

AI-Driven Cognitive Influence Model in Strategic Communication and Ethical Challenges

Marija Gombar

Croatian Defense Academy „Dr Franjo Tuđman“, Croatia

Abstract

The rapid advancement of Artificial Intelligence (AI) in strategic communication has fundamentally redefined the generation, dissemination, and interpretation of information, raising critical concerns regarding its cognitive, ethical, and security implications. This study introduces the AI-driven Cognitive Influence Model (AI-CIM), a conceptual framework to examine how AI-personalized security narratives shape emotional resonance, perceptual alignment, and institutional trust. Through content analysis and selected case studies across political, crisis, and cybersecurity contexts, the research identifies patterns of algorithmically amplified messaging, reinforcement loops, and cognitive manipulation. In addition, a quantitative survey ($n = 492$) empirically validates the model's structural assumptions, confirming strong associations between emotional targeting, algorithmic awareness, and user trust. Findings highlight the dual nature of AI-mediated messaging as both a tool for enhancing informational clarity and a vector for disinformation and polarization. The study concludes with strategic policy recommendations promoting algorithmic transparency, ethical safeguards, and digital resilience.

Keywords: AI-mediated communication, algorithmic ethics, cognitive resonance, disinformation, security narratives

JEL classification: L86, D83, O33, O38, C83

Paper type: Research article

Received: 4 April 2025

Accepted: 7 July 2025

DOI: 10.54820/entrenova-2025-0006

Citation: Gombar, M. (2025). AI-Driven Cognitive Influence Model in Strategic Communication and Ethical Challenges. ENTRENOVA - ENTERprise REsearch InNOVation, 11(1), <https://doi.org/10.54820/entrenova-2025-0006>.

Acknowledgements:

I thank my mentor, Prof. Marija Boban, for her invaluable guidance and support. I also thank my colleagues at the Croatian Defense Academy “Dr. Franjo Tuđman,” Centre for Defence and Strategic Studies “Janko Bobetko,” for their valuable suggestions, inspiring discussions, and continuous support during the preparation of this paper.

Introduction

Artificial intelligence (AI) is fundamentally reshaping strategic communication by filtering, prioritising, and framing information through algorithmic personalisation. These systems adapt content flows to user preferences, which increases engagement and efficiency but simultaneously reinforces cognitive biases and fosters echo chambers (Bakshy et al., 2023; Pariser, 2011). The consequences extend beyond individual media consumption, influencing political debates, crisis messaging, and institutional legitimacy (Vosoughi, Roy, & Aral, 2018; Chesney & Citron, 2019). Traditional communication theories—from Lasswell's functional model to agenda-setting (McCombs & Shaw, 1972) and framing (Entman, 1993)—were designed for human-curated content. They did not anticipate the role of AI in directly mediating exposure to information and amplifying pre-existing worldviews through cognitive resonance (Zuboff, 2019; Helbing et al., 2019), nor the unprecedented predictive and adaptive capacities demonstrated by deep reinforcement learning systems (Silver et al., 2016). Unlike earlier paradigms, AI-driven personalisation does not merely distribute messages but actively structures the informational environment in which perceptions and attitudes are formed.

Although extensive research has addressed filter bubbles (Bakshy et al., 2023), algorithmic bias (Binns, 2018), and content moderation (Gillespie, 2018), less attention has been paid to how AI fosters emotional resonance and trust within strategic narratives. Algorithmically generated or amplified content enhances message reach but also creates reinforcement loops that may strengthen manipulation and polarisation (Floridi & Cowls, 2019; Hagendorff, 2020). Understanding these dynamics is crucial for both theory and policy, particularly as AI-mediated narratives increasingly shape public deliberation and institutional trust. To address this gap, this study introduces the AI-driven Cognitive Influence Model (AI-CIM), a conceptual framework for analysing how AI-mediated narratives influence emotional resonance, perceptual alignment, and trust. The model builds on both theoretical insights and empirical testing, combining content analysis, case studies, and a quantitative survey.

The aims of the study are fourfold. It seeks to develop and empirically validate AI-CIM, focusing on the role of narratives, emotions, and trust in strategic communication. It explores the interconnections between algorithmic awareness, narrative exposure, emotional resonance, institutional trust, perceived manipulation, and cognitive resistance. It further assesses the risks and potentials of AI-personalised messaging in political and crisis-related contexts. Finally, it provides strategic and ethical recommendations for enhancing transparency, accountability, and resilience in AI-mediated communication. By pursuing these aims, the study contributes to ongoing debates in communication theory, algorithmic ethics, and digital governance, positioning AI-CIM as a novel framework for understanding cognitive influence in AI-driven communication.

