The Academic
International Journal of Multidisciplinary Research

An Online Peer Reviewed / Refereed Journal
Volume 4 | Special Issue 2 | April 2026
ISSN: 2583-973X (Online)
Website: www.theacademic.in

Artificial Intelligence and Ideological Propaganda Mechanisms of Cognitive Influence and Mitigation Strategies

Amit S. Chingali, Avadhut Manage

D.M.S. Mandal's Bhaurao Kakatkar College, Belgaum, amitchingali.ac@gmail.com

DOI : https://doi.org/10.5281/zenodo.20057115

ARTICLE DETAILS

Research Paper

Accepted: 05-04-2026

Published: 18-04-2026

Keywords:

artificial intelligence, media cells, algorithmic propaganda, deepfakes, cognitive manipulation, disinformation, recommendation systems, media literacy, social bots

ABSTRACT

Artificial intelligence has fundamentally reshaped ideological persuasion by combining scale, personalisation, and apparent authenticity at a level no prior technology achieved. This review draws on cognitive psychology, political communication, computational social science, and information security to analyse how AI-powered systems large language models, synthetic media tools, recommendation engines, and behavioural targeting platforms are actively reshaping how beliefs are formed and manipulated. Equal attention is given to organised media cells, the human networks that supply strategic direction to AI-generated campaigns. The article traces the cognitive vulnerabilities these systems exploit, reviews real-world deployment cases, and evaluates current countermeasures against their actual track record. The core finding is that AI has not invented propaganda but has removed nearly every logistical barrier that previously limited it, making targeted ideological influence cheaper, faster, and harder to detect than at any prior moment in history. Addressing this requires coordinated action across technology platforms, regulatory institutions, and civil society.

1. Introduction

Propaganda is ancient. The printing press industrialised it; radio and television gave it mass reach; the internet made it global and cheap. Artificial intelligence has done something qualitatively different: it has

personalised propaganda at a scale and with a precision simply not achievable before. That shift, rather than novelty alone, makes the present situation worth examining carefully. The infrastructure responsible emerged largely through tools designed for other ends. Recommendation algorithms were built to retain engaged users. Behavioural analytics were developed for advertisers. Large language models were trained for writing assistance. Synthetic media tools were created for film production. None was designed to manipulate elections or radicalise populations, yet each has proven exceptionally well suited to those ends, individually and in combination.

The literature on this topic is sprawling and fragmented across disciplines. What is frequently missing is an account that holds all threads simultaneously and pays equal attention to the human organisational layer through which AI capabilities are deployed in practice. This review attempts that synthesis.[1,2]

2. Historical Context: How We Got Here

States during the World Wars mobilised film, radio, and print as systematic instruments of public persuasion.[3] What constrained those operations was primarily logistical: producing credible content at scale required significant human labour, institutional infrastructure, and capital. The internet began dissolving these constraints from the mid-1990s, democratising publishing while simultaneously removing structural barriers that had moderated the speed and scale at which false or manipulative content could propagate.

The period from roughly 2010 to 2016 introduced computational propaganda: the use of automation, data analytics, and platform algorithms to amplify political messaging at industrial scale.[1] The Internet Research Agency operation during the 2016 US presidential election is the canonical case study. Its distinguishing feature was not the quality of individual content, much of which was crude, but the scale of distribution and the precision with which different messages were targeted at different demographic and psychographic segments.[4]

Generative AI, accelerating sharply from around 2020 onward, added customisability and apparent authenticity to that scale. A language model can produce a persuasive opinion essay, a fake local news article, a simulated eyewitness account, and a personalised social media post—all on the same topic, all in different styles—at negligible marginal cost per item. That combination has changed the economics of influence operations in ways researchers are still working through.[5]

3. Technical Mechanisms of AI-Enabled Propaganda

3.1 Large Language Models as Content Factories

LLMs can generate fluent, contextually coherent text across an enormous range of styles and subject matters. Experimental work has shown that LLM-generated political persuasion content was rated as credible by subjects and produced measurable shifts in stated policy preferences.[5] The practical significance lies in scalability: a single operator can generate thousands of thematically consistent posts in different stylistic registers, making coordinated inauthentic behaviour far harder to detect through linguistic pattern analysis. There is also a subtler mechanism: LLMs deployed as conversational assistants tend to present as balanced while quietly favouring certain framings embedded in their training data, biases that may be invisible to users who experience the system as neutral.[6]

