 500 users, chosen to provide a representative sample across demographic variables such as age, religion and caste. Although this might seem like a small set of users, this is one the largest samples obtained by asking WhatsApp users to donate the data on their phones for research purposes. We also observed messaging patterns over a long enough period to give a reliable snapshot of how political conversations play out. We did this by monitoring all the non-personal WhatsApp groups, downloading any new messages automatically. We took the utmost care to remove any personal identifiers before the data were stored and analysed, and to ensure that the consent process was thorough and informative[3].

The first set of messages was collected between August and October last year, just ahead of major provincial elections. This provided a data set of around two million messages from mostly private WhatsApp groups, giving us, for the first time, a sense of what ordinary WhatsApp users were seeing and discussing. The topics varied broadly, encompassing religion, education, village life and other national and local matters.

We were especially interested in content that spread virally on WhatsApp — marked by the app as 'forwarded many times'. Such a tag is placed on content that is forwarded through a chain that involves at least five hops from the original sender. Five hops could mean that the message has already been distributed to a very large number of users, although WhatsApp does not disclose the exact number.

Automated methods do not yet exist for identifying AI-generated content at scale, so we examined all 1,858 viral messages in our sample manually. To annotate genAI images and videos, we initially concentrated on content that exhibited clear signs of being machine-generated — such as certain textural anomalies or unnatural blending. We also involved specialists who could evaluate the authenticity of each piece of content by focusing on characteristics known to be typical of AI creations. We acknowledge the inevitable limitations in such assessments until better techniques emerge for distinguishing material created by genAI.

Of the 1,858 viral messages, fewer than two dozen contained instances of genAI-created content — a footprint of just 1%. Even though this figure might be a slight undercount and does not include text messages or audio, the finding suggests that the prevalence of genAI-created content in viral messages in India is relatively low. That seems to have remained the case during the 'multi-phase' general election in the country that has just concluded. We are continuing to monitor viral content and have not detected any significant spike in genAI content used for political campaigning. The impact of genAI on election misinformation might not yet be as extensive as feared.

Nonetheless, these are early days for the technology, and our first findings showcase genAI's power to manufacture compelling, culturally resonant visuals and narratives, extending beyond what conventional content creators might achieve.

HIMANSHU SHARMA/AFP/GETTY

A woman casts her ballot in the 2024 Indian general election.

What does genAI content show?

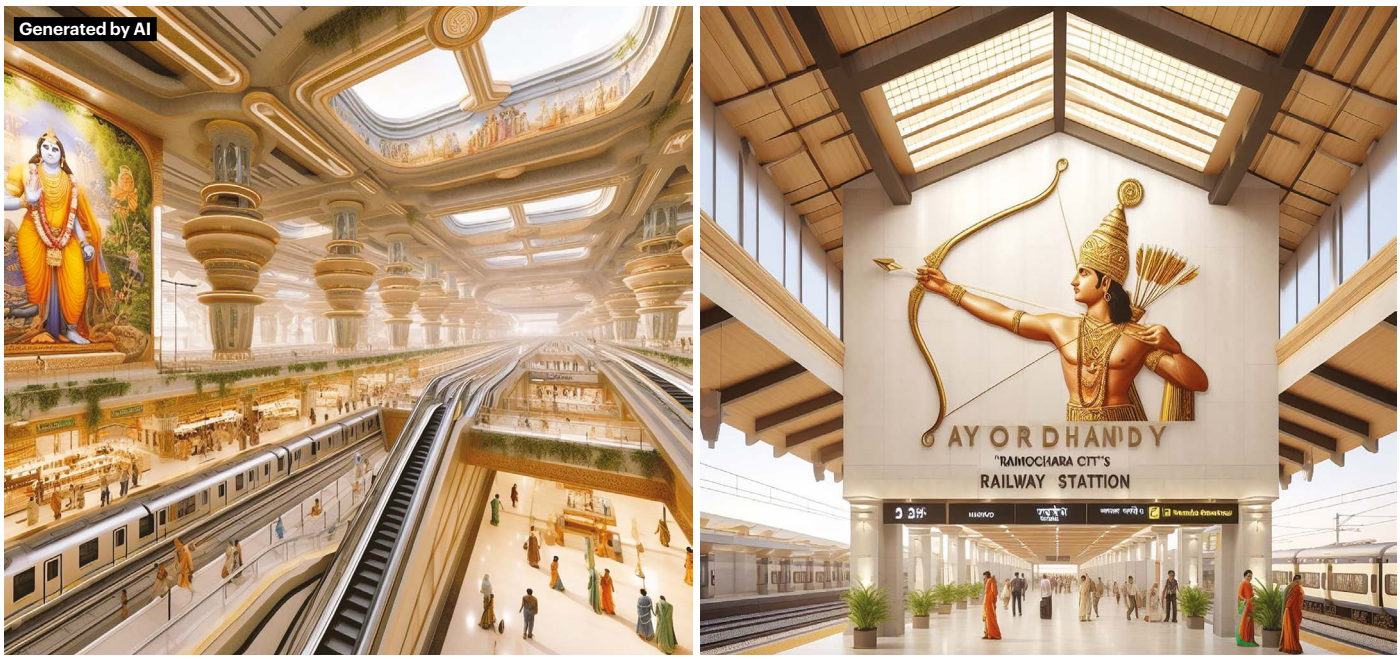
One category of misleading content we identified relates to infrastructure projects. Visually plausible genAI images of a futuristic train station, allegedly depicting a new facility at Ayodhya — a city many Hindus believe to be the birthplace of the god Lord Ram — managed to spread widely. The pictures featured a spotless station showcasing depictions of Lord Ram on its walls. Ayodhya has been the site of religious tensions, particularly since Hindu nationalists demolished a mosque in December 1992 so that they could build a temple over its ruins.