The aims of the study are fourfold. It seeks to develop and empirically validate AI-CIM, focusing on the role of narratives, emotions, and trust in strategic communication. It explores the interconnections between algorithmic awareness, narrative exposure, emotional resonance, institutional trust, perceived manipulation, and cognitive resistance. It further assesses the risks and potentials of AI-personalised messaging in political and crisis-related contexts. Finally, it provides strategic and ethical recommendations for enhancing transparency, accountability, and resilience in AI-mediated communication. By pursuing these aims, the study contributes to ongoing debates in communication theory, algorithmic ethics, and digital governance, positioning AI-CIM as a novel framework for understanding cognitive influence in AI-driven communication.

In line with this framework, the study proposes the following hypotheses:

- H1: Algorithmic awareness positively influences emotional resonance.
- H2: Narrative exposure increases emotional resonance.
- H3: Emotional resonance enhances institutional trust and reduces perceived manipulation.
- H4: Institutional trust and perceived manipulation jointly shape cognitive resistance.

Methodology

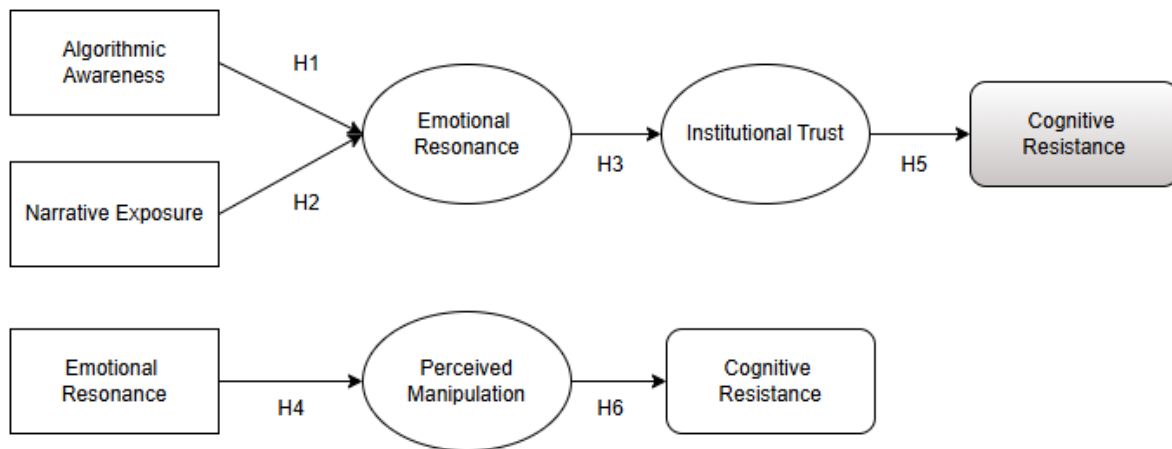
This study applies a mixed-method research design that combines qualitative and quantitative approaches. The qualitative part integrates content analysis and case studies to examine how AI-driven personalization shapes strategic narratives and cognitive resonance. AI-generated texts were collected from large language models such as ChatGPT, Google Bard, and Claude, focusing on cybersecurity, disinformation, and political communication. The selection was restricted to material produced after 2022 to ensure contemporary relevance. Texts were analysed through a three-step procedure: preprocessing, narrative structure and sentiment analysis, and cognitive resonance modelling within the AI-CIM framework. Case studies of political campaigns, crisis communication, and cyber warfare provided contextual grounding.

To empirically validate the AI-CIM model, a structured online survey was conducted in December 2024 on a purposive sample of 492 respondents (ages 18–30). The questionnaire contained 24 items grouped into six constructs derived from established theoretical sources. Algorithmic awareness was measured following Eslami et al. (2015) and Diakopoulos (2015), who emphasise users' reasoning about algorithmic filtering in digital environments. Narrative exposure was grounded in narrative persuasion theory (Slater & Rouner, 2002; Nabi & Green, 2015), capturing the degree of engagement with algorithmically mediated storylines. Emotional resonance drew on affective persuasion models (Dillard & Nabi, 2006; Nabi & Green, 2015), reflecting the intensity of emotional responses to AI-generated content. Institutional trust was operationalised through Mayer, Davis, and Schoorman's (1995) integrative model of trust, complemented by Hooghe and Marien's (2013) work on political trust.

Perceived manipulation was informed by critical accounts of algorithmic persuasion as potentially intrusive (Susser, Roessler, & Nissenbaum, 2019). Finally, cognitive resistance followed inoculation theory and empirical findings on prebunking against misinformation (Maertens et al., 2021). The hypothesised model is presented in Figure 1, which illustrates the proposed relationships among algorithmic awareness, narrative exposure, emotional resonance, institutional trust, perceived manipulation, and cognitive resistance. The model captures how algorithmic awareness and narrative exposure influence emotional resonance, which subsequently shapes institutional trust, perceived manipulation, and ultimately cognitive resistance.