3.2 Deepfakes and the Liar's Dividend

Chesney and Citron identified what they called the liar's dividend: even before any specific deepfake causes harm, public knowledge that convincing synthetic media exists allows genuinely authentic recordings to be dismissed as potentially fabricated.[7] This epistemic damage degrades the evidential value of video and audio as a category. Voice cloning is particularly insidious because audio encountered briefly in a social media story leaves a narrow window for critical evaluation. The 2023 and 2024 election cycles saw fabricated audio attributed to political figures spread during critical news windows precisely when fact-checking organisations were most stretched.[7,8]

3.3 Recommendation Algorithms and the Gradual Drift

Recommendation systems are probably the least visible and most consequential mechanism in this ecosystem. Unlike deepfakes, they are not discrete artefacts that can be identified and debunked; they operate through cumulative exposure effects embedded in the daily experience of billions of users. These systems are optimised for engagement metrics, and emotionally arousing, partisan, and novel content consistently outperforms balanced reporting on those metrics.[9] Research on YouTube viewing data found evidence of consistent movement toward more politically partisan content,[10] and the filter bubble dynamic in which users are progressively insulated from content that challenges their existing positions[11] represents a structural advantage for identity-confirming, emotionally charged messaging over nuanced analysis.

3.4 Micro-targeting and Social Bots

Behavioural data analytics enables messages tailored to individual psychological profiles and inferred emotional states. Research established that digital behaviour traces could predict personality traits with accuracy exceeding that of close friends.[13] If personality can be inferred from behaviour, messages can be engineered to resonate with specific dispositions. Zuboff's analysis frames this infrastructure not as an abuse of advertising technology but as the expression of what it was designed to do: modify human behaviour at scale.[14]

Coordinated bot networks manufacture the appearance of popular support for positions that may command little genuine public endorsement, exploiting the well-understood heuristic by which people use apparent consensus to evaluate the validity of claims.[15] The arms race between detection and evasion has accelerated sharply with the availability of LLMs, which can generate bot content sophisticated enough to evade tests designed to catch automated output.

4. Cognitive Vulnerabilities That Propaganda Exploits

People are not neutral evaluators of evidence. When information touches beliefs tied to identity or group membership, cognitive effort is preferentially invested in defending the prior belief. AI propaganda systems exploit this by delivering identity-consistent content at high volume while the recommendation architecture simultaneously reduces exposure to disconfirming evidence, forming a feedback loop that narrows the effective information environment without any censorship occurring.[16]

Repeated exposure to a claim reliably increases its perceived truth one of the most robustly replicated findings in cognitive psychology.[17] AI systems can generate thousands of distinct formulations of a single false claim, varied in phrasing, context, and apparent source, and distribute them across many platforms. The recipient experiences repeated exposure to essentially the same claim while perceiving source diversity, so the illusory truth effect operates on each exposure while the apparent independence of sources prevents the scepticism that blatant repetition would otherwise trigger.[17,18]

Emotional states spread through social networks in ways that parallel contagion. Content that systematically induces fear, anger, or disgust spreads not just a belief but an affective state in which critical evaluation is compromised by the very feelings the content was designed to induce.[19] Affective polarisation—growing political hostility between groups independent of actual policy disagreement—is

systematically exacerbated by AI-optimised content that maximises emotional arousal, even when every factual claim is technically accurate.[20]

Tajfel and Turner's social identity theory established that people derive significant portions of their self-concept from group memberships, and that perceived threats to group status generate strong defensive responses.[21] Micro-targeting allows propagandists to identify individuals for whom a particular group identity is especially salient and to deliver content calibrated to reinforce that identity while heightening perception of out-group threat. The result is not merely a changed belief but a deeper affective alignment considerably more resistant to subsequent correction than ordinary factual error.

5. How Media Cells Spread Propaganda Through Social and Mainstream Media

The focus on AI tools can create the impression that contemporary propaganda is primarily automated. This is a misleading picture. Behind the majority of successful influence operations there are organised human networks—media cells—that supply strategic direction, cultural knowledge, and tactical coordination that no algorithm can provide alone.