Rapid infrastructure development, including the modernization of train stations, is a key policy goal for the Bharatiya Janata Party (BJP), which seeks to portray India as a swiftly modernizing economy to both domestic and international audiences. However, our data do not provide direct evidence linking the BJP, or any other political party, with the creation or dissemination of AI-generated images.

Another theme evident in the genAI content seemed to be projections of a brand of Hindu supremacy. For example, we uncovered AI-generated videos showing muscular Hindu saints mouthing offensive statements against Muslims, often referencing historical grievances. AI was also used to create images glorifying Hindu deities and men with exaggerated physiques sporting Hindu symbols — continuing a long-standing and well-documented propaganda tactic of promoting Hindu dominance in India. We also found media depicting fabricated scenes from the ongoing war in Gaza, subtly equating the 7 October 2023 attacks by Hamas against Israel with alleged violence perpetrated by Muslims against Hindus in India. Using current events to target minority groups is a well-documented trend in the Indian social and political context[5].

What should be done?

Although our study suggests that genAI technology is unlikely to single-handedly shape elections at the moment, the cases we document show its potential for personalizing and



K. GARIMELLA. PREPRINT AT GO.NATURE.COM/4DYJ3AD

AI-generated images spread through social media include a mock-up of a modern railway station at Ayodhya.

turbocharging disinformation campaigns in the future. Even if such content resembles animation, it can still be highly effective, because its hyper-idealized imagery resonates on an emotional level, especially with viewers who already have sympathetic beliefs. This emotional engagement, combined with the visual credibility that modern AI provides[6] – blurring the line between animation and reality – could allow such content to be persuasive, although this requires further study.

It is important to note that misinformation can easily be created through low-tech means, such as by misattributing an old, out-of-context image to a current event, or through the use of off-the-shelf photo-editing software. Thus, genAI might not change the nature of misinformation much.

Nonetheless, as the technology matures and becomes more widely accessible, the need for vigilance and countermeasures will rise. GenAI can be used to generate a large volume of images rapidly at low cost and offers the ability to customize content dynamically to resonate with specific cultural or social groups. Its use can also obscure the origin of the content, making it difficult to trace its creators. This combination of scalability, cost-efficiency, customization and anonymity could enhance genAI's impact on public opinion.

As this disruptive technology continues to evolve, sustained monitoring of its prevalence and impact across different contexts will be crucial. In the Indian context and elsewhere, researchers need to look further at three key aspects related to the impact of AI on society and politics.

Can genAI level the playing field? GenAI has the potential to disrupt existing power dynamics in content production. By democratizing content creation, it challenges the dominance of well-resourced political parties.

Can social-media firms do more? WhatsApp's decentralized nature presents both opportunities and challenges for content distribution. Although it enables the widespread dissemination of information, it also complicates efforts to regulate and moderate content. GenAI's ability to create personalized and targeted content at scale might amplify existing biases and echo chambers. Without effective moderation mechanisms, platforms risk becoming breeding grounds for misinformation and propaganda.

How does genAI influence beliefs? Although users today might be able to detect genAI-generated content, it could become increasingly difficult to distinguish what's real from what's fake as the technology improves. This is particularly concerning in the context of messaging apps such as WhatsApp, because users might be more likely to believe the content they receive from trusted groups or individuals, even if it is misleading or false[7].

Policymakers should thus develop comprehensive global and domestic strategies to combat the ill effects of AI-generated content. Public messaging to stress the importance of source verification is key. Building a framework to watermark content as being AI-generated is an urgent necessity. Initiating public awareness campaigns to educate communities, especially inexperienced tech users, about the nuances of AI-generated content is also important.

Collaboratively evolving new regulatory frameworks, involving governments, the technology industry and academic bodies, can ensure that standards keep pace with AI advancements. The final piece of the puzzle is to sustain support for research to build sophisticated technologies that can detect AI at scale – a crucial requirement in the age of synthetic media.

Utmost care must be taken in devising mitigation mechanisms, because compromising digital privacy and security to target AI-powered misinformation can be counterproductive. We firmly caution against proposals that might undermine end-to-end encryption on platforms such as WhatsApp.

The authors

Kiran Garimella is an assistant professor in the School of Communication and Information at Rutgers University in New Brunswick, New Jersey, USA. **Simon Chauchard** is an associate professor of political science and a distinguished researcher at University Carlos III in Madrid, Spain.
e-mails: kg766@comminfo.rutgers.edu; simon.chauchard@uc3m.es

1. Chauchard, S. & Garimella, K. *J. Quant. Descr. Digit. Media* https://doi.org/10.51685/jqd.2022.006 (2022).
2. Garimella, K. & Eckles, D. *Harv. Kennedy Sch. Misinformation Rev*. https://doi.org/10.37016/mr-2020-030 (2020).
3. Garimella, K. & Chauchard, S. Preprint at OSFPreprints https://doi.org/10.31219/osf.io/k6qv5 (2024).
4. Goel, V. 'In India, Facebook's WhatsApp Plays Central Role in Elections' *The New York Times* (14 May 2018).
5. Engineer, A. A. *Oriente Moderno* **84**, 71–82 (2004).
6. Davenport, T. H. & Mittal, N. 'How generative AI is changing creative work' *Harvard Business Review* (14 November 2022).
7. Gursky, J., Riedl, M. J., Joseff, K. & Woolley, S. *Soc. Media Soc.* **8**, 20563051221094773 (2022).

The authors declare no competing interests.