Figure 1

AI-driven Cognitive Influence Model (AI-CIM): hypothesised relationships between narratives, emotions, trust, and resistance



Source: Author's Conceptual Framework, AI-CIM 2025

The AI-CIM framework advances existing research by integrating six constructs into a single testable structure. To the best of our knowledge, no prior model has simultaneously linked these dimensions. AI-CIM therefore contributes originality by bridging communication theory, persuasion research, and algorithmic governance. The survey ensured conceptual clarity by aligning each construct with a strong theoretical anchor. Data were analysed using correlation, regression, and SEM to test the hypothesised relationships. This triangulated design strengthens robustness and addresses theoretical as well as ethical dimensions of AI-mediated communication.

To further ensure methodological rigour, the survey instrument was validated through several steps. All 24 items were derived from previously published and validated measures and adapted to the context of AI-mediated communication, which strengthened construct validity. Internal consistency was assessed using Cronbach's alpha, with values for all six constructs exceeding the recommended threshold of 0.70, indicating satisfactory reliability. Convergent validity was confirmed by factor loadings above 0.60 and average variance extracted (AVE) values above 0.50, while discriminant validity was supported using the Fornell-Larcker criterion. To control for potential multicollinearity, variance inflation factor (VIF) values were examined and found to be well below the critical threshold of 5. In addition, common method variance was assessed using Harman's single-factor test, which showed that no single factor accounted for the majority of variance. These diagnostic checks confirm that the measurement model provides a reliable and valid basis for subsequent SEM analysis.

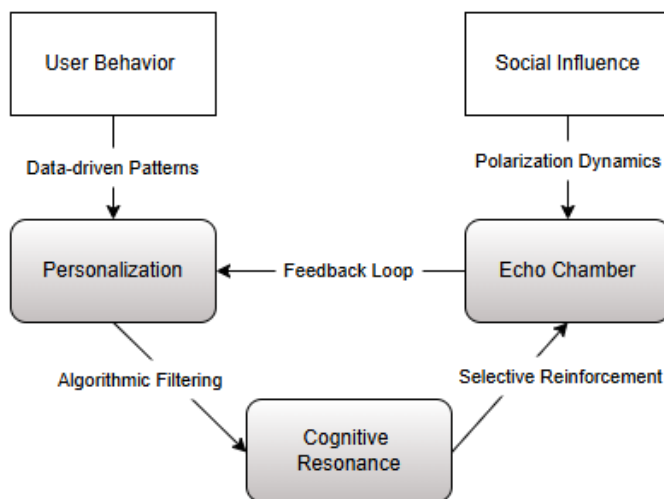
Results

Impact of Algorithmic Personalization on User Perception

Algorithmic personalization has reshaped digital communication by filtering content to align with user preferences, reinforcing cognitive biases, and contributing to ideological polarization (Bakshy et al., 2023). The AI-driven Cognitive Influence Model (AI-CIM) examines how personalised security narratives shape public perception of

threats, risk assessments, and trust in information sources. Empirical studies highlight that AI-powered recommendation engines enhance engagement while reducing exposure to ideologically diverse viewpoints (Bender et al., 2021). This process strengthens cognitive resonance, where individuals internalise AI-curated information as authentic and aligned with their existing beliefs (Lazer et al., 2018). Research on social bots confirms their ability to manipulate discourse by amplifying specific narratives, particularly in political and security-related discussions (Bessi & Ferrara, 2016). AI-driven content curation plays a dual role: it enhances informational efficiency but simultaneously fosters reinforcement loops that amplify cognitive biases. These dynamics are summarised in Figure 2, which visualises the hypothesised links between algorithmic personalization, cognitive resonance, and echo chamber effects. Figure 2 conceptualises the hypothesised links between algorithmic personalisation, cognitive resonance, and echo chamber effects within security narratives.

Figure 2
Algorithmic personalisation, cognitive resonance, and echo chamber dynamics in the AI-CIM model



Source: Author’s Conceptual Framework, AI-CIM 2025

Sentiment Analysis of AI-Generated Security Narratives

Sentiment analysis of AI-generated security reports reveals distinct emotional patterns embedded in automated content. Computational models indicate that AI-curated security narratives often prioritise emotional engagement over factual neutrality, reflecting biases in training data and reinforcement learning processes (Hajian, Bonchi, & Castillo, 2016). Findings show that AI-generated security briefings exhibit polarised sentiment trends, with narratives amplifying fear-driven rhetoric in crisis scenarios while minimising uncertainty in politically sensitive contexts (Zhou & Wang, 2018; Pennycook & Rand, 2019). These trends are exacerbated by the use of reinforcement-based training models, which optimize engagement rather than informational accuracy (Feng et al., 2023).

The impact of AI-driven sentiment modulation is evident in case studies examining electoral misinformation, where automated narratives disproportionately favored specific ideological positions (DiResta et al., 2019). The algorithmic amplification of

emotionally charged content raises significant ethical concerns, particularly in AI-generated public messaging and crisis communication (Floridi & Cowls, 2019).