5.1 Structure and Operation

A media cell is a small coordinated group tasked with producing, distributing, and managing the reception of ideological content across multiple platforms. Sophistication varies widely, from volunteer networks sharing pre-written talking points to professional communications agencies with defined roles for content creation, platform management, audience analytics, and rapid response.[1,4] The IRA's documented structure operating around the clock in shifts, assigning specific audience segments to team members, and maintaining quality-control processes for cultural plausibility provides the clearest public example of the sophisticated end of the spectrum.[4] The availability of AI content generation tools has dramatically reduced the labour cost for smaller cells, allowing a group of three or four motivated individuals to produce content volumes that would previously have required professional staffing.

5.2 Platform-Specific Tactics

Media cells do not treat platforms as interchangeable. On Twitter and X, the primary goal is typically creating the impression of trending consensus through coordinated posting and strategic hashtag use, triggering self-amplifying dynamics when organic users begin engaging. Facebook remains important for reaching older demographics and penetrating closed community spaces where content is processed with

less critical scrutiny than impersonal feeds; cells invest in building community trust over years before deploying groups for ideological purposes. YouTube's algorithm rewards watch time, so cells invest in documentary-style content that appears balanced while systematically excluding opposing viewpoints.[10] WhatsApp and Telegram's encryption allows content that would be removed from public platforms to circulate freely, later laundered into the broader information environment through screenshots shared to public social media.[22] Instagram and TikTok prioritise aesthetic appeal and emotional resonance; a fifteen-second video pairing an evocative image with a politically charged caption can reach millions without making a single verifiable factual claim.

5.3 Narrative Laundering and Organic Amplifiers

One of the most consequential strategies is the deliberate movement of narratives from fringe online spaces into mainstream media coverage. A media cell generates and amplifies a narrative on social media, creating the appearance of widespread public controversy. A journalist observing a trending topic may write a story about the controversy without endorsing the underlying claim giving it substantial additional exposure while maintaining journalistic neutrality. Debunking stories create the same effect: they repeat the claim even as they correct it, contributing to illusory truth dynamics.[17,23] The sheer volume of AI-equipped cells produces a firehose effect that overwhelms fact-checkers' capacity to address each claim.[24]

Perhaps the most valuable asset a media cell can develop is not an automated account but an authentic, trusted voice who will amplify its narratives without any coordination being visible. Micro-influencers with follower bases in the tens of thousands typically command far more audience trust than institutional sources, and content shared by a trusted personal voice is processed with the cognitive generosity extended to recommendations from friends.[25] Some amplifiers are approached directly; others are identified as ideologically sympathetic and fed validating content with no explicit coordination; others become entirely unwitting amplifiers. This last category means influence operations can achieve genuine organic spread through people with no idea they are participating in a coordinated campaign.

6. Documented Cases

The IRA operation during the 2016 US election remains the most thoroughly documented large-scale computational propaganda campaign in the public record. Operating across Facebook, Twitter, YouTube, and Instagram, IRA-linked accounts created and amplified divisive content on immigration, race relations, and gun control while organising real-world events and infiltrating genuine activist

communities.[4] The operation's success lay in the scale, precision of targeting, and duration rather than in the quality of any individual content item.

Chinese state-linked influence operations documented by the Stanford Internet Observatory have employed AI-generated profile images and LLM-assisted content creation in campaigns targeting Taiwan, Hong Kong, and Western democratic publics, with increasing sophistication in cover identities and cultural literacy over time.[26] A significant proportion of AI-generated disinformation is also commercially motivated: false news spreads faster and further than true news[27] because emotional novelty drives engagement, creating consistent financial incentives for publishers who generate high-engagement false content. Decentralised extremist communities have adopted AI content generation tools to accelerate recruitment and automate targeting of potentially receptive individuals, requiring no central coordination or state funding shared ideology and shared tools are sufficient to produce coordinated effects.[10,28]