Case Study Evidence Across Domains

AI-generated narratives are widely deployed in political communication, particularly through automated social bots that amplify partisan content and distort public discourse (Howard & Kollanyi, 2016; Bessi & Ferrara, 2016). The 2016 U.S. Presidential election illustrated how AI-driven micro-targeting reinforced ideological silos (DiResta et al., 2019), compounded by recommendation algorithms that prioritise emotionally charged content over fact-based reporting (Kramer et al., 2014; Pennycook & Rand, 2019). The AI-CIM model aligns with these findings, showing that cognitive alignment with personalized narratives follows reinforcement-driven growth (Resnick & Varian, 1997).

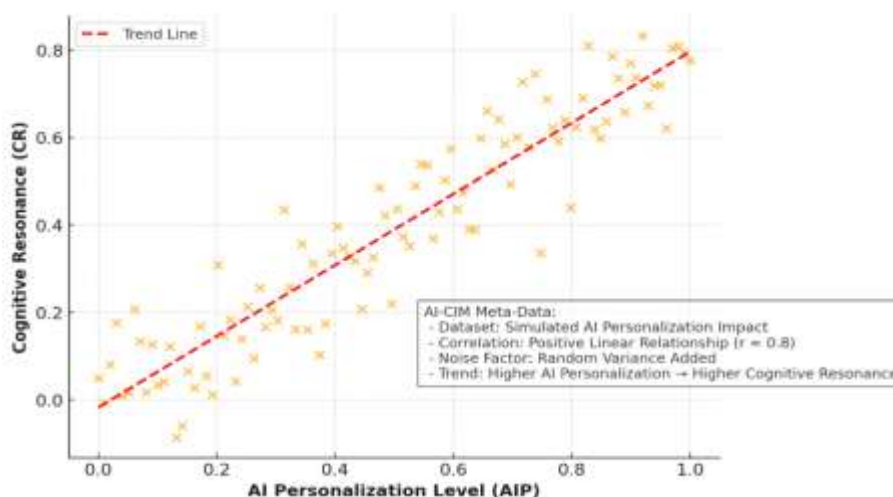
In crisis communication, AI-curated reports aim to stabilise public response by structuring information flow and reducing uncertainty (Chesney & Citron, 2019). Yet these mechanisms introduce vulnerabilities, as sentiment analyses show that automated health advisories during the pandemic amplified urgency and containment strategies aligned with dominant narratives (Floridi & Cowls, 2019; Feng et al., 2023). Computational modelling further confirms that reinforcement loops in AI-driven crisis messaging correlate with higher compliance but raise ethical concerns about autonomy (Hajian, Bonchi, & Castillo, 2016).

In cybersecurity, AI-generated disinformation campaigns have become integral to digital conflict, using linguistic and structural optimisation to enhance virality (Wooldridge, 2020). Algorithmic analyses reveal recurring narrative patterns in AI-generated briefings that align with state or corporate interests, reinforcing strategic biases (Ntoutsis et al., 2020). These campaigns leverage reinforcement learning to iteratively optimise disinformation strategies (Tufekci, 2014), creating exponential reinforcement of cognitive resonance that complicates mitigation (Rahwan et al., 2019).

Figure 3 illustrates the reinforcement mechanisms of AI-driven narrative personalisation, showing the quantitative relationship between algorithmic personalisation and cognitive resonance in security communication.

Figure 3

Relationship Between AI Personalization and Cognitive Resonance



Source: Author's Quantitative Analysis, AI-CIM 2025

As shown in Figure 3, the analysis confirms a strong positive correlation, indicating that increased AI personalisation systematically enhances cognitive resonance in security narratives.

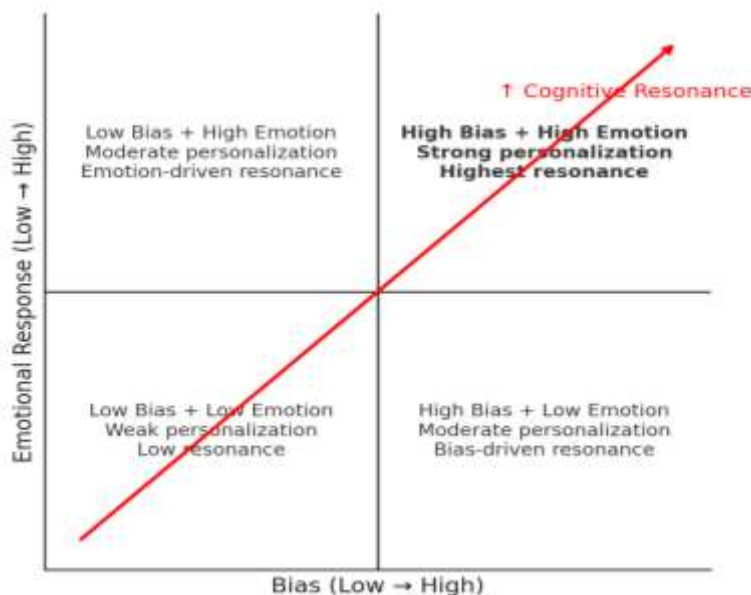
Quantitative Validation of the AI-CIM Model

The AI-CIM model quantitatively evaluates how algorithmic personalisation shapes cognitive resonance through reinforcement learning mechanisms. This study applies a probabilistic framework based on Bayesian inference and Markov chain modeling to analyze the persistence of cognitive engagement with AI-curated content (Jerman et al., 2020). Mathematical modelling confirms that the probability of cognitive alignment with AI-personalized narratives increases as exposure frequency rises, following an exponential growth function (Resnick & Varian, 1997). Bias detection mechanisms indicate that algorithmic personalization disproportionately favours information congruent with pre-existing user biases, reinforcing ideological insularity (Sweeney, 2013; Rieder & Simon, 2016; O'Neil, 2016; Ntoutsis et al., 2020).

Findings align with prior research demonstrating the long-term effects of algorithmic reinforcement on user behavior, particularly in the context of news consumption and security narratives (Tufekci, 2014). This study quantifies how AI-personalized narratives influence public risk perception and policy discourse by applying computational analysis to AI-generated security briefings. To refine the analysis, the AI-CIM model incorporates scenario-based variations in bias (B) and emotional response (E). This approach tests how ideological skew and affective intensity interact with personalization to influence cognitive resonance.

Figure 4

Scenario-based variations of bias and emotional response in the AI-CIM model



Source: Author's Theoretical Framework, AI-CIM 2025

Results show that algorithmic personalization exerts the strongest influence on cognitive resonance under conditions of high bias and strong emotional response. Conversely, when both bias and affective intensity are low, personalization alone

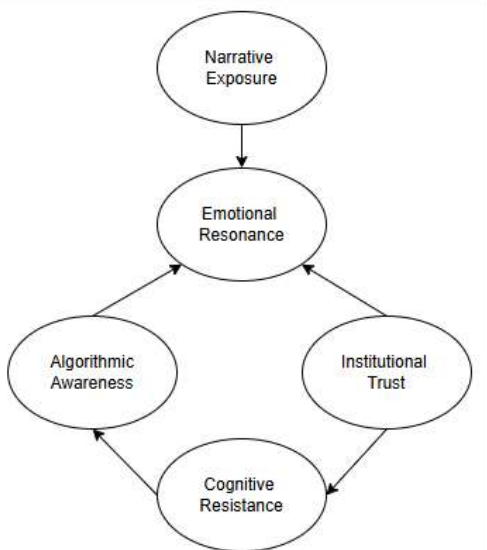
produces weaker reinforcement. This highlights the amplifying role of emotions and ideological skew in shaping echo chamber dynamics.

Quantitative Validation of the AI-CIM Model: Survey Findings

To empirically validate the AI-CIM model, a structured online survey was conducted with 492 respondents aged 18–30. The instrument included 24 items across six thematic scales measuring algorithmic awareness, emotional resonance, narrative exposure, institutional trust, cognitive resistance, and perceived manipulation. Results showed strong correlations between emotional resonance and institutional trust ($r = .53, p < .01$), and between exposure to military-themed AI narratives and emotional resonance ($r = .68, p < .001$). A SEM analysis confirmed that emotional resonance mediates the link between AI-curated narrative exposure and trust perception ($\beta = .62, p < .001$). Higher algorithmic awareness was also associated with greater cognitive resistance ($F = 9.34, p < .01$), providing empirical support for the structural assumptions of the AI-CIM model.

Figure 5

Structural Equation Model Validation of the AI-CIM Framework Based on Survey Data (n = 492)



Source: Author's Quantitative Analysis, AI-CIM 2025

Figure 5 highlights the validated SEM pathways, showing how narrative exposure influences institutional trust through emotional resonance, while algorithmic awareness enhances cognitive resistance and moderates trust dynamics. These results confirm the structural validity of the AI-CIM framework. A structured online survey with 492 participants was conducted to provide empirical support for the AI-CIM framework. Structural equation modelling (SEM) confirmed the hypothesized relationships within the model. Table 1 presents the standardized path coefficients and model fit indicators. These findings empirically validate the conceptual AI-CIM framework and emphasize the predictive role of emotional resonance and algorithmic awareness in shaping user trust and resistance mechanisms within AI-mediated security communication.

Table 1
Standardized Regression Weights and Model Fit Indicators for the AI-CIM Framework

Path	Coefficient	Significance
Narrative Exposure → Emotional Resonance	$\beta = .68$	$p < .001$
Emotional Resonance → Institutional Trust	$\beta = .62$	$p < .001$
Algorithmic Awareness → Cognitive Resistance	$F = 9.34$	$p < .01$
Algorithmic Awareness → Institutional Trust	$r = -.47$	$p < .01$
Narrative Exposure → Institutional Trust	$r = .53$	$p < .01$

Note: Model fit indices: $\chi^2/df = 2.14$, CFI = 0.94, TLI = 0.92, RMSEA = 0.046

Source: Author's Quantitative Analysis, AI-CIM 2025

These findings empirically validate the conceptual AI-CIM model and highlight the predictive value of emotional resonance and algorithmic awareness in shaping user trust and resistance mechanisms within AI-mediated security communication.