7. Mitigation Strategies and What They Actually Achieve

Automated deepfake and synthetic text detection systems achieve high accuracy on controlled benchmark datasets. Their real-world performance against content specifically optimised to evade detection is considerably weaker, and the fundamental asymmetry between generation and detection where generation can be iterated far faster than detection can be validated represents a structural problem for this approach.[29] Content authentication frameworks based on cryptographic provenance (notably the C2PA standard) offer a more structurally sound approach by establishing a verifiable chain of custody from creation to consumption, but a watermarked image that is screenshotted and reposted loses its watermark while retaining its content.[30]

Platform-level interventions represent the most consequential lever available, since the reach of propaganda is largely a function of algorithmic amplification. Friction-based sharing interventions—prompts that ask users to read content before sharing or display accuracy cues at the point of sharing—have produced modest but statistically significant reductions in misinformation sharing.[31] The EU's Digital Services Act, in force since 2024, requires large platforms to assess and mitigate systemic disinformation risks and provide regulators with access to algorithmic audit processes, though effectiveness will depend on regulatory capacity and platform cooperation.[32] The EU AI Act classifies systems designed for subliminal manipulation or exploitation of psychological vulnerabilities as

prohibited or high-risk, but enforcement remains the central challenge because identifying violations requires detection capabilities that regulators largely do not currently possess.[33]

Traditional media literacy programmes face a fundamental problem: the quality of AI-generated content increasingly defeats the cues these programmes teach users to look for, since credible-looking sites, professional profile images, and fluent prose are all now generatable at negligible cost. Prebunking approaches drawing on inoculation theory have accumulated a more genuinely promising evidence base; randomised controlled trials found that prebunking interventions produced significant reductions in susceptibility to specific persuasion techniques including emotional appeals and false balance framing.[34] Lateral reading immediately seeking external information about a source rather than reading it in depth has been validated as a more effective real-world fact-checking strategy than conventional approaches, though motivational barriers limit adoption outside formal educational contexts.[35]

No single actor possesses the combination of capabilities, legitimacy, and reach required to address AI propaganda at scale. Multi-stakeholder frameworks including the Santa Clara Principles, the International Fact-Checking Network, and the Hiroshima AI Process have produced meaningful principles and voluntary commitments, though their effectiveness against determined state-sponsored actors is necessarily limited by their voluntary character.[36]

8. Research Gaps and Priorities

The empirical literature is heavily concentrated in high-income, English-speaking countries. Dynamics in environments with weaker institutional countermeasures and different cultural vulnerability profiles may be substantially different, and research investment outside these settings is urgently needed. Laboratory findings on susceptibility to AI-generated content do not straightforwardly translate into understanding of real-world behaviour change: evidence that subjects rate synthetic content as credible does not tell us whether that content would change votes or produce other downstream effects.[37] The mitigation literature predominantly studies isolated interventions rather than integrated systems, and systemic evaluation methods capable of capturing interaction effects across the full information environment are still lacking. Finally, the economics of AI propaganda who funds influence operations, what outcomes they target, what the market for commercial disinformation services looks like remain poorly mapped, and clearer economic analysis would substantially improve the targeting of regulatory and governance interventions.

9. Conclusion

AI did not invent propaganda, but it has removed nearly every logistical barrier that previously constrained propaganda's scale, personalisation, and speed. A small media cell with access to current AI tools can now produce and distribute influence content at volumes and with the cultural precision that would have required a substantial professional operation just a decade ago. This is not a minor incremental change; it is a structural shift in the economics of mass persuasion. The cognitive vulnerabilities exploited are not new either. Confirmation bias, the illusory truth effect, emotional contagion, and identity-based processing have always been features of human cognition. What is new is the precision and scale with which these vulnerabilities can now be targeted, and the organised human infrastructure the media cells operating across every significant platform that translates AI-generated content into coordinated campaigns migrating across the entire media ecosystem. The current portfolio of countermeasures is real but incomplete. Technical detection tools face a structural asymmetry against generation methods. Regulatory frameworks have made genuine progress but face serious enforcement challenges. Media literacy education helps but cannot match the volume and sophistication of AI-generated content. Platform governance reform has produced meaningful changes in specific contexts but has not addressed the fundamental tension between engagement-maximising business models and information quality.[29,31,34] What is ultimately at stake is the quality of the epistemic environment in which democratic deliberation occurs. Addressing this challenge seriously requires treating it as the democratic emergency it is, rather than as a technical problem that will eventually yield to a sufficiently clever algorithmic solution.