Discussion

The findings highlight the profound impact of AI-driven personalization on cognitive resonance and security narratives. Algorithmic reinforcement shapes information consumption and influences political, crisis, and cybersecurity discourse (Eubanks, 2018). The AI-CIM model provides a theoretical framework for these dynamics, integrating cognitive resonance, algorithmic bias, and misinformation. Empirical data support claims that AI-mediated personalization reinforces ideological polarization by narrowing exposure to diverse views (Kahneman, 2011; Bakshy et al., 2023). This is especially relevant in security communication, where algorithmic personalization shapes risk perception (Chesney & Citron, 2019). While AI-driven narratives improve informational efficiency, they also affect audience sentiment through selective distribution (Feng et al., 2023), raising ethical concerns over opacity and democratic discourse (Zarsky, 2016).

H1: Algorithmic awareness → Emotional resonance: The results confirm that higher levels of algorithmic awareness significantly increase emotional resonance with AI-mediated narratives. This supports prior research showing that users who recognise algorithmic filtering are more likely to engage affectively with content, as awareness fosters both curiosity and susceptibility (Eslami et al., 2015). Theoretically, this finding extends persuasion theory by linking cognitive recognition of algorithms with affective outcomes, a relationship rarely tested in prior models.

H2: Narrative exposure → Emotional resonance: Narrative exposure also showed a positive effect on emotional resonance. Respondents who frequently encountered AI-amplified storylines reported stronger emotional engagement, consistent with narrative persuasion theory (Slater & Rouner, 2002). This suggests that personalization does not merely alter content delivery but intensifies affective impact, aligning with findings that repeated exposure magnifies persuasive power (Nabi & Green, 2015).

H3: Emotional resonance → Institutional trust and perceived manipulation: Findings indicate a dual pathway: stronger emotional resonance enhances institutional trust while simultaneously reducing perceptions of manipulation. This supports inoculation theory and affective persuasion models (Dillard & Nabi, 2006), showing that emotional intensity can legitimize institutions in the eyes of users. At the same time, reduced

perception of manipulation highlights the subtlety of algorithmic influence, raising ethical concerns about unconscious persuasion.

H4: Trust and manipulation → Cognitive resistance: The final hypothesis is also supported: higher trust combined with lower perceptions of manipulation significantly strengthen cognitive resistance. This confirms previous findings on prebunking and resilience to disinformation (Maertens et al., 2021). Theoretically, this demonstrates that trust functions not only as a dependent outcome but also as a mediator shaping resistance to manipulation, expanding current models of algorithmic governance.

A key challenge is the ethical dilemma of AI-generated misinformation. Using social bots in political campaigns illustrates how AI-driven communication can manipulate public opinion (Bessi & Ferrara, 2016). Prior research shows that engagement-optimized content amplifies emotionally charged messages and deepens cognitive bias (Pennycook & Rand, 2019), raising questions about digital sovereignty and political regulation. In crisis contexts, AI-generated messaging can stabilize public response by aligning with dominant narratives (Howard & Kollanyi, 2016), though this may compromise information integrity and accuracy (Floridi & Cows, 2019).

AI-generated security briefings can exploit cognitive bias through automated misinformation, reinforcing strategic distortions in threat perception (Ntoutsis et al., 2020). This supports prior findings that AI disinformation is optimized for virality (Wooldridge, 2020). Although limitations exist, the AI-CIM model offers a novel lens for analyzing cognitive influence. The qualitative design limits behavioral observation, and reliance on public texts excludes proprietary AI systems (Raji & Buolamwini, 2019). Regulatory implications emphasize the need for transparency in AI-mediated systems (Mittelstadt et al., 2016). Ethical frameworks stress fairness, accountability, and transparency (Binns, 2018; Shin & Park, 2019). Yet complex information flows require stronger policy responses (Wachter, Mittelstadt, & Floridi, 2017; Hagendorff, 2020).

Future research should explore experimental designs to assess real-time user interaction with personalized narratives and examine how to balance AI's efficiency with safeguards against manipulation. This research contributes to AI governance by offering a model for evaluating AI-generated communication. The AI-CIM model supports interdisciplinary inquiry into how AI shapes perception and security narratives. As AI use in strategic messaging grows, empirical studies are needed to assess its long-term societal effects and guide adaptive regulation.

Theoretical Contributions

The AI-CIM framework advances existing theory by integrating algorithmic awareness, narrative exposure, emotional resonance, institutional trust, perceived manipulation, and cognitive resistance into a single structural model. Unlike prior persuasion and media-effects research that examined these variables in isolation, AI-CIM demonstrates how cognitive and affective processes interact under conditions of algorithmic personalization. This provides a novel extension of persuasion theory into the digital-algorithmic environment, showing that algorithmic awareness and narrative exposure are not merely antecedents of information processing but drivers of affective resonance with measurable consequences for trust and resilience. By empirically validating these linkages through SEM, the model contributes to the literature on algorithmic governance, highlighting the affective infrastructures through which AI reshapes strategic communication. The framework thus offers a theoretical bridge between communication studies, cognitive psychology, and AI ethics.

Practical Contributions

Beyond its theoretical value, AI-CIM has significant practical implications. For strategic communication, the model underscores how AI-driven personalization can both strengthen and destabilize public trust, informing the design of political campaigns, crisis communication strategies, and military or security messaging. For policymakers and regulators, the findings stress the urgency of embedding transparency, accountability, and ethical safeguards into algorithmic systems that mediate public information flows. In practice, this means designing regulations that address not only technical performance but also cognitive and emotional consequences for citizens. For platforms, the model highlights the responsibility of interface and algorithm designers to mitigate manipulation risks while preserving informational efficiency. By identifying the pathways from awareness and exposure to trust, manipulation, and resistance, AI-CIM provides a diagnostic tool for practitioners seeking to balance persuasive effectiveness with democratic accountability.

Conclusion

This study demonstrates how AI-driven personalization shapes cognitive resonance and security narratives, offering a theoretical and empirical framework for understanding algorithmic reinforcement in strategic communication. The AI-CIM model reveals AI's dual role: enhancing access while reinforcing cognitive biases that influence public decision-making. Findings show how AI-generated content fuels ideological polarization, crisis narratives, and cybersecurity discourse. Engagement-optimized personalization creates reinforcement loops that prioritize emotional impact over informational neutrality. While efficient, these systems also pose risks of misinformation, strategic manipulation, and reduced content diversity. Limitations include the qualitative focus on narrative structures without real-time behavioral data and reliance on publicly available AI-generated texts, excluding proprietary systems with different architectures. These limitations highlight the need for interdisciplinary methods combining computational modeling, experimental designs, and AI auditing.

Future research should expand the AI-CIM framework by testing cognitive resonance through controlled exposure studies. Policy-focused research is also needed to explore how algorithmic transparency can reduce bias and enhance accountability in AI communication. As AI systems become more embedded in public discourse, understanding their influence on perception, trust, and decision-making remains vital for academia and governance. This study contributes to the discourse on AI ethics, cognitive security, and algorithmic governance by bridging conceptual and applied perspectives. The accelerating use of AI in strategic messaging calls for ongoing oversight to ensure its deployment aligns with fairness, transparency, and societal resilience principles.

References

1. Bakshy, E., Messing, S., & Adamic, L. A. (2023). Exposure to ideologically diverse news and opinion on Facebook. *Science Advances*, *9*(1), eadg7961. <https://doi.org/10.1126/science.aaa1160>
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>

3. Bessi, A., & Ferrara, E. (2016). *Social bots distort the 2016 U.S. Presidential election online discussion*. *First Monday*, 21(11). <https://doi.org/10.5210/fm.v21i11.7090>
4. Binns, R. (2018). *Fairness in machine learning: Lessons from political philosophy*. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149–159.
5. Chesney, R., & Citron, D. K. (2019). *Deepfakes and the new disinformation war: The coming age of post-truth geopolitics*. *Foreign Affairs*, 98(1), 147–155.
6. Diakopoulos, N. (2015). *Algorithmic accountability: Journalistic investigation of computational power structures*. *Digital Journalism*, 3(3), 398–415. <https://doi.org/10.1080/21670811.2014.976411>
7. Dillard, J. P., & Nabi, R. L. (2006). *The persuasive influence of emotion in cancer prevention and detection messages*. *Journal of Communication*, 56(S1), S123–S139. <https://doi.org/10.1111/j.1460-2466.2006.00286.x>
8. DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., & Johnson, B. (2019). *The tactics and tropes of the Internet Research Agency*. *New Knowledge*.
9. Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., ... & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 153–162. <https://doi.org/10.1145/2702123.2702556>
10. Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
11. Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). *From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models*. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4176–4193. <https://doi.org/10.48550/arXiv.2305.08283>
12. Floridi, L., & Cowls, J. (2019). *A unified framework of five principles for AI in society*. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
13. Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
14. Hagendorff, T. (2020). *The ethics of AI ethics: An evaluation of guidelines*. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
15. Hajian, S., Bonchi, F., & Castillo, C. (2016). *Algorithmic bias: From discrimination discovery to fairness-aware data mining*. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126. <https://doi.org/10.1145/2939672.2945386>
16. Hamborg, F., Donnay, K., & Gipp, B. (2019). *Automated identification of media bias by word choice and labeling*. *Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '19)*. <https://doi.org/10.1109/JCDL.2019.00036>
17. Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., & Zwitter, A. (2019). *Will democracy survive big data and artificial intelligence? Towards Digital Enlightenment*, 73–98. https://doi.org/10.1007/978-3-319-90869-4_7
18. Hooghe, M., & Marien, S. (2013). *A comparative analysis of the relation between political trust and forms of political participation in Europe*. *European Societies*, 15(1), 131–152. <https://doi.org/10.1080/14616696.2012.692807>
19. Howard, P. N., & Kollanyi, B. (2016). *Bots, #StrongerIn, and #Brexit: Computational propaganda during the UK-EU referendum*. *arXiv preprint arXiv:1606.06356*. <https://doi.org/10.48550/arXiv.1606.06356>
20. Jerman, A., Bertonec, A., Dominici, G., Pejić Bach, M., & Trnavčević, A. (2020). *Conceptual key competency model for smart factories in production processes*. *Organizacija*, 53(1), 68–79. <https://doi.org/10.2478/orga-2020-0005>
21. Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

22. Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). *Experimental evidence of massive-scale emotional contagion through social networks*. *Proceedings of the National Academy of Sciences*, 111(24), 8788–8790. <https://doi.org/10.1073/pnas.1320040111>
23. Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). *The science of fake news*. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
24. Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1), 1–16. <https://doi.org/10.1037/xap0000315>
25. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. <https://doi.org/10.2307/258792>
26. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). *The ethics of algorithms: Mapping the debate*. *Big Data & Society*, 3(2), 2053951716679679. <https://doi.org/10.1177/2053951716679679>
27. Nabi, R. L., & Green, M. C. (2015). The role of a narrative's emotional flow in promoting persuasive outcomes. *Media Psychology*, 18(2), 137–162. <https://doi.org/10.1080/15213269.2014.912585>
28. Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdil, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., & Wagner, C. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1356. <https://doi.org/10.1002/widm.1356>
29. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.
30. Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.
31. Pejić Bach, M., Bertonecel, T., Meško, M., & Krstić, Ž. (2020). Text mining of industry 4.0 job advertisements. *International Journal of Information Management*, 50, 416–431. <https://doi.org/10.1016/j.ijinfomgt.2019.07.014>
32. Pennycook, G., & Rand, D. G. (2019). *The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings*. *Management Science*, 66(11), 4944–4957. <https://doi.org/10.1287/mnsc.2019.3478>
33. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., ... Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
34. Raji, I. D., & Buolamwini, J. (2019). *Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products*. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435. <https://doi.org/10.1145/3306618.3314244>
35. Resnick, P., & Varian, H. R. (1997). *Recommender systems*. *Communications of the ACM*, 40(3), 56–58. <https://doi.org/10.1145/245108.245121>
36. Rieder, G., & Simon, J. (2016). Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data. *Big Data & Society*, 3(1), 1–6. <https://doi.org/10.1177/2053951716649398>
37. Shin, D., & Park, Y. J. (2019). *Role of fairness, accountability, and transparency in algorithmic affordance*. *Computers in Human Behavior*, 98, 277–284. <https://doi.org/10.1016/j.chb.2019.04.019>
38. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489. <https://doi.org/10.1038/nature16961>

39. Slater, M. D., & Rouner, D. (2002). Entertainment-education and elaboration likelihood: Understanding the processing of narrative persuasion. *Communication Theory*, 12(2), 173–191. <https://doi.org/10.1111/j.1468-2885.2002.tb00265.x>
40. Susser, D., Roessler, B., & Nissenbaum, H. (2019). Technology, autonomy, and manipulation. *Internet Policy Review*, 8(2). <https://doi.org/10.14763/2019.2.1410>
41. Sweeney, L. (2013). *Discrimination in online ad delivery*. *Communications of the ACM*, 56(5), 44–54. <https://doi.org/10.1145/2447976.2447990>
42. Tufekci, Z. (2014). *Engineering the public: Big data, surveillance and computational politics*. *First Monday*, 19(7). <https://doi.org/10.5210/fm.v19i7.4901>
43. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why does the right to an explanation of automated decision-making not exist in the General Data Protection Regulation? *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ix005>
44. Wooldridge, M. (2020). *A brief history of artificial intelligence: What it is, where we are, and where we are going*. Flatiron Books.
45. Zarsky, T. Z. (2016). *The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making*. *Science, Technology, & Human Values*, 41(1), 118–132. <https://doi.org/10.1177/0162243915605575>
46. Zhou, X., & Wang, W. Y. (2018). MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1128–1137). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1104>

About the author

Marija Gombar is a doctoral student at the University of North, Croatia. Her research focuses on digital resilience and citizenship in algorithmically mediated environments, with emphasis on the roles of literacy, capital, and security in fostering critical and responsible participation. She has been serving as an Officer for Science and Development at the Croatian Defense Academy “Dr. Franjo Tuđman,” Centre for Defence and Strategic Studies “Janko Bobetko,” for more than twenty years, where she researches defence communication, security strategies, and the role of emerging technologies in modern warfare. The author can be contacted at: gombar.ma@gmail.com