Acknowledgements

The authors acknowledge the academic resources made available through D.M.S. Mandal's Bhaurao Kakatkar College library. No external funding was received. The authors declare no conflicts of interest.

References

- Bradshaw, S., and Howard, P. N. (2019). The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation. Oxford Internet Institute, University of Oxford.

- Wardle, C., and Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe Report DGI(2017)09.
- Taylor, P. M. (1990). Munitions of the Mind: A History of Propaganda from the Ancient World to the Present Era. Manchester University Press.
- Benkler, Y., Faris, R., and Roberts, H. (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press.
- Goldstein, J. A., et al. (2023). Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv:2301.04246.
- Sunstein, C. R. (2017). Republic: Divided Democracy in the Age of Social Media. Princeton University Press.
- Chesney, R., and Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.
- Toews, R. (2023). Deepfakes are going to wreak havoc on society. Forbes Technology Council.
- Pariser, E. (2011). The Filter Bubble: What the Internet Is Hiding from You. Penguin Press.
- Hosseinmardi, H., et al. (2021). Examining the consumption of radical content on YouTube. PNAS, 118(32), e2101967118.
- Pariser, E. (2011). [See ref. 9]
- Cadwalladr, C., and Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica. The Guardian, 17 March 2018.
- Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. PNAS, 110(15), 5802–5805.
- Zuboff, S. (2019). The Age of Surveillance Capitalism. PublicAffairs.
- Ferrara, E., et al. (2016). The rise of social bots. Communications of the ACM, 59(7), 96–104.
- Kunda, Z. (1990). The case for motivated reasoning. Psychological Bulletin, 108(3), 480–498.

- Hasher, L., Goldstein, D., and Toppino, T. (1977). Frequency and the conference of referential validity. Journal of Verbal Learning and Verbal Behavior, 16(1), 107–112.
- Pennycook, G., Cannon, T. D., and Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. Journal of Experimental Psychology: General, 147(12), 1865–1880.
- Kramer, A. D. I., et al. (2014). Experimental evidence of massive-scale emotional contagion through social networks. PNAS, 111(24), 8788–8790.
- Iyengar, S., et al. (2019). The origins and consequences of affective polarization in the United States. Annual Review of Political Science, 22, 129–146.
- Tajfel, H., and Turner, J. C. (1986). The social identity theory of intergroup behavior. In Psychology of Intergroup Relations (pp. 7–24). Nelson-Hall.
- Gupta, A., et al. (2013). Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. Proceedings of WWW 2013, 729–736.
- Benkler, Y., Faris, R., and Roberts, H. (2018). [See ref. 4]
- Paul, C., and Matthews, M. (2016). The Russian “firehose of falsehood” propaganda model. RAND Corporation Perspective.
- Khamis, S., Ang, L., and Welling, R. (2017). Self-branding, “micro-celebrity” and the rise of Social Media Influencers. Celebrity Studies, 8(2), 191–208.
- Stanford Internet Observatory (2022). Unheard Voice: Evaluating five years of pro-Western covert influence operations. Stanford Digital Repository.
- Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.
- Moonshot CVE (2021). Online Radicalisation: From Hate Speech to Terrorism. Moonshot CVE Research Report.
- Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.

- Coalition for Content Provenance and Authenticity (C2PA) (2022). C2PA Technical Specification. https://c2pa.org/specifications/
- Pennycook, G., and Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. PNAS, 116(7), 2521–2526.
- European Commission (2022). Digital Services Act. Regulation (EU) 2022/2065.
- European Commission (2024). AI Act. Regulation (EU) 2024/1689.
- van der Linden, S., Roozenbeek, J., and Compton, J. (2020). Inoculating against fake news about COVID-19. Frontiers in Psychology, 11, 566790.
- Wineburg, S., and McGrew, S. (2019). Lateral reading and the nature of expertise. Teachers College Record, 121(11), 1–40.
- G7 Hiroshima AI Process (2023). International Guiding Principles for Advanced AI Systems. G7 Leaders Statement.
- Jerit, J., and Zhao, Y. (2020). Political misinformation. Annual Review of Political Science, 23(1), 77–94.